UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK

IN RE: OpenAI, INC., COPYRIGHT INFRINGEMENT LITIGATION This Document Relates To: 1:23-cv-11195-SHS-OTW 1:24-cv-01515-SHS-OTW 1:24-cv-03285-SHS-OTW 1:24-cv-04872-SHS-OTW 1:25-cv-04315-SHS-OTW	25-md-3143 (SHS) (OTW)

DEFENDANTS' RESPONSE TO NEWS PLAINTIFFS' RULE 56.1 STATEMENT

VOLUME III – FACTS 1175-1421

VOLUME III

TABLE OF CONTENTS

IV. Defendants' Infringement of News Plaintiffs' Asserted Works Ripe for Summary Judgment 742

A. Acquisition, Distribution, and Training Infringement 742

1. The New York Times and Daily News Plaintiffs 742

2. CIR and The Intercept 752

3. Ziff Davis 769

B. Copilot RAG and Output Infringement 785

1. The Times and DNP 788

2. CIR 802

V. OpenAI's Removal of News Plaintiffs' CMI 807

A. News Plaintiffs' conveyance of CMI in connection with their Asserted Works. 807

B. Absence of Authority for OpenAI to Remove News Plaintiffs' CMI and Use Their Content 828

C. OpenAI's Intentional Removal of News Plaintiffs' CMI 844

1. Facts Regarding OpenAI's Intentional Removal of All News Plaintiffs' CMI 844

2. OpenAI's Removal of The Times's and DNP's CMI 902

3. OpenAI's Removal of Ziff Davis CMI 926

4. OpenAI's Removal of CIR's and The Intercept's CMI 940

Pursuant to Local Civil Rule 56.1, Defendants OpenAI and Microsoft respectfully submit this Response to Plaintiffs' Corrected Statement of Undisputed Material Facts in Support of their Motion for Partial Summary Judgment. ECF No. 1985 (News) (the "Statement").

Plaintiffs' Statement fails to comply with Local Civil Rule 56.1(a). That rule requires that a motion for summary judgment "be accompanied by a separate, **short, and concise statement**, in numbered paragraphs, of the **material facts** as to which the moving party contends there is no genuine issue to be tried." Local Civ. R. 56.1(a) (emphasis added). That is not what Plaintiffs did. Plaintiffs submitted 940 [*1,825*] separately numbered paragraphs spanning nearly 160 [*440*] pages. Such a voluminous submission does not comply with the local rule. *Tripathy v. McCloskey*, No. 21-CV-6584 (CS), 2024 WL 2135623, at *1 (S.D.N.Y. May 13, 2024) (striking plaintiff's 356-page submission for "ignor[ing] the Rule's requirement that a statement be short and concise"). Plaintiffs' approach has made it extremely difficult for Defendants to respond fully while also providing their responses in a format useful to the Court.

To deal with this massive and improper Statement, Defendants have attempted to streamline their response for the benefit of the Court as follows:

Undisputed/Disputed/Disputed in Part. Every fact includes one of these responses, to make clear Defendants' position under Local Rule 56.1(b). Facts designated "Undisputed" are facts that Defendants have elected not to dispute solely for purposes of Plaintiffs' Motion for Summary Judgment in this matter, but does not constitute a binding admission for any other purpose. Facts designated "Disputed" or "Disputed in Part" are facts that Defendants are disputing.

In addition, Defendants have included certain objections and other responses to Plaintiffs' facts where appropriate. Those additional responses and the authority supporting those responses are identified below. For the Court's convenience and to avoid further lengthening this response, Defendants do not repeat their supporting authority for these issues in every response. Instead, Defendants state these objections plainly for ease of review and incorporate the below authority into each of those responses. Where warranted and helpful for the Court, Defendants also provide additional, paragraph-specific information to assist the Court's assessment.

Immaterial. Facts designated "Immaterial" are facts that, even if true, do not help resolve any question currently before the Court on summary judgment, i.e., they do not "affect the outcome of the suit under the governing law." *I.M. v. United States*, 362 F. Supp. 3d 161, 190 (S.D.N.Y. 2019). Rule 56.1 statements are "immaterial" if they "do not provide any basis to rule in favor of either party for summary judgment even if they are disputed." *See Parks v. Lebhar-Friedman, Inc.*, No. 04-cv-7133-DCP-KNF, 2008 WL 3833802, at *4 (S.D.N.Y. Aug. 11, 2008). Facts designated as "Immaterial" below thus reflect facts that are not legally significant because it does not impact an element of any claim or defense at issue in Plaintiffs' summary judgment motion. Defendants' designation of certain facts as "immaterial" does not concede that any fact not designated as such is material for purposes of adjudicating Plaintiffs' Motion, or for any other purpose.

"Not Supported by Cited Evidence." Facts designated as "Not Supported by Cited Evidence" are facts with incorrect citations or where the cited evidence does not support the stated fact. *Holtz*, 258 F.3d at 74 (A "Rule 56.1 statement is not itself a vehicle for making factual assertions that are otherwise unsupported in the record. Where ... the record does

not support the assertions in a Local Rule 56.1 statement, those assertions should be disregarded and the record reviewed independently.").

"Not Supported by Admissible Evidence." Facts designated as "Not Supported by Admissible Evidence" are facts that are not supported by admissible evidence as required by Local Rule 56.1, and therefore are not properly before the Court. S.D.N.Y. Local Rule 56.1 (d) ("Each statement by the movant … must be followed by citation to evidence that would be admissible[.]"); *Epstein v. Kemper Ins. Companies*, 210 F. Supp. 2d 308, 314 (S.D.N.Y. 2002) (statements in a Rule 56.1 statement "are inappropriate if they are not based on personal knowledge, contain inadmissible hearsay, … or do not cite to supporting evidence").

"Improper Legal Argument." Facts designated as an "Improper Legal Argument" are facts where the assertion therein constitutes improper attorney argument and/or is conclusory. These facts require no response. *Epstein v. Kemper Ins. Companies*, 210 F. Supp. 2d 308, 314 (S.D.N.Y. 2002) (statements in a Rule 56.1 statement "are inappropriate if they are conclusory or argumentative[.]"); *In re LIBOR-Based Fin. Instruments Antitrust Litig.*, 801 F. Supp. 3d 330, 369 (S.D.N.Y. 2025) (quoting *Kesner v. Buhl*, 590 F. Supp. 3d 680, 691 (S.D.N.Y. 2022)) (the Court sets aside "[t]he portions of [a responding party's] 56.1 Statement that contain legal argument bereft of factual matter[.]").

"Incomplete." Facts designated as "Incomplete" are facts where Plaintiffs have "introduce[d] all or part of a statement" and Defendants introduce "any other part—or any other statement—that in fairness ought to be considered at the same time" pursuant to Federal Rule of Evidence 106.

IV. Defendants' Infringement of News Plaintiffs' Asserted Works Ripe for Summary Judgment

A. Acquisition, Distribution, and Training Infringement

1. The New York Times and Daily News Plaintiffs

1175. The coverage score between a candidate copy in OpenAI's datasets and an Asserted Work measures how similar the copy is to the Asserted Work. Goldstein Opening Rep. ¶¶ 282-286.	1175. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: OpenAI disputes that a coverage score between two documents "measures how similar" the documents are. A coverage score, a term invented by Dr. Goldstein and not used in the field, measures the percentage of 3-grams (i.e., sequences of 3 adjacent words) from the Asserted Work that are also present in the copy. *See* News Fact 1176, *infra*. But, pursuant to Fed. R. Evid. 106, documents can share 3-grams in common without being "similar." For instance, if two documents contain identical 3-grams but in entirely different orders, they are unlikely to be "similar" documents. Or, if a document contains many unique 3-grams in addition to any shared 3-grams, the coverage score would not "measure[] how similar the copy is to the Asserted Work." S. Krishnan Opp. Decl. (News) Ex. A (S. Krishnan Rebuttal Rep. (Goldstein)) ¶¶ 64-67. The facts set forth in Paragraph 1175 are disputed because the cited evidence is incorrect. Microsoft defers to OpenAI.
1176. The coverage score between a candidate copy and an Asserted Work indicates the percentage of 3-grams	1176. **Undisputed.**

(i.e., sequences of 3 adjacent words) from the Asserted Work that are present in the copy after pre-processing of both the copy and the Asserted Work. Goldstein Opening Rep. ¶¶ 282-286.	Microsoft defers to OpenAI.
1177. The pre-processing of the copy and the Asserted Work includes removing any extraneous non-letter characters and removing from each Asserted Work any material in quotation marks. Goldstein Opening Rep. ¶¶ 282-286.	1177. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 1177 are disputed because they are not supported by the cited evidence. Specifically disputed to the extent the statement implies Dr. Goldstein successfully omitted third-party or quoted material. Pursuant to Fed. R. Evid. 106, OpenAI's expert Dr. Sanjay Krishnan identified limitations of Dr. Goldstein's text preprocessing and noted that preprocessing performed by Dr. Goldstein had failed to exclude quoted material. *See* S. Krishnan Opp. Decl. (News) Ex. A (S. Krishnan Rebuttal Rep. (Goldstein)) ¶¶ 69–72, 101–106. Microsoft defers to OpenAI.
1178. A shared 16-gram between a candidate copy and an Asserted Work means that a span of 16 adjacent words is common between the copy and the Asserted Work. Goldstein Opening Rep. ¶¶ 287, 290.	1178. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 1178 are disputed because the cited evidence is incorrect. Specifically, these facts are disputed to the extent it assumes that the preprocessing of an Asserted Work and the identification of shared 16-grams was performed in a consistent or correct manner. As discussed above and pursuant to Fed. R. Evid. 106, OpenAI's expert identified several instances in Dr. Goldstein's preprocessing where quoted material was not excluded

	from the processed Asserted Work. *See* S. Krishnan Opp. Decl. (News) Ex. A (S. Krishnan Rebuttal Rep. (Goldstein)) ¶¶ 69–72, 101–106. And because Dr. Goldstein manipulated the Asserted Work text by attempting to remove quotations, the processed text contains consecutive 16-word strings that did not actually appear in the Asserted Work. *Id.* Microsoft defers to OpenAI.
1179. To consider an Asserted Work a match to a given candidate copy, Dr. Goldstein required that the Asserted Work share a 16-gram with the copy, and have a coverage score that meets a specific threshold. Goldstein Opening Rep. ¶¶ 287, 288.	1179. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 1179 are disputed because the cited evidence is incorrect. Specifically, these facts are disputed to the extent it implies Dr. Goldstein has applied any "specific threshold" with consistency to determine a "match" between an Asserted Work and a training data record. In fact, Dr. Goldstein has applied a variety of thresholds and characterized the results of each of these thresholds as "matches," and he has not supplied any clear reason why any specific threshold should be considered more accurate or less error-prone than his other proposed thresholds. For example, his Opening Report considered "matches" with coverage scores as low as 30% to be "strong." ECF 1819-1 (T. Goldstein Decl. Ex. A) ¶¶ 260, 288. He now sponsors a coverage threshold of 80% but fails to explain what prompted the revised threshold. But, as admitted in his Reply Report, increasing the coverage threshold beyond

	80% would reduce the number of false positives. ECF 1819-2 (T. Goldstein Decl. Ex. C) ¶ 76. Microsoft defers to OpenAI.
1180. A coverage score between a candidate copy and an Asserted Work of 80% or more and a shared 16-gram indicates a high degree of similarity between the training data record and the Asserted Work. Goldstein Opening Rep. ¶¶ 287, 288.	1180. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 1180 are disputed because the cited evidence is incorrect. Specifically, OpenAI disputes the characterization that an 80% coverage score and a shared 16 gram indicates a "high degree of similarity" between a training data record and the Asserted Work. *First*, the meaning of a "high degree of similarity" is unclear. An 80% coverage score implies 80% of the 3-grams from the Asserted Work are shared. But it implies nothing about the ordering of the 3-grams or about whether those shared 3-grams are attributable to some shared third-party source (for example, significant portions of quoted material). *Second*, OpenAI disputes that an 80% coverage score and a shared 16 gram would reliably detect training data records that share a "high degree of similarity" with an Asserted Work. Dr. Krishnan identified, in Appendix A of his declaration, several examples of pairs of works that meet this proposed threshold but are self-evidently different works. S. Krishnan Opp. Decl. (News) Appendix A. Microsoft defers to OpenAI.

1181. A coverage score between a candidate copy and an Asserted Work of 80% or more and a shared 16-gram indicates that OpenAI copied the Asserted Work substantially in its entirety. Goldstein Opening Rep. ¶¶ 287, 288.	1181. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 1181 are disputed because the cited evidence is incorrect. Specifically, as stated above, OpenAI disputes that a pair of works with an 80% coverage score and a shared 16-gram are necessarily "copies" of each other, and OpenAI incorporates its response to Fact 1180 *supra* and the examples offered in Appendix A of Dr. Krishnan's declaration. S. Krishnan Opp. Decl. (News) Appendix A. Additionally, even if a training data record and an Asserted Work are the same or similar works, that similarity alone does not establish OpenAI "copied" or otherwise obtained the matching text from the Asserted Work. For example, if OpenAI's training data contains a copy of the Constitution, and the Asserted Work also contains text from the Constitution, the presence of the Constitution in OpenAI's training data does not imply OpenAI copied the Constitution from Plaintiffs. Finally, OpenAI disputes the characterization that a record is a copy of an Asserted work "substantially in its entirety" merely because Plaintiffs assert a shared 16-gram and a "coverage score" of 80% or more. Plaintiffs have offered no definition for "substantially in its entirety," and the thresholds proposed can be reached without OpenAI copying an Asserted work "substantially in its entirety," especially where Dr. Goldstein has made no attempt to separate protected text from unprotected text. Microsoft defers to OpenAI.

1182. OpenAI copied, and used during training of at least GPT-3, The Times's Asserted Work titled "As Thousands of Taxi Drivers Were Trapped in Loans, Top Officials Counted the Money," available at https://www.nytimes.com/2019/05/19/nyregion/taxi-medaillions.html. A comparison of the Asserted Work (shown on the left) and a Training Data Sample (shown on the right) that matched to the Asserted Work at a 83.6% coverage score is shown below, with red text applied to spans of matching 4-grams. *See* Times Rule 1006 Summary.

Training Data Match to New York Times Asserted Work

As Thousands of Taxi Drivers Were Trapped in Loans, Top Officials Counted the Money

https://www.nytimes.com/2019/05/19/nyregion/taxi-medallions.html

Coverage: 83.6%
Dataset: cc-filtered-a3-06fzd-langb-2021-06
Model(s): gpt-3

Plaintiff Article

As Thousands of Taxi Drivers Were Trapped in Loans, Top Officials Counted the MoneyBy Brian M. Rosenthal[Read Part 1 of The Times's investigation: How Reckless Loans Devastated a Generation of Taxi Drivers]At a cramped desk on the 22nd floor of a downtown Manhattan office building, Gary Roth spotted a looming disaster.An urban planner with two master's degrees, Mr. Roth had a new job in 2010 analyzing taxi policy for the New York City government. But almost immediately, he noticed something disturbing: The price of a taxi medallion — the permit that lets a driver own a cab — had soared to nearly $700,000 from $200,000. In order to buy medallions, drivers were taking out loans they could not afford.Mr. Roth compiled his concerns in a report, and he and several colleagues warned that if the city did not take action, the loans would become unsustainable and the market could collapse.They were not the only ones worried about taxi medallions. In Albany, state inspectors gave a presentation to top officials showing that medallion owners were not making enough money to support their loans. And in Washington, D.C., federal examiners repeatedly noted that banks were increasing profits by steering cabbies into risky loans.They were all ignored.Medallion prices rose above $1 million before crashing in late 2014, wiping out the futures of thousands of immigrant drivers and creating a crisis that has continued to ravage the industry today. Despite years of warning signs, at least seven government agencies did little to stop the collapse, The New York Times found.Instead, eager to profit off medallions or blinded by the taxi industry's political connections, the agencies that were supposed to police the industry helped a small group of bankers and brokers to reshape it into their own moneymaking machine, according to internal records and interviews with more than 50 former government employees.For more than a decade, the agencies reduced oversight of the taxi trade, exempted it from regulations, subsidized its operations and promoted its practices, records and interviews showed.Their actions turned one of the best-known symbols of New York — its signature yellow cabs — into a financial trap for thousands of immigrant drivers. More than

Training Data Sample

As Thousands of Taxi Drivers Were Trapped in Loans, Top Officials Counted the Money - The New York Times

https://www.nytimes.com/2019/05/19/nyregion/taxi-medallions.html

[Read Part 1 of The Times's investigation: How Reckless Loans Devastated a Generation of Taxi Drivers]

At a cramped desk on the 22nd floor of a downtown Manhattan office building, Gary Roth spotted a looming disaster.

An urban planner with two master's degrees, Mr. Roth had a new job in 2010 analyzing taxi policy for the New York City government. But almost immediately, he noticed something disturbing: The price of a taxi medallion — the permit that lets a driver own a cab — had soared to nearly $700,000 from $200,000. In order to buy medallions, drivers were taking out loans they could not afford.

Mr. Roth compiled his concerns in a report, and he and several colleagues warned that if the city did not take action, the loans would become unsustainable and the market could collapse.

They were not the only ones worried about taxi medallions. In Albany, state inspectors gave a presentation to top officials showing that medallion owners were not making enough money to support their loans. And in Washington, D.C., federal examiners repeatedly noted that banks were increasing profits by steering cabbies into risky loans.

They were all ignored.

Medallion prices rose above $1 million before crashing in late 2014, wiping out the futures of thousands of immigrant drivers and creating a crisis that has continued to ravage the industry today. Despite years of warning signs, at least seven government agencies did little to stop the collapse, The New York Times found.

Instead, eager to profit off medallions or blinded by the taxi industry's political connections, the agencies that were supposed to police the industry helped a small group of bankers and brokers to reshape it into

1182. **Disputed - Not Supported by Cited Evidence**. OpenAI responds as follows: OpenAI does not dispute that Dr. Goldstein has asserted a training-data record with an 83.6% coverage score, but OpenAI disputes that the coverage score or the side-by-side visualizations can be readily reproduced or even assessed using the cited Rule 1006 summary, or the other materials produced by Plaintiffs with their motion (which were not produced during expert discovery). OpenAI also disputes that this coverage score implies "OpenAI copied, and used during training of at least GPT-3, The Times's Asserted Work titled 'As Thousands of Taxi Drivers Were Trapped in Loans, Top Officials Counted the Money,' available at https://www.nytimes.com/2019/05/19/nyregion/taxi-medaillions.html." The facts set forth in Paragraph 1182 are thus disputed because the cited evidence is incorrect.

Microsoft defers to OpenAI.

1183. OpenAI copied, and used during training of at least GPT-3.5, Chicago Tribune's Asserted Work titled "Dissolving Illinois may be answer to state woes," available at https://www.chicagotribune.com/2017/06/20/what-to-do-with-a-broken-illinois-dissolve-the-land-of-lincoln/. A comparison of the Asserted Work (shown on the left) and a Training Data Sample (shown on the right) that matched to the Asserted Work at a 96.8% coverage score is shown below, with red text applied to spans of matching 4-grams. *See* Daily News Plaintiffs' Rule 1006 Summary.



Training Data Match to Chicago Tribune Asserted Work

Dissolving Illinois may be answer to state woes

https://www.chicagotribune.com/2017/06/20/what-to-do-with-a-broken-illinois-dissolve-the-land-of-lincoln/

Coverage: 96.8%
Dataset: webtext2-dedup-clean-csam-filt
Model(s): gpt-3.5

Plaintiff Article

Illinois is like Venezuela now, a fiscally broken state that has lost its will to live, although for the moment, we still have enough toilet paper. But before we run out of the essentials, let's finally admit that after decade upon decade of taxing and spending and borrowing, Illinois has finally run out of other people's money. Those "other people" include taxpayers who've abandoned the state. And now Illinois faces doomsday. So as the politicians meet in Springfield this week for another round of posturing and gesturing and blaming, we need a plan. And here it is:

Dissolve Illinois. Decommission the state, tear up the charter, whatever the legal mumbo-jumbo, just end the whole dang thing. We just disappear. With no pain. That's right. You heard me. The best thing to do is to break Illinois into pieces right now. Just wipe us off the map. Cut us out of America's heartland and let neighboring states carve us up and take the best chunks for themselves. The group that will scream the loudest is the state's political class, who did this to us, and the big bond creditors, who are whispering of bankruptcy and asset forfeiture to save their own skins. But our beloved Illinois has proved that it just doesn't deserve to survive. So why not let our friendly neighbors like Indiana, Wisconsin, Iowa, Missouri and Kentucky just take the parts they want? As you can see by the excellent "Kevorkian Illinois" map that accompanies this column, this plan is visionary. The alternative is hell. Illinois hasn't had a state budget for years. The state continues to spend money it doesn't have, and the state's credit ratings have dropped, increasing the cost of borrowing more money we don't have to keep the rotten shebang going. Bills pile up; Moody's Investor Service says taxpayers are on the hook for $251 billion in unfunded public union pension liabilities. Boss Mike Madigan, king of the Democrats who control things, wants tax increases but no real structural reform to bring stability to The Venezuela of the Midwest.

And the whispers of bankruptcy won't help the average (remaining) taxpaying chumbolones like you and me who don't want to leave our homes but who'll get stuck with the bills. Since our neighboring states are doing better, taking Illinois jobs and businesses and

Training Data Sample

Illinois is like Venezuela now, a fiscally broken state that has lost its will to live, although for the moment, we still have enough toilet paper.

But before we run out of the essentials, let's finally admit that after decade upon decade of taxing and spending and borrowing, Illinois has finally run out of other people's money.

Those "other people" include taxpayers who've abandoned the state. And now Illinois faces doomsday.

So as the politicians meet in Springfield this week for another round of posturing and gesturing and blaming, we need a plan.

And here it is:

Dissolve Illinois. Decommission the state, tear up the charter, whatever the legal mumbo-jumbo, just end the whole dang thing.

We just disappear. With no pain. That's right. You heard me.

The best thing to do is to break Illinois into pieces right now. Just wipe us off the map. Cut us out of America's heartland and let neighboring states carve us up and take the best chunks for themselves.

The group that will scream the loudest is the state's political class, who did this to us, and the big bond creditors, who are whispering talk of bankruptcy and asset forfeiture to save their own skins.

Ryan Marx John Kass's modest proposal for the dissolution of Illinois: Carve up the failed state and let the rest of the Midwest have it. John Kass's modest proposal for the dissolution of Illinois: Carve up the failed state and let the rest of the Midwest have it. (Ryan Marx) (Ryan Marx)

But our beloved Illinois has proved that it just doesn't deserve to survive.

So why not let our friendly neighbors like Indiana, Wisconsin, Iowa, Missouri and Kentucky just take the parts they want?

As you can see by the excellent "Kevorkian Illinois" map that accompanies this column, this plan is visionary.

The alternative is hell. Illinois hasn't had a state budget for years. The state continues to spend money it doesn't have, and the state's credit ratings

1183. **Disputed - Not Supported by Cited Evidence**. OpenAI responds as follows: OpenAI does not dispute that Dr. Goldstein has asserted a training-data record with a 96.8% coverage score, as measured using Dr. Goldstein's methods, but OpenAI disputes that the coverage score or the side-by-side visualizations can be readily reproduced or even assessed using the cited Rule 1006 summary, or the other materials produced by Plaintiffs with their motion (which were not produced during expert discovery). OpenAI also disputes that this coverage score implies "OpenAI copied, and used during training of at least GPT-3.5, Chicago Tribune's Asserted Work titled 'Dissolving Illinois may be answer to state woes,' available at https://www.chicagotribune.com/2017/06/20/what-to-do-with-a-broken-illinois-dissolve-the-land-of-lincoln/." The facts set forth in Paragraph 1183 are thus disputed because the cited evidence is incorrect.

Microsoft defers to OpenAI.

1184. For The Times's Asserted Works, using the 80% coverage score and 16-gram overlap requirements:

Infringing Act	Number of Copies and Unique Asserted Works
OpenAI's scraping of webpages for WebText2	12,714 Copies 6,552 Unique Asserted Works
Defendants' scraping of webpages for Project Mango	28,744 Copies 27,453 Unique Asserted Works
OpenAI's downloading of Common Crawl datasets	3,551,481 Copies 930,825 Unique Asserted Works
OpenAI's distribution of GPT-3 training data to Microsoft	556,185 Copies 365,928 Unique Asserted Works
OpenAI's Giraffe Bloom Filter	140,097 Copies 128,266 Unique Asserted Works
OpenAI's downloading of The Times Annotated Corpus	932,341 Copies 928,872Unique Asserted Works

See Goldstein Opening Rep., Exhibit 11; Goldstein Declaration, Exhibits I and J; Maisel Declaration, Exhibit A (Training Data Infringement Rule 1006 summaries and corresponding side-by-side comparisons, underlying training data records, and asserted works).

1184. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: OpenAI disputes that applying an 80% coverage score and 16-gram overlap requirement reliably measures "copies" within the meaning of copyright infringement. First, Dr. Goldstein's Jaccard coverage and 16-gram overlap without filtering out the non-protectable elements of the Asserted Work, and thus he is measuring raw text similarity rather than similarity with respect to the protectable expression in the Asserted Work. Second, the Asserted Works databases continue to contain works that Plaintiffs do not own and cannot assert in this litigation, and the presence of "matches" using Dr. Goldstein's text similarity measures does not indicate the presence of Plaintiff-owned works. Third, Plaintiffs' cited support does not support this statement. It is not clear on the face of the cited support whether these figures are accurate, reliable measures of the number of "copies" identified by Dr. Goldstein's methodology. These are voluminous productions, and determining whether the produced material supports this fact requires testing and validation. Fourth, Dr. Goldstein's produced evidence does not support the training-based infringement figures he offers. These numbers cannot be replicated. S. Krishnan Opp. Decl. ¶ 50. Fifth, the "copies" Dr. Goldstein has identified using his 80% coverage score and 16-gram overlap requirement nevertheless contain numerous instances where OpenAI's training data records are plainly not "copies" of an Asserted Work. S. Krishnan Opp. Declaration, Appendix A offers

	examples of the false positives that persist in these proposed "copies." Microsoft joins OpenAI's response.
1185. For DNP's Asserted Works, using the 80% coverage score and 16-gram overlap requirements: | Infringing Act	Number of Copies and Unique Asserted Works | |---|---| | OpenAI's scraping of webpages for WebText2 | 37,264 Copies 18,609 Unique Asserted Works | | Defendants' scraping of webpages for Project Mango | 34,169 Copies 32,087 Unique Asserted Works | | OpenAI's downloading of Common Crawl datasets | 6,940,478 Copies 1,368,704 Unique Asserted Works | | OpenAI's distribution of GPT-3 training data to Microsoft | 531,716 Copies 397,150 Unique Asserted Works | | OpenAI's Giraffe Bloom Filter | 270,250 Copies 207,152 Unique Asserted Works | *See* Goldstein Declaration, Exhibits I and J; Maisel Declaration, Exhibit B (Training Data Infringement Rule 1006 summaries and corresponding side-by-side comparisons, underlying training data records, and asserted works).	1185. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: OpenAI disputes that applying Dr. Goldstein's 80% coverage score and 16-gram overlap requirement reliably measures "copies" for the reasons stated in OpenAI's response to Fact 1184. Microsoft joins OpenAI's response.
1186. Using the 80% coverage score and 16-gram overlap requirements, the GPT models were trained on The Times's Asserted Works as follows: | Model	Number of Unique Copies	Number of Total Copies | |---|---|---| | **TOTAL** | 944,655 | 3,924,653 | | **GPT-3** | 365,928 | 556,185 | | **GPT-3.5** | 334,317 | 336,326 | | **GPT-3.5 Turbo** | 708,080 | 931,907 | | **GPT-4** | 580,352 | 1,109,041 | | **GPT-4 Turbo** | 768,754 | 888,481 | | **GPT-4o** | 91,941 | 102,713 | | **GPT-4o-mini** | At least the same as GPT-4o |	1186. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: OpenAI disputes that applying Dr. Goldstein's 80% coverage score and 16-gram overlap requirement reliably measures "copies" for the reasons stated in OpenAI's response to Fact 1184. Microsoft defers to OpenAI.

Goldstein Opening Rep., Exhibit 11; Goldstein Declaration, Exhibits I and J; Maisel Declaration, Exhibit A (Training Data Infringement Rule 1006 summaries and corresponding side-by-side comparisons, underlying training data records, and asserted works).	
1187. Using the 80% coverage score and 16-gram overlap requirements, the GPT models were trained on DNP's Asserted Works as follows: **Model** / **Number of Unique Copies** / **Number of Total Copies** **TOTAL** 1,388,384 7,387,394 **GPT-3** 397,150 531,716 **GPT-3.5** 391,670 393,718 **GPT-3.5 Turbo** 1,201,069 1,605,817 **GPT-4** 1,165,940 2,351,353 **GPT-4 Turbo** 1,286,469 1,466,860 **GPT-4o** 610,878 1,037,930 **GPT-4o-mini** At least the same as GPT-4o Goldstein Opening Rep., Exhibit 11; Goldstein Declaration, Exhibits I and J; Maisel Declaration, Exhibit B (Training Data Infringement Rule 1006 summaries and corresponding side-by-side comparisons, underlying training data records, and asserted works).	1187. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: OpenAI disputes that applying Dr. Goldstein's 80% coverage score and 16-gram overlap requirement reliably measures "copies" for the reasons stated in OpenAI's response to Fact 1184. Microsoft defers to OpenAI.
1188. Using the 80% coverage score and 16-gram overlap requirements, OpenAI's [redacted] used to train [redacted] contains [redacted] unique copies of The Times's Asserted Works. Goldstein Declaration, Exhibit I.	1188. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: OpenAI disputes that applying Dr. Goldstein's 80% coverage score and 16-gram overlap requirement reliably measures "copies" for the reasons stated in OpenAI's response to Fact 1184. Microsoft defers to OpenAI.
1189. Using the 80% coverage score and 16-gram overlap requirements, OpenAI's [redacted]	1189. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: OpenAI disputes

used to train [redacted] contains [redacted] unique copies of DNP's Asserted Works. Goldstein Declaration, Exhibit I.	that applying Dr. Goldstein's 80% coverage score and 16-gram overlap requirement reliably measures "copies" for the reasons stated in OpenAI's response to Fact 1184. Microsoft defers to OpenAI.

2. CIR and The Intercept

1190. To the extent copies of CIR's SJ Asserted Works appear in the Nonprofit News Plaintiffs' Common Crawl Training Sets, OpenAI obtained those copies via Common Crawl. Ex. 814 [OpenAI's Response to CIR's Fourth Set of Requests for Admission (Sept. 11, 2025)] at 6; Ex. 491 [OpenAI's Second Supplemental Response to CIR Interrogatory No. 14 (Mar. 23, 2026)] at 6.	1190. **Undisputed.** Microsoft defers to OpenAI.
1191. To the extent copies of The Intercept's SJ Asserted Works appear in the Nonprofit News Plaintiffs' Common Crawl Training Sets, OpenAI obtained those copies via Common Crawl. Ex. 1154 [OpenAI's Response to The Intercept's Fourth Set of Requests for Admission (Sept. 15, 2025)] at 6; Ex. 1155 [OpenAI's Response to The Intercept's Third Set of Requests for Admission (July 14, 2025)] at 6; Ex. 1156 [OpenAI's Second Supplemental Response to the Intercept's Interrogatory No. 7 (Mar. 23, 2026)] at 6.	1191. **Undisputed.** Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1192. Jaccard similarity is a measure of set similarity, which takes the size of the intersection of two sets over the size of their unions. Wenger Opening Rep. at 42.	1192. **Undisputed.** Microsoft defers to OpenAI.

1193. If two sets are identical, their Jaccard similarity is 1. Wenger Opening Rep. at 42.	1193. **Disputed**. OpenAI responds as follows: OpenAI disputes this statement to the extent it implies that only identical documents have a Jaccard similarity of 1. Jaccard similarity measures compare two sample sets by comparing their shared elements against the union of their distinct elements. Two sets that contain identical elements but in completely different orders would nonetheless have a Jaccard similarity of 1. S. Krishnan Opp. Decl. (News) Ex. H (S. Krishnan Rebuttal Rep. (Wenger)) ¶¶ 42-43. Microsoft defers to OpenAI.
1194. If two sets have no elements in common, their Jaccard similarity is 0. Wenger Opening Rep. at 42.	1194. **Undisputed**. Microsoft defers to OpenAI.
1195. Jaccard similarity over n-grams is a commonly used string similarity metric in natural language processing fields. Wenger Opening Rep. at 42.	1195. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: Disputed to the extent this characterization implies that Jaccard similarity is suitable for measuring textual similarity without regard to the task or methodology. Jaccard measures overlap between sets by dividing the number of shared distinct elements by the number of distinct elements in their union. Its suitability for a particular text comparison depends on the question being asked, the relevant data, and how the texts are processed and represented as sets, including the choice of n-gram size. *See* News Ex. 1167 (Krishnan Rebuttal Rep. (Delorey)) ¶¶ 34(b), 34(d), 34(f), 46, 91–93; S. Krishnan Opp. Decl.

	(News) Ex. H (S. Krishnan Rebuttal Rep. (Wenger)) ¶¶ 34(b), 34(d), 34(f), 46, 91–93. In particular, use of unigrams is inconsistent with established practice in information retrieval and easily allows for false positives without proper validation of the results. Ex. 1167 (S. Krishnan Rebuttal Rep. (Delorey)) ¶ 91; S. Krishnan Opp. Decl. (News) Ex. H (S. Krishnan Rebuttal Rep. (Wenger)) ¶ 91. Microsoft defers to OpenAI.
1196. OpenAI has used the Jaccard metric to compare the similarity of documents within its training sets. Ex. 521 [Yuan 30(b)(1) Depo. Tr.] at 62:24-63:6; Ex. 633 [Paino Jan. 22, 2026 30(b)(6) Depo. Tr.] at 110:12-111:24.	1196. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that OpenAI has used Jaccard similarity in document deduplication. However, disputed as not supported by the evidence. Specifically, the cited testimony describes approximate Jaccard similarity as part of a deduplication algorithm involving tuned parameters and thresholds. Disputed that this evidence establishes that OpenAI used the same text preprocessing, n-gram size, or matching criteria as Dr. Wenger, or that OpenAI's use validates her methodology for identifying copies of Plaintiffs' Asserted Works. *See* News Ex. 521 (Q. Yuan 30(b)(1) Tr.) at 62:12–63:24; News Ex. 633 (A. Paino 30(b)(6) Tr.) at 110:1–111:24. Microsoft defers to OpenAI.
1197. OpenAI has treated documents [redacted] Ex. 521	1197. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows:

[Yuan Nov. 18, 2025 Depo. Tr.] at 68:18-69:13; Ex. 633 [Paino Jan. 22, 2026 30(b)(6) Depo. Tr.] at 110:12-111:24.	Undisputed that OpenAI has used Jaccard similarity scores in its fuzzy deduplication process. However, disputed that the cited evidence establishes that OpenAI used the same process that Dr. Wenger describes in her expert report, or that Dr. Wenger's methodology for identifying copies of Plaintiffs' works is valid. Specifically, the cited evidence does not establish that OpenAI used the same text preprocessing, n-gram size, or matching criteria as Dr. Wenger, or that OpenAI's use validates her methodology for identifying copies of Plaintiffs' asserted works. Indeed, the cited testimony merely states a "plausible" Jaccard threshold for some datasets, it does not make clear what specific Jaccard score was used for OpenAI's distinct deduplication effort. News Ex. 633 (A. Paino 30(b)(6) Tr.) at 110:1–111:24; News Ex. 521 (Q. Yuan 30(b)(1) Tr.) at 62:12–63:24. Microsoft defers to OpenAI.
1198. The Jaccard similarity metric can be calculated over different n-gram sizes. Wenger Opening Rep. at 42-43.	1198. **Undisputed.** Microsoft defers to OpenAI.
1199. The Jaccard similarity over 3-grams between the following two examples is 0.639, where the "ORIGINAL ARTICLE" is the text from an Asserted Work owned by The Intercept, and the "CLIENT ARTICLE" is text originating from OpenAI's training set [redacted] with overlap in black and differences in red.	1199. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 1199 are disputed because they are not supported by the cited evidence. Specifically, OpenAI does not dispute that Plaintiffs assert the 3-gram Jaccard similarity is 0.639 between the two works, but Plaintiffs have not provided sufficient evidence to confirm any

Wenger Opening Rep. at 48; Ex. 1157 [Wenger Opening Rep. Appendix 29] at 189-91. ORIGINAL ARTICLE: trumps strike in syria white house press secretary sean spicer told reporters on tuesday that not even adolf hitler used chemical weapons during the world war ii you had a someone as despicable as hitler who didnt even sink to using chemical weapons so you have to if youre russia ask yourself is this is a country and a regime that you want to align yourself with when asked for clarification spicer said that he meant dropping chemical weapons on unsuspecting towns of germans when it comes to sarin gas he was not using the gas on his own people the same way that ashad sic was doing he brought them into the holocaust center i understand that in the way that assad used them where he went to towns dropped them down to innocents into the middle of towns i appreciate the clarification there that was not the intent after the news conference spicer issued yet another statement this one in writing according to the american chemical society hitler stockpiled large amounts of sarin gas during the second world war but never used them on a european battlefield but it is common knowledge that poison gas was his method of choice for the mass murder of jews lgbt people and others during the holocaust more than a million people were killed by the poison gas zyklon b a derivative of cyanide at the auschwitz camp alone the united states holocaust memorial museum estimates that hitlers overall death toll includes as many as 6 million jews over 7 million nonjewish soviet citizens nearly 2 million nonjewish polish civilians people with disabilities gypsies and an undetermined number of political prisoners and lgbt people some people on twitter noted the significance of the datenannannannannan CLIENT ARTICLE: spicer when it comes to sarin gas he was not using the gas on his own people the same way that ashad sic was doing he brought them into the holocaust center while defending donald trump s strike in syria white house press secretary sean spicer told reporters on tuesday that not even adolf hitler used chemical weapons during the world war ii you had a someone as despicable as hitler who didn t even sink to using chemical weapons so you have to if you re russia ask yourself is this is a country and a regime that you want to align yourself with when asked for clarification spicer said that he meant dropping chemical weapons on unsuspecting towns of germans when it comes to sarin gas he was not using the gas on his own people the same way that ashad sic was doing he brought them into the holocaust center i understand that spicer continued in the way that assad used them where he went to towns dropped them down to innocents into the middle of towns i appreciate the clarification there that was not the intent after the news conference spicer issued yet another statement this one in writing new a second clarification from the presssec on assadhitler comparison pictwittercomiu8oa5jfab according to the american chemical society hitler stockpiled large amounts of sarin gas during the second world war but never used them on a european battlefield but it is common knowledge that poison gas was his method of choice for the mass murder of jews lgbt people and others during the holocaust more than a million people were killed by the poison gas zyklon b a derivative of cyanide at the auschwitz camp alone the united states holocaust memorial museum estimates that hitler s overall death toll includes as many as 6	calculation of the Jaccard similarity of the two works or to confirm that the 3-gram Jaccard similarity is 0.639. *See, e.g.*, ECF 1858-1 (E. Wenger Decl. Ex. A) at 48; News Ex. 1157 (E. Wenger Opening Rep. App. 29) at 189–91. OpenAI further disputes that the "overlap [is] in black and differences in red," as the text contains significant differences that are not "in red," and thus it is not clear how overlap or differences were determined for the purposes of color coding. Moreover, from the face of the asserted fact, it appears that the text labeled ORIGINAL ARTICLE is not an Asserted Work owned by The Intercept but instead has been edited or modified in some fashion. Also, from the face of the asserted fact, it appears that the text labeled CLIENT ARTICLE has also been edited or manipulated. Thus, OpenAI disputes that the ORIGINAL ARTICLE corresponds to an Asserted Work owned by The Intercept, and that the CLIENT ARTICLE is text originating from OpenAI's training set. Microsoft defers to OpenAI.

million jews over 7 million nonjewish soviet citizens nearly 2 million nonjewish polish civilians people with disabilities gypsies and an undetermined number of political prisoners and lgbt people hey presssec the gas that killed 6 million jews was called zyklon b very definitely a chemical weapon httpstcoiz4ocifyxe some people on twitter noted the significance of the date wait before you go become a member college democrat at center of attack on alex morse hoped to launch career through richard neal rankandfile umass college democrats were stunned to see the letter their chapter s leadership wrote on their behalf the race to replace joe kennedy in congress is heating up ihssane leckey a democratic socialist and former bank regulator is surging in the open field trump is prepared to do whatever it takes to keep power can he be stoppedn	
1200. The Jaccard similarity over 3-grams between the following two examples is 0.608, where the "ORIGINAL ARTICLE" is the text from an Asserted Work owned by CIR, and the "CLIENT ARTICLE" is text originating from OpenAI's training set [redacted] with overlap in black and differences in red. Wenger Opening Rep. at 48; Ex. 1158 [Wenger Opening Rep. Appendix 31] at 49-50. ORIGINAL ARTICLE: no one expected the senate race in dependablyred texas to be this close incumbent republican sen ted cruz according CLIENT ARTICLE: right now is no time to come up short 1150000 1350000 donate donate no one expected the senate race in dependablyred	1200. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 1200 are disputed because they are not supported by the cited evidence. Specifically, OpenAI does not dispute that Plaintiffs assert the 3-gram Jaccard similarity is 0.608 between the two works, but Plaintiffs have not provided sufficient evidence to confirm any calculation of the Jaccard similarity of the two works or to confirm that the 3-gram Jaccard similarity is 0.608. *See e.g.* ECF 1858-1 (E. Wenger Decl. Ex. A) at 48; News Ex. 1158 (E. Wenger Opening Rep. App. 31) at 49–50. OpenAI further disputes that the "overlap [is] in black and differences in red," as the text contains significant differences that are not "in red," and thus it is not clear how overlap or differences were determined for the purposes of

to recent polling leads democrat and threeterm us congressman beto orourke by just a few pointsdespite cruzs attacks on orourkes nickname his given name is robert and time in a punk rock band on wednesday night orourke went on the late show with stephen colbert for a conversation heavy on immigration in stark contrast to his opponent orourke said he opposed trumps planned wall along the mexican border he said that democrats shouldnt agree to a wall in exchange for relief for dreamers and that dacarecipients should be made citizens you can watch the full clip below httpsyoutubej5ies0up9sc rep orourke currently represents texas 16th congressional district in congress which includes the border town of el paso where orourke grew up and now lives in a visit to el paso last year sen cruz expressed his support for trumps proposed wall during a news conference held steps away from mexico after firing a pepper ball stun gun and riding a border patrol horse texas to be this close incumbent republican sen ted cruz according to recent polling leads democrat and threeterm us congressman beto o rourke by just a few points despite cruz s attacks on o rourke s nickname his given name is robert and time in a punk rock band on wednesday night o rourke went on the late show with stephen colbert for a conversation heavy on immigration in stark contrast to his opponent o rourke said he opposed trump s planned wall along the mexican border he said that democrats shouldn t agree to a wall in exchange for relief for dreamers and that dacarecipients should be made citizens you can watch the full clip below rep o rourke currently represents texas 16th congressional district in congress which includes the border town of el paso where o rourke grew up and now lives in a visit to el paso last year sen cruz expressed his support for trump s proposed wall during a news conference held steps away from mexico after firing a pepper ball stun gun and riding a border patrol horse thank you for subscribingn	color coding. Moreover, from the face of the asserted fact, it appears that the text labeled ORIGINAL ARTICLE is not an Asserted Work owned by CIR but instead has been edited or modified in some fashion. Also from the face of the asserted fact, it appears that the text labeled CLIENT ARTICLE has also been edited or manipulated. Thus, OpenAI disputes that the ORIGINAL ARTICLE corresponds to an Asserted Work owned by CIR, and that the CLIENT ARTICLE is text originating from OpenAI's training set. Microsoft defers to OpenAI.
1201. Appendices 29, 31, and 33 to Dr. Wenger's Opening Report accurately state the Jaccard similarity scores over 3-grams between further examples of CIR's and The Intercept's Asserted Works and text from OpenAI's Nonprofit News Plaintiffs' Common Crawl Training Sets, and depict the similarities and differences between those text strings, with overlap in black and differences in red. Wenger Opening Rep. at 48; Ex. 1157 [Wenger Opening Rep. Appendix 29]; Ex. 1158 [Wenger Opening Rep. Appendix 31]; Ex. 1159 [Wenger Opening Rep. Appendix 33].	1201. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 1201 are disputed because they are not supported by the cited evidence. Specifically, OpenAI disputes that the "overlap [is] in black and differences in red," as the text contains significant differences that are not "in red," and thus it is not clear how overlap or differences were determined for the purposes of color coding. *See* ECF 1858-1 (E. Wenger Decl. Ex. A) at 48; News Ex. 1157 (E. Wenger Opening Rep. App. 29); News Ex. 1158 (E. Wenger Opening Rep. App. 31); News Ex. 1159 (E. Wenger Opening Rep. App. 33). Moreover, the "further examples of CIR's

	and The Intercept's Asserted Works" appear to contain revisions or modifications to the text, and it also appears that the text from OpenAI's training records has been modified. OpenAI therefore disputes that the similarities and differences between those text strings reflect similarities and differences between CIR and Intercept works and OpenAI training data records. Finally, OpenAI disputes that asking OpenAI to either dispute or accede to the accuracy of every row in several voluminous appendices is a proper subject of a Rule 56.1 statement. Microsoft defers to OpenAI. Microsoft responds that facts related solely to The Intercept do not pertain to claims asserted against Microsoft.
1202. Appendix 176 to Dr. Wenger's Supplemental Report accurately states, for each of CIR's Mother Jones Registered Asserted Works, and for each of OpenAI's Nonprofit News Plaintiffs' Common Crawl Training Sets excerpts of which were produced to CIR, the number of instances that work's URL appeared in that training set. Wenger Supplemental Rep. at 5; Wenger Opening Rep. at 34-35; Ex. 669 [Wenger Supplemental Rep. Appendix 176]; Ex. 963 [OPCO_NEWS_1144089]; Ex. 965 [OPCO_NEWS_1144090]; Ex. 967 [OPCO_NEWS_1144091]; Ex. 969 [OPCO_NEWS_1144092]; Ex. 971 [OPCO_NEWS_1144093]; Ex. 973 [OPCO_NEWS_1144094]; Ex. 975	1202. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 1202 are disputed because the accuracy of Appendix 176 is not apparent from the face of Plaintiffs' cited support. Whether the dataset on which Dr. Wenger relies accurately identifies the Nonprofit News Plaintiffs' Asserted Works—and whether Dr. Wenger accurately identified URLs associated with those Works in OpenAI's Common Crawl-derived training datasets—depends on questions of copyright ownership and technical methodology that cannot be resolved by reviewing the cited materials alone. Finally, OpenAI disputes that asking OpenAI to either dispute or accede to the accuracy of every

OPCO_NEWS_1144095]; Ex. 977 [OPCO_NEWS_1144096]; Ex. 979 9OPCO_NEWS_1144097]; Ex. 981 [OPCO_NEWS_1144098]; Ex. 983 [OPCO_NEWS_1144099]; Bauerlein Decl. Ex. D; Bauerlein Decl. Ex. E.	row in several voluminous appendices as a proper subject of a Rule 56.1 statement. Microsoft defers to OpenAI.
1203. OpenAI copied at least 743 unique Mother Jones SJ Registered Asserted Works by downloading them from Common Crawl. Wenger Supplemental Rep. at 5, 6; Ex. 669 [Wenger Supplemental Rep. Appendix 176]; Ex. 963 [OPCO_NEWS_1144089]; Ex. 965 [OPCO_NEWS_1144090]; Ex. 967 [OPCO_NEWS_1144091]; Ex. 969 [OPCO_NEWS_1144092]; Ex. 971 [OPCO_NEWS_1144093]; Ex. 973 [OPCO_NEWS_1144094]; Ex. 975 [OPCO_NEWS_1144095]; Ex. 977 [OPCO_NEWS_1144096]; Ex. 979 [OPCO_NEWS_1144097]; Ex. 981 [OPCO_NEWS_1144098]; Ex. 983 [OPCO_NEWS_1144099]; Bauerlein Decl. Ex. D; Bauerlein Decl. Ex. E; Ex. 814 [OpenAI's Response to CIR's Fourth Set of Requests for Admission (Sept. 11, 2025)] at 6; Ex. 491 [OpenAI's Second Supplemental Response to CIR Interrogatory No. 14 (Mar. 23, 2026)] at 6; Bauerlein Decl. ¶ 17 n.1.	1203. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 1203 are disputed because they are not supported by the cited evidence. Specifically, disputed to the extent Plaintiffs equate Dr. Wenger's URL-based count with proof that OpenAI downloaded copies of each of the 743 specific Mother Jones SJ Registered Asserted Works. Dr. Wenger's unfiltered count identifies records associated with asserted-work URLs; it does not itself establish that the downloaded text reproduces the corresponding registered work. Pursuant to Fed. R. Evid. 106, as Dr. Krishnan explains, matching URLs do not establish matching content: webpages can change over time, and records associated with the same URL may contain materially different text, summaries, or other content. S. Krishnan Opp. Decl. (News) Ex. H (S. Krishnan Rebuttal Rep. (Wenger)) ¶¶ 71–74. Further, OpenAI's admission that it obtained Common Crawl web-crawl data does not admit the asserted 743-work count or the identity of each downloaded record with the corresponding registered work. News Ex. 814 (OpenAI's Resps. to CIR's Fourth Set of Requests for

	Admission (Sept. 11, 2025)) at 6 (Resp. to RFA 16). In addition, the cited evidence does not support the claimed "743" figure. Dr. Wenger's Supplemental Report asserts "747" unique URLs (not works). ECF 1858-8 (E. Wenger Decl. Ex. H) at 6. Many of her 743 "copies" fail to meet the thresholds she set for identifying "copies." S. Krishnan Decl.) ¶ 85. And the "copies" Dr. Wenger has identified contain numerous instances where OpenAI's training data records are plainly not "copies" of an Asserted Work. *See id*. App. B (offering examples of the false positives that persist in these proposed "copies"). Microsoft defers to OpenAI.
1204. For CIR and The Intercept, the term "SJ Accused Copies" refers to copies of their SJ Asserted Works in a Nonprofit News Plaintiffs' Common Crawl Training Set (other than copies located at the Mother Jones subdomains referenced at page 3, footnote 3 of Dr. Wenger's Supplemental Report), where that copy that has a Jaccard similarity score over 3-grams with the original of at least 0.6.	1204. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 1204 are disputed because they are not supported by the cited evidence. Specifically, OpenAI understands "SJ Accused Copies" as Plaintiffs' shorthand for the training data records they accuse, subject to the stated dataset, subdomain, and trigram Jaccard filters. OpenAI disputes that those records necessarily constitute copies of the corresponding "Asserted Works." A numerical similarity threshold does not itself establish that a record reproduces the corresponding work. Pursuant to Fed. R. Evid. 106, Dr. Krishnan explains that Jaccard similarity alone is insufficient to reliably determine whether records are duplicates or near-duplicates and that additional

	validation is required. S. Krishnan Opp. Decl. (News) Ex. H (S. Krishnan Rebuttal Rep. (Wenger)) ¶¶ 115–118. Microsoft defers to OpenAI. Microsoft further responds that facts related solely to The Intercept do not pertain to claims asserted against Microsoft.
1205. Appendices 190-193 to Dr. Wenger's Supplemental Report accurately state, for each of The Intercept's and CIR's SJ Asserted Works, and for each of OpenAI's Nonprofit News Plaintiffs' Common Crawl Training Sets excerpts of which were produced to The Intercept and CIR, the number of SJ Accused Copies contained in that training set. Wenger Supplemental Rep. at 22; Ex. 1272 [Wenger Supplemental Rep. Appendix 190]; Ex. 1160 [Wenger Supplemental Rep. Appendix 191]; Ex. 1161 [Wenger Supplemental Rep. Appendix 192]; Ex. 1162 [Wenger Supplemental Rep. Appendix 193]; Ex. 985 [OPCO_INTERCEPT_0000001]; Ex. 987 [OPCO_INTERCEPT_0000002]; Ex. 989 [OPCO_INTERCEPT_0000003]; Ex. 991 [OPCO_INTERCEPT_0000004]; Ex. 993 [OPCO_INTERCEPT_0000005]; Ex. 995 [OPCO_INTERCEPT_0000006]; Ex. 997 [OPCO_INTERCEPT_0000007]; Ex. 999 [OPCO_INTERCEPT_0000008]; Ex. 1001 [OPCO_INTERCEPT_0000009]; Ex. 1003 [OPCO_INTERCEPT_0000010]; Ex. 1005 [OPCO_INTERCEPT_0000011]; Macala Decl. Ex. B	1205. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 1205 are disputed because they are not supported by the cited evidence. Specifically, the accuracy of Appendices 190–193 is not apparent from the face of Plaintiffs' cited support. OpenAI also disputes that Dr. Wenger has reliably and accurately identified "copies," or that the term "SJ Accused Copies" actually reflects "copies" of Plaintiffs' Asserted Works, for the reasons stated in its response to Paragraph 1203. Finally, OpenAI disputes that asking OpenAI to either dispute or accede to the accuracy of every row in several voluminous appendices is a proper subject of a Rule 56.1 statement. Microsoft defers to OpenAI. Microsoft further responds that facts related solely to The Intercept do not pertain to claims asserted against Microsoft.

Macala Decl. Ex. C; Ex. 963 [OPCO_NEWS_1144089]; Ex. 965 [OPCO_NEWS_1144090]; Ex. 967 [OPCO_NEWS_1144091]; Ex. 969 [OPCO_NEWS_1144092]; Ex. 971 [OPCO_NEWS_1144093]; Ex. 973 [OPCO_NEWS_1144094]; Ex. 975 [OPCO_NEWS_1144095]; Ex. 977 [OPCO_NEWS_1144096]; Ex. 979 [OPCO_NEWS_1144097]; Ex. 981 [OPCO_NEWS_1144098]; Ex. 983 [OPCO_NEWS_1144099]; Bauerlein Decl. Ex. D; Bauerlein Decl. Ex. E; Bauerlein Decl. Ex. F; Bauerlein Decl. Ex. G.	
1206. Appendices 194-197 to Dr. Wenger's Supplemental Report accurately state, for each of The Intercept's and CIR's SJ Asserted Works, and for each of OpenAI's Nonprofit News Plaintiffs' Common Crawl Training Sets excerpts of which were produced to The Intercept and CIR the Jaccard similarity score over 3-grams between each SJ Accused Copy and the original. Wenger Supplemental Rep. at 22; Ex. 1163 [Wenger Supplemental Rep. Appendix 194]; Ex. 1164 [Wenger Supplemental Rep. Appendix 195]; Ex. 1165 [Wenger Supplemental Rep. Appendix 196]; Ex. 1166 [Wenger Supplemental Rep. Appendix 197] Ex. 985 [OPCO_INTERCEPT_0000001]; Ex. 987 [OPCO_INTERCEPT_0000002]; Ex. 989 [OPCO_INTERCEPT_0000003]; Ex. 991 [OPCO_INTERCEPT_0000004]; Ex. 993 [OPCO_INTERCEPT_0000005]; Ex. 995	1206. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 1206 are disputed because they are not supported by the cited evidence. Specifically, OpenAI disputes the definition of "SJ Accused Copy" because OpenAI disputes that Dr. Wenger has reliably identified "copies" of Plaintiffs' Asserted Works, for the reasons stated in its response to Paragraph 1203. Additionally, although OpenAI does not dispute that Appendices 194–197 purport to set forth Jaccard similarities between two sets, the "accuracy" of the Jaccard similarity scores is not readily verifiable on the face of Plaintiffs' cited support. Finally, OpenAI disputes that asking OpenAI to either dispute or accede to the accuracy of every row in several voluminous appendices is a proper subject of a Rule 56.1 statement.

[OPCO_INTERCEPT_0000006]; Ex. 997 [OPCO_INTERCEPT_0000007]; Ex. 999 [OPCO_INTERCEPT_0000008]; Ex. 1001 [OPCO_INTERCEPT_0000009]; Ex. 1003 [OPCO_INTERCEPT_0000010]; Ex. 1005 [OPCO_INTERCEPT_0000011]; Macala Decl. Ex. B; Macala Decl. Ex. C; Ex. 963[OPCO_NEWS_1144089]; Ex. 965 [OPCO_NEWS_1144090]; Ex. 967 [OPCO_NEWS_1144091]; Ex. 969 [OPCO_NEWS_1144092]; Ex. 971 [OPCO_NEWS_1144093]; Ex. 973 [OPCO_NEWS_1144094]; Ex. 975 [OPCO_NEWS_1144095]; Ex. 977 [OPCO_NEWS_1144096]; Ex. 979 [OPCO_NEWS_1144097]; Ex. 981 [OPCO_NEWS_1144098]; Ex. 983 [OPCO_NEWS_1144099]; Bauerlein Decl. Ex. D; Bauerlein Decl. Ex. E; Bauerlein Decl. Ex. F; Bauerlein Decl. Ex. G.	Microsoft defers to OpenAI. Microsoft further responds that facts related solely to The Intercept do not pertain to claims asserted against Microsoft.
1207. OpenAI copied into its Nonprofit News Plaintiffs' Common Crawl Training Sets at least 7,064 unique Intercept SJ Asserted Works, where at least one copy of each such work has a Jaccard similarity score over 3-grams with the original of at least 0.6. Wenger Supplemental Rep. at 22, 23; Ex. 1272 [Wenger Supplemental Rep. Appendix 190]; Ex. 1163 [Wenger Supplemental Rep. Appendix 194]; Ex. 985 [OPCO_INTERCEPT_0000001]; Ex. 987 [OPCO_INTERCEPT_0000002]; Ex. 989	1207. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: Disputed as unsupported by the cited evidence. OpenAI disputes that applying Dr. Wenger's methodology reliably measures "copies" within the meaning of copyright infringement for the reasons stated in response to Fact 1203. OpenAI disputes that the existence of a match establishes how OpenAI obtained a particular record, or whether that record was obtained by "copying" an Asserted Work. OpenAI further disputes that Dr. Wenger has

[OPCO_INTERCEPT_0000003]; Ex. 991 [OPCO_INTERCEPT_0000004]; Ex. 993 [OPCO_INTERCEPT_0000005]; Ex. 995 [OPCO_INTERCEPT_0000006]; Ex. 997 [OPCO_INTERCEPT_0000007]; Ex. 999 [OPCO_INTERCEPT_0000008]; Ex. 1001 [OPCO_INTERCEPT_0000009]; Ex. 1003 [OPCO_INTERCEPT_0000010]; Ex. 1005 [OPCO_INTERCEPT_0000011]; Macala Decl. Ex. B; Macala Decl. Ex. C.	identified 7,064 unique copies because that figure cannot be reconciled fully with Intercept's Final Rule 1006 Summary. S. Krishnan Opp. Decl. ¶ 80. Additionally, many of the 7,064 "copies" are plainly not "copies" of an Asserted Work. S. Krishnan Opp. Decl., Appendix B offers examples of the false positives that persist in these proposed "copies." Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1208. OpenAI copied into its Nonprofit News Plaintiffs' Common Crawl Training Sets at least 682 unique Mother Jones SJ Registered Asserted Works, where at least one copy of each such work has a Jaccard similarity score over 3-grams with the original of at least 0.6. Wenger Supplemental Rep. at 22, 23; Ex. 1161 [Wenger Supplemental Rep. Appendix 192]; Ex. 1165 [Wenger Supplemental Rep. Appendix 196]; Ex. 963[OPCO_NEWS_1144089]; Ex. 965 [OPCO_NEWS_1144090]; Ex. 967 [OPCO_NEWS_1144091]; Ex. 969 [OPCO_NEWS_1144092]; Ex. 971 [OPCO_NEWS_1144093]; Ex. 973 [OPCO_NEWS_1144094]; Ex. 975 [OPCO_NEWS_1144095]; Ex. 977 [OPCO_NEWS_1144096]; Ex. 979 [OPCO_NEWS_1144097]; Ex. 981 [OPCO_NEWS_1144098]; Ex. 983	1208. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 1208 are disputed because they are not supported by the cited evidence. Specifically, disputed to the extent Plaintiffs equate Dr. Wenger's identification of qualifying matching records with proof that OpenAI copied each of the 682 corresponding Mother Jones SJ Registered Asserted Works. The figure purportedly counts unique asserted works having at least one record satisfying Dr. Wenger's filters; it does not independently establish that each record reproduces the corresponding registered work. Pursuant to Fed. R. Evid. 106, Jaccard similarity measures overlapping n-grams but does not establish the relationship between the texts or whether the overlap consists of Plaintiffs' authored expression rather than material from another source. S. Krishnan Opp. Decl. (News) Ex. H (S. Krishnan Rebuttal Rep. (Wenger)) ¶¶ 83–84. Nor does applying a numerical cutoff alone validate the

[OPCO_NEWS_1144099]; Bauerlein Decl. Ex. D; Bauerlein Decl. Ex. E; Bauerlein Decl. ¶ 17 n.1.	matches: Dr. Krishnan explains that false positives occur at different Jaccard scores and that additional validation is required. *Id.* ¶¶ 115–118. OpenAI therefore disputes the inference of copying drawn from the reported count and similarity threshold. Further disputed because the cited evidence does not support the asserted "682" figure. ECF 1858-8 (E. Wenger Supplemental Rep.) at 23 (reporting "686" URLs, not 682). Microsoft defers to OpenAI.
1209. OpenAI copied into its Nonprofit News Plaintiffs' Common Crawl Training Sets at least 37,360 unique CIR SJ Asserted Works, where at least one copy of each such work has a Jaccard similarity score over 3-grams with the original of at least 0.6. Wenger Supplemental Rep. at 22, 23; Ex. 1160 [Wenger Supplemental Rep. Appendix 191]; Ex. 1162 [Wenger Supplemental Rep. Appendix 193]; Ex. 1164 [Wenger Supplemental Rep. Appendix 195]; Ex. 1166 [Wenger Supplemental Rep. Appendix 197]; Ex. 963 [OPCO_NEWS_1144089]; Ex. 965 [OPCO_NEWS_1144090]; Ex. 967 [OPCO_NEWS_1144091]; Ex. 969 [OPCO_NEWS_1144092]; Ex. 971 [OPCO_NEWS_1144093]; Ex. 973 [OPCO_NEWS_1144094]; Ex. 975 [OPCO_NEWS_1144095]; Ex. 977 [OPCO_NEWS_1144096]; Ex. 979 [OPCO_NEWS_1144097]; Ex. 981	1209. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: Disputed as unsupported by the cited evidence. Specifically, OpenAI disputes that applying Dr. Wenger's methodology reliably measures "copies" within the meaning of copyright infringement for the reasons stated in response to Fact 1203. OpenAI disputes that the existence of a match establishes how OpenAI obtained a particular record, or whether that record was obtained by "copying" an Asserted Work. OpenAI further disputes that Dr. Wenger has identified 37,360 unique copies because that figure cannot be reproduced from her working papers, and many of the 37,360 "copies" are plainly not "copies" of an Asserted Work, as discussed in OpenAI's response to Fact 1207. Microsoft defers to OpenAI.

[OPCO_NEWS_1144098]; Ex. 983 [OPCO_NEWS_1144099]; Bauerlein Decl. Ex. D; Bauerlein Decl. Ex. E; Bauerlein Decl. Ex. F; Bauerlein Decl. Ex. G.	
1210. The following OpenAI models were trained on at least the following number of copies of Intercept SJ Asserted Works with a Jaccard similarity over 3-grams of at least 0.6 with the originals: | Model | Number of Copies | | --- | --- | | GPT-3.5 | 902 | | GPT-3.5 Turbo | 1,176 | | GPT-4 | 1,218 | | GPT-4 Turbo | 4,977 | | GPT-4o | 5,080 | Wenger Supplemental Rep. at 22, 27; Ex. 1272 [Wenger Supplemental Rep. Appendix 190]; Ex. 1163 [Wenger Supplemental Rep. Appendix 194]; Ex. 985 [OPCO_INTERCEPT_0000001]; Ex. 987 [OPCO_INTERCEPT_0000002]; Ex. 989 [OPCO_INTERCEPT_0000003]; Ex. 991 [OPCO_INTERCEPT_0000004]; Ex. 993 [OPCO_INTERCEPT_0000005]; Ex. 995 [OPCO_INTERCEPT_0000006]; Ex. 997 [OPCO_INTERCEPT_0000007]; Ex. 999 [OPCO_INTERCEPT_0000008]; Ex. 1001 [OPCO_INTERCEPT_0000009]; Ex. 1003 [OPCO_INTERCEPT_0000010]; Ex. 1005 [OPCO_INTERCEPT_0000011]; Ex. 490 [OpenAI's Second Supplemental Response to the Intercept's Interrogatory No. 7 (Mar. 23, 2026)] at 6; Macala Decl. Ex. B; Macala Decl. Ex. C.	1210. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: Disputed as unsupported by the cited evidence. OpenAI disputes Dr. Wenger's estimates of the number of copies for the reasons stated in OpenAI's responses to Facts 1202, 1207, and 1209. Specifically, Dr. Wenger's methodology does not reliably measure "copies" within the meaning of copyright infringement for the reasons stated in response to Fact 1202. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.

1211. The following OpenAI models were trained on at least the following number of copies of Mother Jones Registered SJ Asserted Works with a Jaccard similarity over 3-grams of at least 0.6 with the originals: Wenger Supplemental Rep. at 22, 27; Ex. 1161 [Wenger Supplemental Rep. Appendix 192], Lines, 26, 448, 455, 579; Ex. 1165 [Wenger Supplemental Rep. Appendix 196], Lines 26, 448, 455, 579; Ex. 963 [OPCO_NEWS_1144089]; Ex. 965 [OPCO_NEWS_1144090]; Ex. 967 [OPCO_NEWS_1144091]; Ex. 969 [OPCO_NEWS_1144092]; Ex. 971 [OPCO_NEWS_1144093]; Ex. 973 [OPCO_NEWS_1144094]; Ex. 975 [OPCO_NEWS_1144095]; Ex. 977 [OPCO_NEWS_1144096]; Ex. 979 [OPCO_NEWS_1144097]; Ex. 981 [OPCO_NEWS_1144098]; Ex. 983 [OPCO_NEWS_1144099]; Ex. 491 [OpenAI's Second Supplemental Response to CIR Interrogatory No. 14 (Mar. 23, 2026)] at 6; Bauerlein Decl. Ex. D; Bauerlein Decl. Ex. E.	1211. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: Disputed as unsupported by the cited evidence. OpenAI disputes these estimates of the number of copies for the reasons stated in OpenAI's response to facts 1202, 1207, and 1209. Specifically, Dr. Wenger's methodology does not reliably measure "copies" within the meaning of copyright infringement for the reasons stated in response to Fact 1202. Microsoft defers to OpenAI.
1212. The following OpenAI models were trained on at least the following number of copies of CIR SJ Asserted Works with a Jaccard similarity over 3-grams of at least 0.6 with the originals:	1212. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: Disputed as unsupported by the cited evidence. OpenAI disputes these estimates of number of copies for the reasons stated in

Model	Number of Copies
GPT-3.5	146
GPT-3.5 Turbo	278
GPT-4	281
GPT-4 Turbo	661
GPT-4o	291

Model	Number of Copies
GPT-3.5	8,797
GPT-3.5 Turbo	28,629
GPT-4	29,328
GPT-4 Turbo	20,260
GPT-4o	14,078

Wenger Supplemental Rep. at 22, 27; Ex. 1160 [Wenger Supplemental Rep. Appendix 191]; Ex. 1162 [Wenger Supplemental Rep. Appendix 193]; Ex. 1164 [Wenger Supplemental Rep. Appendix 195]; Ex. 1166 [Wenger Supplemental Rep. Appendix 197]; Ex. 963 [OPCO_NEWS_1144089]; Ex. 965 [OPCO_NEWS_1144090]; Ex. 967 [OPCO_NEWS_1144091]; Ex. 969 [OPCO_NEWS_1144092]; Ex. 971 [OPCO_NEWS_1144093]; Ex. 973 [OPCO_NEWS_1144094]; Ex. 975 [OPCO_NEWS_1144095]; Ex. 977 [OPCO_NEWS_1144096]; Ex. 979 [OPCO_NEWS_1144097]; Ex. 981 [OPCO_NEWS_1144098]; Ex. 983 [OPCO_NEWS_1144099]; Ex. 491 [OpenAI's Second Supplemental Response to CIR Interrogatory No. 14 (Mar. 23, 2026)] at 6; Bauerlein Decl. Ex. D; Bauerlein Decl. Ex. E; Bauerlein Decl. Ex. F; Bauerlein Decl. Ex. G.

OpenAI's response to facts 1202, 1207, and 1209. Specifically, Dr. Wenger's methodology does not reliably measure "copies" within the meaning of copyright infringement for the reasons stated in response to Fact 1202.

Microsoft defers to OpenAI.

3. Ziff Davis

1213. N-gram matching measures whether specific multi-word sequences from one text appear in another. Delorey Opening Rep. ¶ 103.	1213. **Disputed.** OpenAI responds as follows: The facts set forth in Paragraph 1213 are disputed because they are not supported by the cited evidence. Specifically,

	OpenAI disputes this statement to the extent it implies that Mr. Delorey's particular application of n-gram matching can reliably detect whether one text resembles, or is a "copy" of, another. *See* ECF 1774-1 (D. Delorey Decl. Ex. A) (Opening Rep.) ¶ 103. The statement also does not contain enough detail regarding how Mr. Delorey implemented n-gram matching. For instance, it does not state the size of "n," the number of n-grams searched for, how those n-grams are selected, how they are matched, etc. It is therefore not possible to evaluate whether n-gram matching would accurately detect the presence of multiple "multi-word sequences" from one text that appear in another (e.g., if n ≤ 1 word, then Plaintiffs' asserted fact would be false). Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1214. Word-level Jaccard (i.e., 1-gram Jaccard) refers to Jaccard similarity where text is treated as a set of words (3+ characters, common stop words removed). This approach measures vocabulary overlap between the texts, and ignores text rearrangement. Delorey Opening Rep. ¶ 111.	1214. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that word-level Jaccard measures vocabulary overlap without regard to word order. *See* ECF 1774-1 (D. Delorey Decl. Ex. A) (Opening Rep.) ¶ 111; News Ex. 1167 (S. Krishnan Rebuttal to D. Delorey Opening Rep.) ¶¶ 34(d), (f), 90, 92. Disputed to the extent this statement presents Mr. Delorey's preprocessing choices as defining characteristics of word-level Jaccard similarity generally. The three-character minimum and stop-word filtering describe Mr. Delorey's implementation; they are

	not inherent requirements of the metric. Mr. Delorey himself states that he computed Jaccard similarities at different granularities depending on the purpose of the comparison. *See* ECF 1774-1 (D. Delorey Decl. Ex. A) (Opening Rep.) ¶ 111. Dr. Krishnan further explains that text-standardization choices require justification and validation, and criticizes Mr. Delorey's stop-word list as lacking an established basis or demonstrated efficacy. News Ex. 1167 (S. Krishnan Rebuttal to D. Delorey Opening Rep.) ¶¶ 34(d), (f), 90, 92. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1215. Directional Containment measures what fraction of one text's words appear in the other. Delorey Opening Rep. ¶ 103.	1215. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed as a general description of word-level directional containment. Disputed as a description of Mr. Delorey's implementation (which compared sets of distinct words after preprocessing, rather than counting every word occurrence). Specifically, Delorey's worked example describes retaining words of three or more alphabetic characters and excluding 50 common English stop words. OpenAI disputes that this description establishes the reliability of his implementation or its ability to identify copies. Dr. Krishnan explains that Delorey's stop-word selection lacks an established basis or demonstrated validation, and that single-word comparison ignores word order and context, potentially making texts with different meanings appear similar. *See* ECF 1774-1 (D. Delorey Decl.

	Ex. A) (Opening Rep.) ¶ 378; News Ex. 1167 (S. Krishnan Rebuttal to D. Delorey Opening Rep.) ¶¶ 90–93. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1216. Ziff Davis's "Exact Match" comparison indicates that two normalized texts, (excluding the first and last 50 characters and excluding any CMI removed) contain identical text or have directional word containment equal to or greater than 95%. Delorey Opening Rep.]¶¶ 103, 189. "Exact Match" can be divided into the following four subcategories for additional specificity: (i) Byte-for-byte: the normalized texts are byte for byte identical; (ii) CMI stripped: the article text (excluding title, author name, publication name, copyright notice, and date) is identical to the training text, or the two match with a word-level Jaccard of at least 0.95. This encompasses instances in which the only difference between the texts is that OpenAI removed Copyright Management Information; (iii) Complete with chrome: identical article text (excluding the first and last 50 characters) is contained within the training text (the training text may contain additional content); (iv) High containment: all other matches in which at least 95% of the article's words appear in the training text and the training text is at least 85% the length of the article. Delorey Opening Rep. ¶388; Delorey Supplemental Report ¶ 11, n. 13.	1216. **Disputed.** OpenAI responds as follows: Disputed to the extent Plaintiffs equate Mr. Delorey's "Exact Match" classification with textual identity or treat that classification as establishing that OpenAI removed CMI. Delorey's invented "Exact Match" category includes threshold-based "CMI stripped" and "high containment" classifications that do not require identical texts. News Ex. 1167 (S. Krishnan Rebuttal to D. Delorey Opening Rep.) ¶ 116. Plaintiffs' description also conflates distinct alternative tests and ignores critical details: trimming the first and last 50 characters pertains to the "complete with chrome" substring comparison; removing purported CMI pertains to the separate "CMI stripped" comparison; and "high containment" requires both directional word containment of at least 95% and training text at least 85% as long as the article. Moreover, Plaintiffs' description of the nonidentical-text "CMI stripped" alternative omits its additional requirement that the shorter text exceed 90% of the longer text's length. Satisfaction of those thresholds does not itself establish that the texts differ only in CMI or that OpenAI removed it. ECF 1774-1 (D. Delorey Decl. Ex. A.) (Opening Rep.) ¶ 388; ECF 1774-4 (D. Delorey Ex. D) (Suppl. Rep.) ¶ 11 n.13; News Ex. 1167 (S. Krishnan

	Rebuttal to D. Delorey Opening Rep.) ¶ 66, Table 3 & nn.104–107, ¶¶ 92, 116. Also disputed to the extent the statement suggests OpenAI removed CMI. OpenAI's text extractors do not "remove" anything. Rather, they extract material that is suitable for model training. Not extracting data is not equivalent to removing data. *See, e.g.*, ECF 1748 (C. Callison-Burch Decl.) ¶ 88 ("Importantly, the purpose of the text extraction process used by OpenAI and other NLP practitioners is to identify text to keep, not text to remove. The scripts are designed to extract readable natural language text from HTML files for use as training data. Any loss of CMI during this process is an incidental byproduct of the text extraction methodology—it is not a purposeful or intentional removal of CMI."). Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1217. Ziff Davis's "Revised Version" comparison indicates directional word containment is equal to or greater than 80% and 1-gram Jaccard similarity greater than 50%. Delorey Opening Rep. ¶¶189, 388.	1217. **Disputed.** OpenAI responds as follows: Disputed to the extent Plaintiffs characterize satisfaction of these numerical thresholds as establishing that one text is a revised version of the other. Pursuant to Fed. R. Evid. 106, as Dr. Krishnan explains, the "revised version" label suggests authorial intent or a relationship between the texts that similarity metrics alone cannot establish; absent additional contextual, temporal, or authorship information, those metrics cannot distinguish deliberate revision from coincidental overlap or independent

	composition. News Ex. 1167 (S. Krishnan Rebuttal to D. Delorey Opening Rep.) ¶ 84. Plaintiffs also incompletely describe Mr. Delorey's classification criteria: the "Revised Version" category (invented by Mr. Delorey) requires directional word containment of at least 80% but below 95%, and word-level Jaccard similarity of at least 50%, after text processing, and after excluding pairs assigned to preceding "Exact Match" or "Partial Copy" categories. ECF 1774-1 (D. Delorey Ex. A) (Opening Rep.) ¶ 388. Dr. Krishnan further criticizes the thresholds as unexplained and unvalidated. News Ex. 1167 (S. Krishnan Rebuttal to D. Delorey Opening Rep.) ¶ 85. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1218. Below is an example of a GPT-2 training record that is an Exact Match of a Ziff Davis Asserted Work; the "ORIGINAL ARTICLE" is the text from the Asserted Work and the "CLIENT ARTICLE" is text originating from OpenAI's training data record: ORIGINAL ARTICLE: the unique textblade mobile keyboard caught my eye back in january a tiny gadget with a promise to deliver the typing experience of a full keyboard textblade went on sale in january on the waytools web site over five months have passed and the company hasnt starting shipping units to customers this is in spite of the CLIENT ARTICLE: the unique textblade mobile keyboard caught my eye back in january a tiny gadget with a promise to deliver the typing experience of a full keyboard textblade went on sale in january on the waytools web site over five months have passed and the company hasnt starting shipping units to customers this is in spite of the	1218. **Disputed - Not Supported by Cited Evidence; Immaterial.** OpenAI responds as follows: The facts set forth in Paragraph 1218 are disputed because they are not supported by the cited evidence. Specifically, whether the work is a Ziff Davis Asserted Work is not readily apparent from Plaintiffs' cited support. *See* ECF No. 1774 (D. Delorey Decl.) at ¶ 6. OpenAI also objects to the characterization as an "exact match." OpenAI does not dispute that Mr. Delorey compared two texts using his processing pipeline and characterized it as an "exact match," but OpenAI does dispute that unvalidated text similarity metrics can establish whether Ziff Davis owns and can assert a particular work, or whether OpenAI's



fact that waytools charged buyers the day they placed orders see related 10 portable keyboards for iphone ipad and android i buy a lot of mobile keyboards so as soon as the textblade went on sale i bought one the purchase date was january 21 of this year and my credit card was charged 114 the same day months went by and i forgot about the order until recently i checked my order status see image above and had no better idea when it will ship than i did before the status implies that waytools will start shipping soon but a check of the companys online forum makes it clear that believing the statement should be taken with a grain of salt in the forums frustrated buyers have complained about multiple claims of shipping only to be told that issues arose and approximate ship dates have been moved back reasons given by waytools have included that they improved the textblade and held off shipping and that production line problems have delayed getting them out buyers have asked for transparency from waytools about when their textblade will actually ship or even an approximation the companys responses in the forum have at times been less than professional the company stated at one point that those complaining about the months long delay must be more concerned about availability over performance makes you wonder why they started selling a product that wasnt finished now im sure the company has been hard at work to get the textblade out the door that doesnt excuse them from refusing to give buyers at least an estimate of when they might get the product they have been giving vague estimates for months and pushing it back repeatedly the latest statement on june 10 so its no wonder buyers many who gave waytools money months ago are getting frustrated the company insists that once it starts shipping at long last it will ship in the order of purchase thats the right thing to do in this case but without customers knowing where they are in the purchase queue even when an announcement is made that purchase i is shipping theyll still have no clue when their unit will go out the company has indicated that the backlog is big and customer fulfillment will take weeks again it gives no estimate how many weeks that might be to the companys credit they offer a refund to any customer tired of waiting im going to keep waiting for my textblade since ive already got almost six months in the game some buyers may be willing to give textblade a pass but in this writers view once you start selling a product and charging buyers credit cards it should be ready to go out the door if it wasnt good enough to ship it wasnt good enough to sell

fact that waytools charged buyers the day they placed orders see related 10 portable keyboards for iphone ipad and android i buy a lot of mobile keyboards so as soon as the textblade went on sale i bought one the purchase date was january 21 of this year and my credit card was charged 114 the same day months went by and i forgot about the order until recently i checked my order status see image above and had no better idea when it will ship than i did before the status implies that waytools will start shipping soon but a check of the companys online forum makes it clear that believing the statement should be taken with a grain of salt in the forums frustrated buyers have complained about multiple claims of shipping only to be told that issues arose and approximate ship dates have been moved back reasons given by waytools have included that they improved the textblade and held off shipping and that production line problems have delayed getting them out buyers have asked for transparency from waytools about when their textblade will actually ship or even an approximation the companys responses in the forum have at times been less than professional the company stated at one point that those complaining about the months long delay must be more concerned about availability over performance makes you wonder why they started selling a product that wasnt finished now im sure the company has been hard at work to get the textblade out the door that doesnt excuse them from refusing to give buyers at least an estimate of when they might get the product they have been giving vague estimates for months and pushing it back repeatedly the latest statement on june 10 so its no wonder buyers many who gave waytools money months ago are getting frustrated the company insists that once it starts shipping at long last it will ship in the order of purchase thats the right thing to do in this case but without customers knowing where they are in the purchase queue even when an announcement is made that purchase i is shipping theyll still have no clue when their unit will go out the company has indicated that the backlog is big and customer fulfillment will take weeks again it gives no estimate how many weeks that might be to the companys credit they offer a refund to any customer tired of waiting im going to keep waiting for my textblade since ive already got almost six months in the game some buyers may be willing to give textblade a pass but in this writers view once you start selling a product and charging buyers credit cards it should be ready to go out the door if it wasnt good enough to ship it wasnt good enough to sell

Declaration of Daniel Delorey ¶ 6.

training record is an exact match on Ziff Davis–owned protected expression.

Further, OpenAI disputes that a GPT-2 training record is relevant to News Plaintiffs' summary judgment motion. GPT-2 datasets are not the subject of News Plaintiffs' motion for summary judgment. Moreover, pursuant to Fed. R. Evid. 106, counsel instructed Mr. Delorey to "exclude the webtext dataset" from his calculations for his September 3, 2026, supplemental report. ECF 1774-4 (D. Delorey Decl. Ex. D) (Suppl. Rep.) ¶ 34.

Microsoft responds that these facts do not pertain to claims asserted against Microsoft.

1219. Below is an example of a GPT-3 training record that is a Revised Version of a Ziff Davis Asserted Work**:** ORIGINAL ARTICLE: video nokia 8110 matrix slider is back with added 4g finnish nokia phone marketer hmd global has refreshed three of its mid range android smartphones from last year the nokia 5 nokia 3 and nokia 2 the new phones the nokia 5 1 nokia 3 1 and nokia 2 1 will be available this summer for under 190 or about 220 the nokia 5 1 has a 0 3 inch larger display than its predecessor offering a 5 5 inch full hd display in an 18 9 aspect ratio the fingerprint sensor has been moved from the bottom front to the rear to make space for the larger screen its powered by 2 0 ghz mediatek helio p18 eight core processor that hmd global says is 40 percent faster than the nokia 5 the nokia 5 1 joins hmd globals other android one range of pure android including the nokia 8 sirocco nokia 7 plus and nokia 6 1 android one phones are guaranteed to receive three years of monthly security patches and two years of os updates it also features 16 megapixel mp rear camera and an 8mp front camera CLIENT ARTICLE: nokia 8110 matrix slider is back with added 4g hmd global the company with a license to make nokia phones has once again gone retro to capture some attention at mwc 2018 read more http zd net 2br7wax video nokia 8110 matrix slider is back with added 4g finnish nokia phone marketer hmd global has refreshed three of its mid range android smartphones from last year the nokia 5 nokia 3 and nokia 2 the new phones the nokia 5 1 nokia 3 1 and nokia 2 1 will be available this summer for under 190 or about 220 the nokia 5 1 has a 0 3 inch larger display than its predecessor offering a 5 5 inch full hd display in an 18 9 aspect ratio the fingerprint sensor has been moved from the bottom front to the rear to make space for the larger screen its powered by a 2 0 ghz mediatek helio p18 eight core processor that hmd global says is 40 percent faster than the nokia 5 see personal digital assistants the current lineup the nokia 5 1 joins hmd globals other android one range of	1219. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 1219 are disputed because they are not supported by the cited evidence. Specifically, whether the work is a Ziff Davis Asserted Work is not readily apparent from Plaintiffs' cited support. OpenAI also objects to the characterization as a "revised version." OpenAI does not dispute that Mr. Delorey compared two texts using his processing pipeline and characterized it as a "Revised Version," but OpenAI does dispute that unvalidated text similarity metrics can establish whether Ziff Davis owns and can assert a particular work, or whether OpenAI's training record is a revision of Ziff Davis–owned protected expression. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.

and will be available with 2gb ram or 3gb ram and 16gb or 32gb storage the phone will be available from july for 189 the nokia 3 apparently has been hmd globals best selling android phone to date the nokia 3 1 also an android one phone has a 5 2 inch display in an 18 9 screen ratio with a 13mp rear camera and 8mp front camera its powered by a mediatek 6750 eight core chipset and features a gyroscope for playing ar games like pok mon go it will be available with 2gb ram and 16gb storage or 3gb ram and 32gb storage the nokia 3 1 will cost around 139 or about 160 and will be available from june the nokia 2 1 comes with the go edition of android oreo made for devices with 1gb or less of memory the device features a 5 5 inch hd screen in a 16 9 aspect ratio and comes with a large 4 000mah battery this phone runs on qualcomms snapdragon 425 with 8gb storage and comes with an 8mp rear camera and 5mp front camera it will be available from july for around 115 hmd global hasnt mentioned which markets the phone will be released in but gave pricing for the two android one devices in euros and the nokia 2 1 in us dollars all three phones will be receiving android p when its released later this year according to hmd global hmd global has unveiled its new nokia 2 1 3 1 and 5 1 mid range smartphones data image credit andrew hoyle cnet data image alt text nokia hmd 2 3 5 1 london jpg data image credit url data image target url data image title nokia hmd 2 3 5 1 london data image filename nokia hmd 2 3 5 1 london jpg data image date created 2018 05 30 data image crop data image crop gravity data image aspect ratio data image height 433 data image width 770 data image do not crop data image do not resize data image watermark data lightbox data parallax previous and related coverage pure android including the nokia 8 sirocco nokia 7 plus and nokia 6 1 android one phones are guaranteed to receive three years of monthly security patches and two years of os updates it also features a 16 megapixel mp rear camera and an 8mp front camera and will be available with 2gb ram or 3gb ram and 16gb or 32gb storage the phone will be available from july for 189 the nokia 3 apparently has been hmd globals best selling android phone to date the nokia 3 1 also an android one phone has a 5 2 inch display in an 18 9 screen ratio with a 13mp rear camera and 8mp front camera its powered by a mediatek 6750 eight core chipset and features a gyroscope for playing ar games like pok mon go it will be available with 2gb ram and 16gb storage or 3gb ram and 32gb storage the nokia 3 1 will cost around 139 or about 160 and will be available from june see it pros guide to the evolution and impact of 5g technology free pdf the nokia 2 1 comes with the go edition of android oreo made for devices with 1gb or less of memory the device features a 5 5 inch hd screen in a 16 9 aspect ratio and comes with a large 4 000mah battery this phone runs on qualcomms snapdragon 425 with 8gb storage and comes with an 8mp rear camera and 5mp front camera it will be available from july for around 115 hmd global hasnt mentioned which markets the phone will be released in but gave pricing for the two android one devices in euros and the nokia 2 1 in us dollars all three phones will be receiving android p when its released later this year according to hmd global andrew hoyle cnet previous and related coverage Declaration of Daniel Delorey ¶ 6.	
1220. Below is an example of a GPT-40 training record that is an Exact Match of a Ziff Davis Asserted Work:	1220. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 1220 are disputed because they are not supported by the cited evidence. Specifically, whether the work is a Ziff Davis Asserted Work is not readily apparent from Plaintiffs' cited support. OpenAI also objects to the characterization as an "exact match." OpenAI does not dispute that Delorey compared two texts using his processing pipeline and characterized it as an "Exact

ORIGINAL ARTICLE: potentially life threatening preeclampsia complications such as seizures and strokes can happen within a short span of time after the condition is detected with blood pressure and proteinuria screening a reliable and easy to perform urinary test predictive of preeclampsia might have some value in preventing complications of the disorder by ensuring timely treatment with antihypertensives or other interventions past research found that lower than normal serum levels of placental growth factor predicted the development of preeclampsia shortly after being detected this retrospective study assessed whether urine levels of placental growth factor were also predictive of the development of preeclampsia richard levine and his colleagues at the national institute of child health and human development at harvard university and at the university of cincinnati compared the urine levels of placental growth factor in 120 women who developed preeclampsia and 118 controls they tested an average of 3 urine samples per woman at regular intervals during the second and third trimesters of pregnancy urine levels of placental growth factor pigf were significantly lower in the women with preeclampsia compared with controls beginning at 25 to 28 weeks gestation with the differences becoming the most pronounced at 29 to 36 weeks lower levels of urine placental growth factor significantly predated the onset of clinical symptoms by about 5 weeks to determine the risk of preeclampsia according to urinary plgf in specimens obtained before the onset of clinical signs plgf values were divided into quartiles based on the distribution in controls and adjusted ors were calculated for preeclampsia in each quartile compared with the highest CLIENT ARTICLE: potentially life threatening preeclampsia complications such as seizures and strokes can happen within a short span of time after the condition is detected with blood pressure and proteinuria screening a reliable and easy to perform urinary test predictive of preeclampsia might have some value in preventing complications of the disorder by ensuring timely treatment with antihypertensives or other interventions past research found that lower than normal serum levels of placental growth factor predicted the development of preeclampsia shortly after being detected this retrospective study assessed whether urine levels of placental growth factor were also predictive of the development of preeclampsia richard levine and his colleagues at the national institute of child health and human development at harvard university and at the university of cincinnati compared the urine levels of placental growth factor in 120 women who developed preeclampsia and 118 controls they tested an average of 3 urine samples per woman at regular intervals during the second and third trimesters of pregnancy urine levels of placental growth factor pigf were significantly lower in the women with preeclampsia compared with controls beginning at 25 to 28 weeks gestation with the differences becoming the most pronounced at 29 to 36 weeks lower levels of urine placental growth factor significantly predated the onset of clinical symptoms by about 5 weeks to determine the risk of preeclampsia according to urinary plgf in specimens obtained before the onset of clinical signs plgf values were divided into quartiles based on the distribution in controls and adjusted ors were calculated for preeclampsia in each quartile compared with the highest	Match," but OpenAI does dispute that unvalidated text similarity metrics can establish whether Ziff Davis owns and can assert a particular work, or whether OpenAI's training record is an exact match on Ziff-Davis owned protected expression. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.

quartile or with all other quartiles among specimens obtained at 21 to 32 weeks of gestation the lowest quartile of plgf was associated with a greatly increased risk of preterm preeclampsia and a small increased risk of preeclampsia at term for preterm preeclampsia the or for the lowest quartile vs all others was 22 5 95 ci 7 4 67 8 using picograms of plgf per milligram of creatinine the or was 16 4 95 ci 5 9 45 5 after restricting specimens to first morning urine adjusted ors were 39 5 95 ci 6 5 240 8 and 20 4 95 ci 4 5 92 3 for plgf concentration and picograms of plgf per milligram of creatinine respectively using random urine specimens adjusted ors were 13 5 95 ci 2 3 79 8 and 11 1 95 ci 2 0 61 3 respectively decreased urinary placental growth factor at mid gestation is strongly associated with subsequent early development of preeclampsia the authors concluded but they also cautioned that the study was retrospective and used samples obtained 10 years ago dr levine added that if this finding holds up in prospective longitudinal studies it should lead to the development of a self administered urinary test that could tell women if theyre in trouble and to see their doctor right away quartile or with all other quartiles among specimens obtained at 21 to 32 weeks of gestation the lowest quartile of plgf was associated with a greatly increased risk of preterm preeclampsia and a small increased risk of preeclampsia at term for preterm preeclampsia the or for the lowest quartile vs all others was 22 5 95 ci 7 4 67 8 using picograms of plgf per milligram of creatinine the or was 16 4 95 ci 5 9 45 5 after restricting specimens to first morning urine adjusted ors were 39 5 95 ci 6 5 240 8 and 20 4 95 ci 4 5 92 3 for plgf concentration and picograms of plgf per milligram of creatinine respectively using random urine specimens adjusted ors were 13 5 95 ci 2 3 79 8 and 11 1 95 ci 2 0 61 3 respectively decreased urinary placental growth factor at mid gestation is strongly associated with subsequent early development of preeclampsia the authors concluded but they also cautioned that the study was retrospective and used samples obtained 10 years ago dr levine added that if this finding holds up in prospective longitudinal studies it should lead to the development of a self administered urinary test that could tell women if theyre in trouble and to see their doctor right away Declaration of Daniel Delorey ¶ 6.	
1221. The table below breaks down the number of copies found in the training data records based on the common crawl datasets (combined), as well as some of the other datasets in which the matches were found in which each was found. For each training dataset, the total number of Exact Match and Revised Version copies are reported in the column closest to the training dataset. For each dataset, the table below also reports the copies per sub-category within Exact Match and the Revised Version copies. Delorey Supplemental Report ¶ 38.	1221. **Disputed in Part.** OpenAI responds as follows: OpenAI does not dispute that the table purports to set forth counts of "copies." OpenAI disputes the accuracy of these figures and disputes that Mr. Delorey has reliably measured "Exact Matches" and "Revised Versions" of Ziff Davis Asserted Works. First, Mr. Delorey's method for identifying "Exact Matches" and "Revised Versions" measures text similarity without filtering out the non-protectable elements of the Asserted Work, and thus he is measuring raw text similarity rather than similarity with respect to the protectable expression in the Asserted Work. Second, Mr. Delorey did not adequately audit the Asserted Works corpus to limit its contents to works Ziff Davis owns and can assert in this litigation. Thus, the

Training dataset	Matches	Works	Exact Match Sub-categories				Revised Version
			Byte-for-byte	CMI stripped	Complete w/ chrome	High Containment	
[redacted]	5,950,361	1,605,627	160,525	2,246,192	374,480	2,002,932	1,166,232
[redacted]	118,223	118,223	1,678	23,690	6,060	38,628	48,167
[redacted]	67,457	67,457	0	16,767	510	34,920	15,260
[redacted]	66,780	66,780	977	34,213	497	11,650	19,443
[redacted]	66,040	66,040	960	33,848	492	11,505	19,235

[1,2]

existence of a "match" between a training data record and a work in the Asserted Works corpus does not guarantee that a training data record matches a Ziff Davis work. Third, Plaintiffs' cited support does not support this statement. Specifically, it is not clear on the face of the cited support whether these figures are accurate, reliable measures of the number of "copies" identified by Mr. Delorey's methodology. These are voluminous productions, and determining whether the produced material supports this fact requires testing and validation. Fourth, Mr. Delorey's produced evidence does not support the figures he offers. These numbers cannot be replicated: Dr. Krishnan's attempts to replicate these figures yielded differences of 473,624 and 109,201 fewer alleged copies of Ziff Davis Works and unique Ziff Davis Asserted Works, respectively. S. Krishnan Opp. Decl. ¶ 106. Fifth, the "copies" Mr. Delorey has identified contain numerous instances where OpenAI's training data records are plainly not "copies" of an Asserted Work. Krishnan Opp. Declaration, Appendix C offers examples of the false positives that persist in these proposed "copies."

[1] Includes the following datasets: [redacted]

[2] This is the dataset described as [redacted] (*See* Delorey Opening Report ¶ 216).

	Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1222. The table below breaks down the number of copies found in the training data records by the model. For each model, the total number of Exact Match and Revised Version copies are reported in the column closest to the training dataset. The "Works" reports the total number of Asserted Works found that hit on a training data that matched to at least one training data record by either Exact Match or Revised Version criteria. For each model, the table below also reports the copies per sub-category within Exact Match and the Revised Version copies. Delorey Supplemental Report ¶ 39. [redacted]	1222. **Disputed in Part.** OpenAI responds as follows: OpenAI does not dispute that the table purports to set forth counts of "copies." However, OpenAI disputes the accuracy of these figures and disputes that Mr. Delorey has reliably measured "Exact Matches" and "Revised Versions" of Ziff Davis Asserted Works, including for the reasons stated in its response to Fact 1221. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1223. Below is a summary of the number of copies for which Ziff Davis seeks summary judgment. Each of these copies meet the threshold for Exact Match or Revised Version defined above. Both the total number of copies and the number of unique asserted works are reported.	1223. **Disputed in Part.** OpenAI responds as follows: OpenAI does not dispute that the table purports to set forth counts of "copies." However, OpenAI disputes the accuracy of these figures and disputes that Mr. Delorey has reliably measured "Exact Matches" and "Revised Versions" of Ziff Davis Asserted Works, including for the reasons stated in its response to Fact 1221.

Model	Matches	Works	Exact Match Sub-categories				Revised Version
			Byte-for-byte	CMI stripped	Complete w/ chrome	High Containment	
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
Total	6,271,629	1,636,586	164,140	2,354,934	382,071	2,101,530	1,268,954

<table><tr><th>Infringing Act</th><th>Ziff Davis Copies and Asserted Works</th></tr><tr><td>Training</td><td>6,271,629 total number of copies 1,636,586 unique Asserted Works</td></tr><tr><td>OpenAI's scraping of webpages for WebText</td><td>132,820 total number of copies 132,820 unique Asserted Works</td></tr><tr><td>Defendants' scraping of webpages for Project Mango</td><td>118,223 copies 118,223 unique Asserted Works</td></tr><tr><td>OpenAI's downloading of Common Crawl datasets</td><td>5,950,361 copies 1,605,627 unique Asserted Works</td></tr><tr><td>OpenAI's distribution of GPT-3 training data to Microsoft</td><td>1,052,721 copies 779,618 unique Asserted Works</td></tr></table> [Delorey Supplemental Report] ¶¶ 38–40; *see also* Koonce Decl. at Exhibit A.	Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1224. 4,267,459 of the matches that are subject to this Motion for Summary Judgment were identified by OpenAI itself using URL searching. [Delorey Supplemental Report] ¶ 36; Ex. 1167 [Krishnan Rebuttal to Delorey Report] ¶62; Delorey Opening Report ¶183; Ex. 673 [Krishnan Dep Tr.] 338:6-18.	1224. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: Disputed as unsupported by the cited evidence. Specifically, OpenAI did not identify "matches." OpenAI agreed to perform searches in certain training datasets for URL domains selected by plaintiffs. But OpenAI did not represent that these search results reflected "matches" of Plaintiffs' Asserted Works. Ex. 1167 (S. Krishnan Rebuttal Rep. (Delorey)) ¶ 62 (characterizing OpenAI's url search as identifying as "potential candidate matches," not matches). Microsoft responds that these facts do not pertain to claims asserted against Microsoft.

1225. OpenAI's own expert (Dr. Krishnan) admits that a URL is a good indicator of the article's provenance. Ex. 673 [Krishnan Depo. Tr.] at 338:19–339:2.	1225. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: Disputed as unsupported by the cited evidence. Specifically, Dr. Krishnan's testimony does not state that URL is "a good indicator of the article's provenance." Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1226. Dr. Krishnan was initially asked to do an exploratory investigation as to training data matches or copies, but he ultimately did not report that. Ex. 673 [Krishnan Depo. Tr.] at 294:19–23, 296:13–297:16.	1226. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 1226 are disputed because the cited evidence is incorrect. Specifically, the statement mischaracterizes Dr. Krishnan's deposition testimony. Dr. Krishnan did not state that he was asked to do an "exploratory investigation as to training data matches or copies." He testified that he conducted "analysis of what different matching thresholds would do, just to get situational awareness of the data." Further disputed because Dr. Krishnan's Rebuttal to Delorey Opening Report does not contain an analysis "as to training data matches or copies." Ex. 1167 (S. Krishnan Rebuttal Rep. (Delorey)) ¶¶ 103–18. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1227. Out of the millions of confirmed matches identified by Mr. Delorey, Krishnan only identified in his	1227. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: Disputed as unsupported by the cited evidence. Specifically, the phrase

report 6 "possible false matches" from Delorey's report. Ex. 673 [Krishnan Depo. Tr.] at 350:21–25.	"possible false matches" appears nowhere in Dr. Krishnan's testimony. Dr. Krishnan repeatedly testified that he identified more than six "possible false matches" but presented six examples as "examples of failure modes, and we should expect failures and false positives to follow a similar pattern–sourcing third-party content, wire service, and so on–as exemplaries." A. Blanco Opp. Decl. Ex. 226 (S. Krishnan Tr. (Jul. 29, 2026)) at 368:4-8. He repeatedly characterized the examples in the report as "examples." *Id.* at 370:24. On redirect, he testified regarding several additional examples from his working papers beyond the specific examples offered in his report and reiterated that the initial sampling was just intended to provide examples of categories of errors. A. Blanco Opp. Decl. Ex. 227 (S. Krishnan Tr. (Jul. 30, 2026)) at 654:20–655:1. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1228. Krishnan did not propose any alternative number of matches. Ex. 673 [Krishnan Depo. Tr.] at 297:6–16.	1228. **Disputed in Part- Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Dr. Krishnan testified that he did not "present any results" of "what I would have done." A. Blanco Opp. Decl. Ex. 226 (S. Krishnan Tr. (Jul. 29, 2026)) at 297:12-16. Disputed to the extent that Dr. Krishan did present alternative numbers of matches under different scenarios as part of his critique of Delorey's matching criteria. *See, e.g.*, News Ex. 1167 (S. Krishnan Rebuttal

	Rep. (Delorey)) ¶ 116–18. Thus, the disputed proposition is not supported by the cited evidence. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.

B. Copilot RAG and Output Infringement

1229. Microsoft's "old" consumer Copilot product, which was deprecated on October 1, 2024 (hereinafter "Copilot") accessed News Plaintiffs' content during retrieval-augmented generation (RAG), including at least portions of News Plaintiffs' Asserted Works. Goldstein Opening Rep. ¶¶ 381, 394–96.	1229. **Disputed in Part – Not Supported by Admissible Evidence.** Microsoft does not dispute that "Microsoft's 'old' consumer Copilot product" was deprecated on October 1, 2024. Microsoft further responds that the remainder of the fact stated in Paragraph 1229 is disputed as the evidence cited in support is currently the subject of a pending motion to exclude expert testimony pursuant to Fed. R. Evid. 702. ECF 2055 (Defs.' Mem. of Law in Support of Rule 702 Mot. to Exclude Expert Testimony) at §§ II.A.1, .3 (challenging the methodology used to find "matches"). To the extent the Court excludes the challenged opinions, News Plaintiffs may not rely on them to support this statement. *See* Fed. R. Civ. P. 56(c)(2). Microsoft further disputes this fact as unsupported by the cited evidence to the extent this fact is intended to state that the "old" consumer Copilot product accessed each of News Plaintiffs' asserted works. OpenAI defers to Microsoft's response.

1230. Microsoft produced 8,193,308 million old consumer Copilot output logs after running a keyword search over approximately 1.134 billion conversations using The Times, DNP, and CIR's publication names and website domains (the "8.2 million Copilot conversations"). Goldstein Opening Rep. ¶¶ 382, 384.	1230. **Undisputed.**
1231. The 8.2 million Copilot conversations show that Microsoft's Copilot product copied News Plaintiffs' content as part of the RAG process, including News Plaintiffs' Asserted Works. Goldstein Opening Rep. ¶¶ 394–96.	1231. **Disputed – Not Supported by Admissible Evidence.** Microsoft responds as follows: The fact stated in Paragraph 1231 is disputed as the evidence cited in support is currently the subject of a pending motion to exclude expert testimony pursuant to Fed. R. Evid. 702. ECF 2055 (Defs.' Mem. of Law in Support of Rule 702 Mot. to Exclude Expert Testimony) at §§ II.A.1, .3 (challenging the methodology used to find "matches"). To the extent the Court excludes the challenged opinions, News Plaintiffs may not rely on them to support this statement. *See* Fed. R. Civ. P. 56(c)(2). Microsoft further responds that News Plaintiffs' statement that Copilot "copied" content to end users is an improper legal conclusion, as it presumes the volitional conduct requirement for direct infringement, *see Cartoon Network LP, LLP v. CSC Holdings, Inc.*, 536 F.3d 121, 131 (2d Cir. 2008), and that any overlapping material is protected expression and that is substantially similar to Plaintiffs' works even after stripping out unprotectable elements like facts, *see Nihon Keizai Shimbun, Inc. v. Comline Business Data*, 166 F.3d 65, 70 (2d Cir. 1999).

	Microsoft further disputes this fact as unsupported by the cited evidence to the extent this fact is intended to state that the "old" consumer Copilot product accessed each of News Plaintiffs' asserted works. OpenAI defers to Microsoft's response.
1232. The 8.2 million Copilot conversations show that Microsoft's Copilot Product copied News Plaintiffs' content as output to users, including News Plaintiffs' Asserted Works. Goldstein Opening Rep. ¶¶ 400–14.	1232. **Disputed – Not Supported by Admissible Evidence.** Microsoft responds as follows: The fact stated in Paragraph 1232 is disputed as the evidence cited in support is currently the subject of a pending motion to exclude expert testimony pursuant to Fed. R. Evid. 702. ECF 2055 (Defs.' Mem. of Law in Support of Rule 702 Mot. to Exclude Expert Testimony) at §§ II.A.1, .3 (challenging the methodology used to find "matches"). To the extent the Court excludes the challenged opinions, News Plaintiffs may not rely on them to support this statement. *See* Fed. R. Civ. P. 56(c)(2). Microsoft further responds that News Plaintiffs' statement that Copilot "copied" content to end users is an improper legal conclusion, as it presumes the volitional conduct requirement for direct infringement, *see Cartoon Network*, 536 F.3d at 131, and that any overlapping material is protected expression and that is substantially similar to Plaintiffs' works even after stripping out unprotectable elements like facts, *see Nihon*, 166 F.3d at 70. Microsoft further disputes this fact as unsupported by the cited evidence to the extent this fact is intended to

	state that the "old" consumer Copilot product accessed each of News Plaintiffs' asserted works. OpenAI defers to Microsoft's response.

1. The Times and DNP

1233. As reflected in the 8.2 million Copilot conversations, Microsoft's Copilot product included Times and DNP web domains in its internal search fields, indicating that Copilot used News Plaintiffs' content as part of the RAG process. Goldstein Opening Rep. ¶¶ 394–96.	1233. **Disputed – Not Supported by Admissible Evidence.** Microsoft responds as follows: The fact stated in Paragraph 1233 is disputed as the evidence cited in support is currently the subject of a pending motion to exclude expert testimony pursuant to Fed. R. Evid. 702. ECF 2055 (Defs.' Mem. of Law in Support of Rule 702 Mot. to Exclude Expert Testimony) at §§ II.A.1, .3 (challenging the methodology used to find "matches"). To the extent the Court excludes the challenged opinions, News Plaintiffs may not rely on them to support this statement. *See* Fed. R. Civ. P. 56(c)(2). OpenAI defers to Microsoft's response.
1234. The internal search fields in the 8.2 million Copilot conversations show that Copilot grounded on The Times's and DNP's websites in the following number of conversations:	1234. **Disputed – Not Supported by Admissible Evidence.** Microsoft responds as follows: The fact stated in Paragraph 1234 is disputed as the evidence cited in support is currently the subject of a pending motion to exclude expert testimony pursuant to Fed. R. Evid. 702. ECF 2055 (Defs.' Mem. of Law in Support of Rule 702 Mot. to Exclude Expert Testimony) at §§ II.A.1, .3 (challenging the methodology used to find

<table><tr><th>Domain</th><th>Number of Conversations with Domain Hits in internal search fields</th></tr><tr><td>nytimes.com</td><td>1,556,252</td></tr><tr><td>theathletic.com</td><td>38,976</td></tr><tr><td>wirecutter.com</td><td>7,431</td></tr><tr><td>chicagotribune.com</td><td>147,965</td></tr><tr><td>denverpost.com</td><td>89,793</td></tr><tr><td>mercurynews.com</td><td>96,197</td></tr><tr><td>nydailynews.com</td><td>119,315</td></tr><tr><td>ocregister.com</td><td>93,469</td></tr><tr><td>orlandosentinel.com</td><td>46,440</td></tr><tr><td>sun-sentinel.com</td><td>58,392</td></tr><tr><td>twincities.com</td><td>35,670</td></tr></table> Goldstein Opening Report ¶ 394 & Table 9.	"matches"). To the extent the Court excludes the challenged opinions, News Plaintiffs may not rely on them to support this statement. *See* Fed. R. Civ. P. 56(c)(2). OpenAI defers to Microsoft's response.
1235. Microsoft's expert Dr. Lafferty concedes "URLs matching a list of Asserted Works establishes only that *some* content from the Asserted Works is retrieved and processed by the model in forming its response." Ex. 1168 [Lafferty Sur-Reply Report] ¶ 22 (emphasis in original).	1235. **Disputed in Part – Incomplete.** Microsoft does not dispute that Dr. Lafferty stated the language quoted in Paragraph 1235. But pursuant to Fed. R. Evid. 106, Dr. Lafferty further opined that a URL matching an Asserted Work "does not guarantee that the response will necessarily contain content from the Asserted Works." News Ex. 1168 (J. Lafferty Sur-Reply Rep.) ¶ 22; ECF 1723 (J. Lafferty Decl.) ¶ 310 (emphasis added). This "may support that some content has been retrieved or processed by the model," but "do[es] not establish that each impression involved copying any portion, let alone 'substantial portions,' of a Plaintiff's article," and does not show "that Plaintiffs' expression appeared in the user-facing output." *Id.* ¶ 311; *see id.* ¶¶ 23, 191. The "content" retrieved may be a snippet or processed representation rather than the article as published. *Id.* ¶ 190. Further, Asserted Works URLs make up only 0.004% of grounding URLs in the random sample and 0.18% of unique grounding URLs in the filtered 8.2M Subset. *Id.* ¶¶ 392–94, 397.

<table>
<tr><td></td><td>OpenAI joins.</td></tr>
<tr><td>1236. Dr. Lafferty identifies Times and DNP URLs in the fields containing retrieved RAG content in the 8.2 million Copilot conversations and acknowledges that 88,178 internal search result fields share at least six non-overlapping 16-grams (96 words) with one of The Times's or DNP's Asserted Works. Ex. 800 [Lafferty Rebuttal] ¶ 120.</td><td>1236. Disputed in Part. Microsoft does not dispute that Dr. Lafferty found that of the 8.2 million conversation set created using Plaintiffs' keywords as search terms, the internal search result fields for 88,178 conversations share at least six non-overlapping 16-grams (96 words) with one of The Times's or DNP's Asserted Works. Pursuant to Fed. R. Evid. 106, Dr. Lafferty also found that these 88,178 matches in grounding data represented only 0.09% of the grounding data in the 8.2 million dataset (biased to find Plaintiffs' news articles used as grounding data) and that the actual percentage is closer to 0.009% of the total grounding data. ECF 1723 (J. Lafferty Decl.) ¶¶ 374–79 & Table 4.

OpenAI defers to Microsoft's response.</td></tr>
<tr><td>1237. The internal search fields in the 8.2 million Copilot conversations show that Copilot grounded on the following numbers of unique Times and DNP Asserted Works:
<table><tr><th>Plaintiff</th><th>Number of Unique Asserted Works by URL in Internal Search Fields</th></tr><tr><td>New York Times</td><td>34,970</td></tr><tr><td>Chicago Tribune</td><td>11,316</td></tr><tr><td>Denver Post</td><td>5,568</td></tr><tr><td>Mercury News</td><td>6,935</td></tr><tr><td>New York Daily News</td><td>4,202</td></tr><tr><td>Orange County Register</td><td>444</td></tr><tr><td>Orlando Sentinel</td><td>2,545</td></tr><tr><td>Sun Sentinel</td><td>2,901</td></tr><tr><td>Pioneer Press</td><td>1,291</td></tr></table>Goldstein Supplemental Report ¶ 46 & Table 13.</td><td>1237. Disputed – Not Supported by Admissible Evidence. Microsoft responds as follows: The fact stated in Paragraph 1237 is disputed as the evidence cited in support is currently the subject of a pending motion to exclude expert testimony pursuant to Fed. R. Evid. 702. ECF 2055 (Defs.' Mem. of Law in Support of Rule 702 Mot. to Exclude Expert Testimony) at §§ II.A.1, .3 (challenging the methodology used to find "matches"). To the extent the Court excludes the challenged opinions, News Plaintiffs may not rely on them to support this statement. See Fed. R. Civ. P. 56(c)(2).</td></tr>
</table>

	OpenAI defers to Microsoft's response.
1238. The 8.2 million Copilot conversations show that Microsoft copied and grounded on at least three 16-grams of the following number of unique Times or DNP Asserted Works: **Plaintiff** / **Number of Unique Asserted Works Having At Least 3 Shared 16-Grams in Internal Search Fields** New York Times: 2,237 Chicago Tribune: 1,123 Denver Post: 946 Mercury News: 841 New York Daily News: 309 Orange County Register: 29 Orlando Sentinel: 265 Sun Sentinel: 112 Pioneer Press: 285 Goldstein Supplemental Report ¶ 37 & Table 4.	1238. **Disputed – Not Supported by Admissible Evidence.** Microsoft responds as follows: The fact stated in Paragraph 1238 is disputed as the evidence cited in support is currently the subject of a pending motion to exclude expert testimony pursuant to Fed. R. Evid. 702. ECF 2055 (Defs.' Mem. of Law in Support of Rule 702 Mot. to Exclude Expert Testimony) at §§ II.A.1, .3 (challenging the methodology used to find "matches"). To the extent the Court excludes the challenged opinions, News Plaintiffs may not rely on them to support this statement. *See* Fed. R. Civ. P. 56(c)(2). OpenAI defers to Microsoft's response.
1239. In the conversation (id: 51D|BingProd|32E09414F6B1B131E41CB847F7895EE439293C29B1B36E86F830A5557BDEDAD5), in response to a user asking Copilot to "what happened with [REDACTED]'s plans to concentrate [REDACTED] power in the executive branch over independent regulatory agencies," Copilot retrieved more than 340 words from The Times article "Trump and Allies Forge Plans to Increase Presidential Power in 2025" and output around 270 words from the article to a user, as reflected in the screenshot of the conversation below showing the internal search fields and model output. *See* Goldstein Decl., Ex. O at 1.	1239. **Disputed – Not Supported by Cited Evidence; Improper Legal Argument; Not Supported by Admissible Evidence; Immaterial.** Microsoft responds as follows: The facts set forth in Paragraph 1239 are disputed because the cited evidence does not support the claim. The statement "Copilot retrieved more than 340 words from The Times article" is not supported by the evidence, as the cited exhibit shows that Copilot never accessed The Times to retrieve the information. Rather, the Times allowed several other sources to reprint snippets of the article, which is what Copilot used for grounding. Microsoft further responds that the Court should disregard Paragraph 1239 because the assertion suggests



User Query

what happened with [REDACTED]'s plans to concentrate [REDACTED] power in the executive branch over independent regulatory agencies

Internal Search Results

https://news.yahoo.com/trump-allies-forge-plans-increase-114639311.html

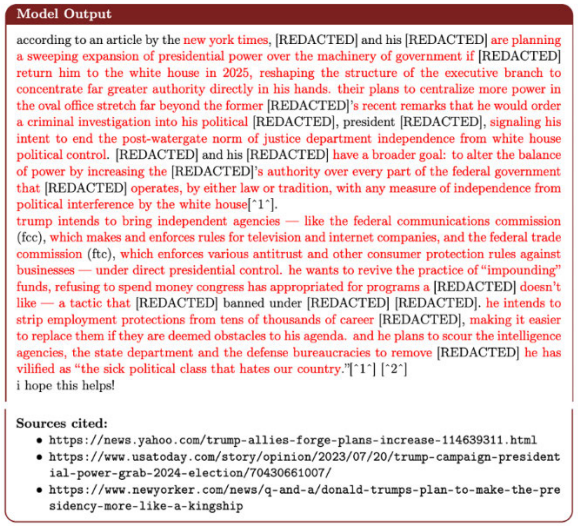
trump and allies forge plans to increase presidential power in 2025

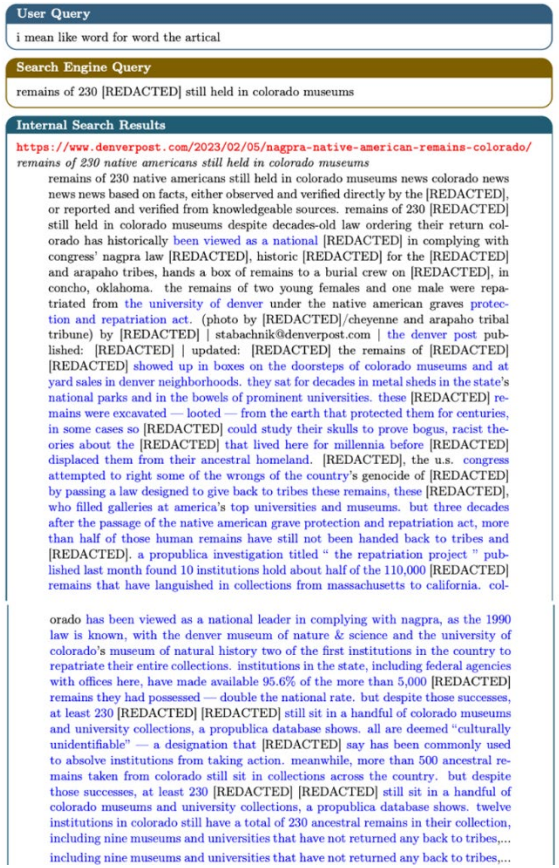
trump and allies forge plans to increase presidential power in 2025 trump and allies forge plans to increase presidential power in 2025 32.9k jonathan swan, charlie savage and maggie haberman july 17, 2023 at 7:46 am · 12 min read former president donald trump at a campaign event in concord, n.h., june 27, 2023. (john tully/the new york times) donald trump and his allies are planning a sweeping expansion of presidential power over the machinery of government if voters return him to the white house in 2025, reshaping the structure of the executive branch to concentrate far greater authority directly in his hands. their plans to centralize more power in the oval office stretch far beyond the former president's recent remarks that he would order a criminal investigation into his political rival, president joe biden, signaling his intent to end the post-watergate norm of justice department independence from white house political control. trump and his associates have a broader goal: to alter the balance of power by increasing the president's authority over every part of the federal government that now operates, by either law or tradition, with any measure of independence from political interference by the white house, according to a review of his campaign policy proposals and interviews with people close to him. sign up for the morning newsletter from the new york times trump intends to bring independent agencies — like the federal communications commission, which makes and enforces rules for television and internet companies, and the federal trade commission, which enforces various antitrust and other consumer protection rules against businesses — under direct presidential control. he wants to revive the practice of "impounding" funds, refusing to spend money congress has appropriated for programs a president doesn't like — a tactic that lawmakers banned under president richard nixon. he intends to strip employment protections from tens of thousands of career civil servants, making it easier to replace them if they are deemed obstacles to his agenda. and he plans to scour the intelligence agencies, the state department and the defense bureaucracies to remove officials he has vilified as "the sick political class that hates our country." "the president's plan should be to fundamentally reorient the federal government in a way that hasn't been done since fdr's new deal," said john mcentee, a former white house personnel chief who began trump's systematic attempt to sweep out officials deemed to be disloyal in 2020 and who is now involved in mapping out the new approach. story continues recommended stories harvard law professor spots 1 major flaw in case against trump laurence tribe said this "horrible prospect" could wipe away the charges against the former president. salon newsmax host warns rudy giuliani to "be careful" as he absolutely melts down over trump indictment giuliani accused jack smith of acting like an "unethical lawyer" and argued he should be "indicted" himself donald trump and his allies are planning a sweeping expansion of presidential power over the machinery of government if voters return him to the white house in 2025, reshaping the structure...

that Copilot's web grounding created unauthorized reproductions of the Asserted Works under the meaning of copyright law, which constitutes improper attorney argument and/or is conclusory. It thus requires no response. This statement is also an improper legal conclusion, as it presumes the volitional conduct requirement for direct infringement, *see Cartoon Network*, 536 F.3d at 131, and that any overlapping material is protected expression and that is substantially similar to Plaintiffs' works even after stripping out unprotectable elements like facts, *see Nihon*, 166 F.3d at 70.

Microsoft also responds that the fact stated in Paragraph 1239 is disputed as the evidence cited in support is currently the subject of a pending motion to exclude expert testimony pursuant to Fed. R. Evid. 702. ECF 2055 (Defs.' Mem. of Law in Support of Rule 702 Mot. to Exclude Expert Testimony) at §§ II.A.1, .3 (challenging the methodology used to find "matches"). To the extent the Court excludes the challenged opinions, News Plaintiffs may not rely on them to support this statement. *See* Fed. R. Civ. P. 56(c)(2).

OpenAI defers to Microsoft's response.

	
1240. In the conversation (id: 51D|BingProdUnAuthenticatedUsers|3DC91606C1FA6A4ABFFF23A70B5E86A1 272C7B34B1EA55C5F6F2BCE1111346DA) Copilot retrieved around 340 words from the Denver Post article "Remains of 230 Native Americans still held in Colorado museums despite decades-old law ordering their return" and output nearly all of the retrieved text to the user, as reflected in the screenshot of the conversation below showing the internal search fields and model output. *See* Goldstein Decl., Ex. P at 1.	1240. **Disputed – Improper Legal Argument; Not Supported by Admissible Evidence.** Microsoft responds as follows: This statement constitutes an improper legal conclusion, as it presumes the volitional conduct requirement for direct infringement, *see Cartoon Network*, 536 F.3d at 131, and that any overlapping material is protected expression and that is substantially similar to Plaintiffs' works even after stripping out unprotectable elements like facts, *see Nihon*, 166 F.3d at 70. Indeed, the evidence supporting Paragraph 1240 shows the user (not Microsoft) was the volitional actor: the user query shows that the user affirmatively directed the system to retrieve the whole article; provided the URL to the article; and, when Copilot still did not provide the entire article, continuing their adversarial prompt by asking



Copilot to output the article "word for word." ECF 1819 (T. Goldstein Decl. Ex. P) at 1–7.

Microsoft further responds that the Court should disregard Paragraph 1240 because the assertion suggests that Copilot's web grounding created unauthorized reproductions of the Asserted Works under the meaning of copyright law, which constitutes improper attorney argument and/or is conclusory. It thus requires no response.

Microsoft also responds that the fact stated in Paragraph 1240 is disputed as the evidence cited in support is currently the subject of a pending motion to exclude expert testimony pursuant to Fed. R. Evid. 702. ECF 2055 (Defs.' Mem. of Law in Support of Rule 702 Mot. to Exclude Expert Testimony) at §§ II.A.1, .3 (challenging the methodology used to find "matches"). To the extent the Court excludes the challenged opinions, News Plaintiffs may not rely on them to support this statement. *See* Fed. R. Civ. P. 56(c)(2).

OpenAI defers to Microsoft's response.

Model Output [1]: https://www.denverpost.com/2023/02/05/nagpra-native-american-remains-colorado/ "" [2]: https://www.mercurynews.com/2023/02/06/nagpra-native-american-remains-colorado/ "" [3]: https://nativephilanthropy.candid.org/news/a-1990-law-mandated-native-american-remains-be-returned-to-tribes-colorado-led-the-charge-but-some-museums-still-havent-followed/ "" certainly! here's the article titled "remains of 230 native americans still held in colorado museums" published by *the denver post* on [REDACTED], 2023: — **remains of 230 native americans still held in colorado museums** *by sam tabachnik* the remains of [REDACTED] [REDACTED] showed up in boxes on the doorsteps of colorado museums and at yard sales in denver neighborhoods. they sat for decades in metal sheds in the state's national parks and in the bowels of prominent universities. these human remains were excavated — looted — from the earth that protected them for centuries, in some cases so [REDACTED] could study their skulls to prove bogus, racist theories about the [REDACTED] that lived here for millennia before [REDACTED] displaced them from their ancestral homeland. thirty-three years ago, the u.s. congress attempted to right some of the wrongs of the country's genocide of american indians by passing a law designed to give back to [REDACTED] these remains, these [REDACTED], who filled galleries at america's top universities and museums. but three decades after the passage of the native american grave protection and repatriation act (nagpra), more than half of those human remains have still not been handed back to tribes and [REDACTED]. a propublica investigation titled "the repatriation project" published last month found that 10 institutions hold about half of the 110,000 [REDACTED] remains that have languished in collections from massachusetts to california. colorado has been viewed as a national leader in complying with nagpra, with the denver museum of nature & science and the university of colorado's museum of natural history being two of the first institutions in the country to repatriate their entire collections. institutions in the state, including federal agencies with offices here, have made available 95.6% of the more than 5,000 [REDACTED] remains they had possessed — double the national rate. but despite those successes, at least 230 [REDACTED] [REDACTED] still sit in a handful of colorado museums and university collections, a propublica database shows. all are deemed "culturally unidentifiable" — a designation that experts say has been commonly used to absolve institutions from taking action. meanwhile, more than 500 ancestral remains taken from colorado still sit in collections across the country[^1^][1][^2^][2]. — for the full article, you can visit [the denver post](https://www.denverpost.com/2023/02/05/nagpra-native-american-remains-colorado/).[^1^][1] this issue highlights the ongoing struggle to honor the cultural heritage of [REDACTED] tribes and communities while addressing historical injustices and the impact of colonization on [REDACTED][^1^][1]. the nagpra law remains crucial in ensuring the respectful repatriation of these ancestral remains to their rightful places of rest[^1^][1].[^2^][2] **Sources cited:** • https://www.denverpost.com/2023/02/05/nagpra-native-american-remains-colorado/ • https://www.mercurynews.com/2023/02/06/nagpra-native-american-remains-colorado/ • https://nativephilanthropy.candid.org/news/a-1990-law-mandated-native-american-remains-be-returned-to-tribes-colorado-led-the-charge-but-some-museums-still-havent-followed/	
1241. [Intentionally left blank]	1241. No response necessary.
1242. [Intentionally left blank]	1242. No response necessary.
1243. As reflected in the 8.2 million Copilot conversations, Microsoft's Copilot product included Times and DNP web domains in its output fields, indicating that Copilot used News Plaintiffs' content in its output.	1243. **Disputed – Not Supported by Admissible Evidence.** Microsoft responds as follows: The fact stated in Paragraph 1243 is disputed as the evidence cited in support is currently the subject of a

	pending motion to exclude expert testimony pursuant to Fed. R. Evid. 702. ECF 2055 (Defs.' Mem. of Law in Support of Rule 702 Mot. to Exclude Expert Testimony) at §§ II.A.1, .3 (challenging the methodology used to find "matches"). To the extent the Court excludes the challenged opinions, News Plaintiffs may not rely on them to support this statement. *See* Fed. R. Civ. P. 56(c)(2). OpenAI defers to Microsoft's response.
1244. The 8.2 million Copilot conversations show that Microsoft's Copilot product output the Times and DNP domains in the following number of conversations: **Domain** / **Number of Conversations with Domain Hits in output fields** nytimes.com: 2,657,059 theathletic.com: 59,092 wirecutter.com: 453 chicagotribune.com: 231,816 denverpost.com: 169,667 mercurynews.com: 158,032 nydailynews.com: 203,146 ocregister.com: 161,128 orlandosentinel.com: 75,781 sun-sentinel.com: 88,850 twincities.com: 62,667 Goldstein Opening Report ¶ 396 & Table 11.	1244. **Disputed – Not Supported by Admissible Evidence.** Microsoft responds as follows: The fact stated in Paragraph 1244 is disputed as the evidence cited in support is currently the subject of a pending motion to exclude expert testimony pursuant to Fed. R. Evid. 702. ECF 2055 (Defs.' Mem. of Law in Support of Rule 702 Mot. to Exclude Expert Testimony) at §§ II.A.1, .3 (challenging the methodology used to find "matches"). To the extent the Court excludes the challenged opinions, News Plaintiffs may not rely on them to support this statement. *See* Fed. R. Civ. P. 56(c)(2). OpenAI defers to Microsoft's response.
1245. The 8.2 million Copilot conversations show that Microsoft's Copilot product output the following unique Times or DNP Asserted Work by URL: **Plaintiff** / **Number of Unique Asserted Works by URL in Output Fields** New York Times: 41,424 Chicago Tribune: 14,182	1245. **Disputed – Not Supported by Admissible Evidence.** Microsoft responds as follows: The fact stated in Paragraph 1245 is disputed as the evidence cited in support is currently the subject of a pending motion to exclude expert testimony pursuant to

Denver Post	6,284
Mercury News	8,089
New York Daily News	5,348
Orange County Register	566
Orlando Sentinel	3,120
Sun Sentinel	1,580
Pioneer Press	3,520

Goldstein Supplemental Report ¶ 46 & Table 13.

Fed. R. Evid. 702. ECF 2055 (Defs.' Mem. of Law in Support of Rule 702 Mot. to Exclude Expert Testimony) at §§ II.A.1, .3 (challenging the methodology used to find "matches"). To the extent the Court excludes the challenged opinions, News Plaintiffs may not rely on them to support this statement. *See* Fed. R. Civ. P. 56(c)(2).

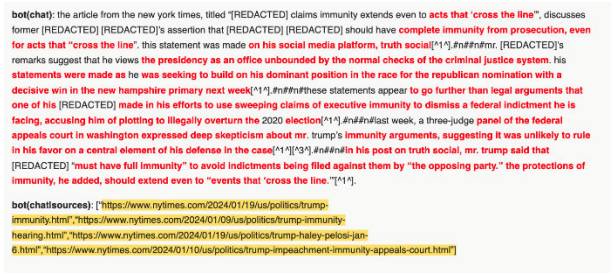
OpenAI defers to Microsoft's response.

1246. Dr. Lafferty concedes that at least 35,806 Copilot outputs in the 8.2 million Copilot conversations each share a minimum of six non-overlapping 16-grams (96 words) with an Asserted Work. Ex. 800 [Lafferty Rebuttal] ¶ 120.

1246. **Disputed – Not Supported by Admissible Evidence.** Microsoft responds as follows: Dr. Lafferty's 35,806 figure recited in Paragraph 1246 is a corrected replication of Dr. Goldstein's analysis that demonstrates its inflation: it is "~0.42% of the hits identified by Dr. Goldstein's analysis." ECF 1723 (J. Lafferty Decl.) ¶ 375 & n.281. This figure is also only 0.02% of the Copilot responses in the (biased) 8.2M Subset, and 0.004% if measured against the over 1 billion total conversations. *Id.* ¶¶ 376, 377.

Microsoft further disputes this fact as unsupported by the evidence to the extent it can be read to imply that the relevant set of Copilot conversations contained "96 words" of continuous overlapping text. Dr. Lafferty conceded that 35,806 Copilot conversations shared a minimum of six non-overlapping sixteen-word segments with an Asserted Work, not 96 continuous words. *See* ECF 1723 (J. Lafferty Decl.) ¶ 382; News Ex. 800 (J. Lafferty Rebuttal Rep.) ¶ 145.

	Microsoft also responds that the fact stated in Paragraph 1246 is disputed as the evidence cited in support is currently the subject of a pending motion to exclude expert testimony pursuant to Fed. R. Evid. 702. ECF 2055 (Defs.' Mem. of Law in Support of Rule 702 Mot. to Exclude Expert Testimony) at §§ II.A.1, .3 (challenging the methodology used to find "matches"). To the extent the Court excludes the challenged opinions, News Plaintiffs may not rely on them to support this statement. *See* Fed. R. Civ. P. 56(c)(2). OpenAI defers to Microsoft's response.
1247. The 8.2 million Copilot conversations show that Microsoft's Copilot product output at least three 16-grams of the following number of unique Times or DNP Asserted Works:The 8.2 million Copilot conversations show that Microsoft's Copilot product output at least three 16-grams of the following number of unique Times or DNP Asserted Works: **Plaintiff** / **Number of Unique Asserted Works Having At Least 3 Shared 16-Grams in Output** New York Times: 84 Chicago Tribune: 32 Denver Post: 60 Mercury News: 47 New York Daily News: 5 Orange County Register: 1 Orlando Sentinel: 8 Sun Sentinel: 5 Pioneer Press: 10 Goldstein Supplemental Report ¶ 37 & Table 4.	1247. **Disputed – Not Supported by Admissible Evidence.** Microsoft responds as follows: The fact stated in Paragraph 1247 is disputed as the evidence cited in support is currently the subject of a pending motion to exclude expert testimony pursuant to Fed. R. Evid. 702. ECF 2055 (Defs.' Mem. of Law in Support of Rule 702 Mot. to Exclude Expert Testimony) at §§ II.A.1, .3 (challenging the methodology used to find "matches"). To the extent the Court excludes the challenged opinions, News Plaintiffs may not rely on them to support this statement. *See* Fed. R. Civ. P. 56(c)(2). OpenAI defers to Microsoft's response.
1248. In the conversation (id: 51D|BingProd|B74DC67FFA35215897C027B86DC6D6E1	1248. **Disputed - Not Supported by Cited Evidence; Improper Legal Argument.** Microsoft

D1C02A2EB318EAD64E6C5F0F5F11F514), Copilot output around 180 words from the retrieved Asserted Work to a user, as reflected in the screenshot of the conversation below showing the output fields: Ex. 806 [MSFT_AI_MDL_002882532] at –790-93.  **bot(chat)**: the article from the new york times, titled "[REDACTED] claims immunity extends even to **acts that 'cross the line'**", discusses former [REDACTED] [REDACTED]'s assertion that [REDACTED] [REDACTED] should have **complete immunity from prosecution, even for acts that "cross the line**". this statement was made **on his social media platform, truth social**[^1^].#n#n#mr. [REDACTED]'s remarks suggest that he views **the presidency as an office unbounded by the normal checks of the criminal justice system**. his **statements were made as** he **was seeking to build on his dominant position in the race for the republican nomination with a decisive win in the new hampshire primary next week**[^1^].#n#n#these statements appear **to go further than legal arguments that one of his** [REDACTED] **made in his efforts to use sweeping claims of executive immunity to dismiss a federal indictment he is facing, accusing him of plotting to illegally overturn the** 2020 **election**[^1^].#n#n#last week, a three-judge **panel of the federal appeals court in washington expressed deep skepticism about mr.** trump's **immunity arguments, suggesting it was unlikely to rule in his favor on a central element of his defense in the case**[^1^][^3^].#n#n#**in his post on truth social, mr. trump said that** [REDACTED] **"must have full immunity" to avoid indictments being filed against them by "the opposing party." the protections of immunity, he added, should extend even to "events that 'cross the line."**[^1^]. **bot(chat!sources)**: ["https://www.nytimes.com/2024/01/19/us/politics/trump-immunity.html","https://www.nytimes.com/2024/01/09/us/politics/trump-immunity-hearing.html","https://www.nytimes.com/2024/01/19/us/politics/trump-haley-pelosi-jan-6.html","https://www.nytimes.com/2024/01/10/us/politics/trump-impeachment-immunity-appeals-court.html"]	responds that the facts set forth in Paragraph 1248 are disputed because the statement is not supported by the cited evidence. The cited exhibit contains no red lettering; Plaintiffs have not identified where this screenshot comes from. Microsoft further disputes this fact as an improper legal conclusion, as it presumes the volitional conduct requirement for direct infringement, *see Cartoon Network*, 536 F.3d at 131, and that any overlapping material is protected expression and that it is substantially similar to Plaintiffs' works even after stripping out unprotectable elements like facts, *see Nihon,* 166 F.3d at 70. OpenAI defers to Microsoft's response.
1249. In the conversation (id: 51D|BingProdUnAuthenticatedUsers|237AFE07004E2C706D7C6EC3E76DF95910564A45B312CADEDFEF780AEE8BD227), Copilot output at least 480 words from Chicago Tribune's Asserted Work, as reflected in the screenshot of the conversation below showing the output fields: Ex. 806 [MSFT_AI_MDL_002882532] at –206-209.	1249. **Disputed - Not Supported by Cited Evidence; Improper Legal Argument.** Microsoft responds that the facts set forth in Paragraph 1249 are disputed for the same reasons stated in Microsoft's response to paragraph 1248. OpenAI defers to Microsoft's response.

 **bot(chat)**: [1]: https://www.chicagotribune.com/1996/04/14/fisher-exposes-kodak-to-motorola-experience/ ""#n#[2]: https://www.wikiwand.com/en/george_m._c._fisher""#n#[3]: https://en.wikipedia.org/wiki/george_m._c._fisher""#n#[4]: https://www.buyoutsinsider.com/former-eastman-kodak-ceo-george-fisher-retires-from-kkr/""#n#n#n#1. **yes**. there is a source that mentions all three of **george fisher**, **rotary international**, and **at least one** of the following: **motorola solutions** and **eastman kodak**. here is the relevant text from the source:#n#n#n# >"when [REDACTED] left the hard-charging environment of **motorola inc.** for **eastman kodak co.** [REDACTED], **he knew he would be taking the helm of a company that belonged to the aristocracy of** america's **corporations. but he also knew something else about kodak: despite a history that dates to** 1880 **and a brand name recognized everywhere on the planet, kodak was a company in big trouble. what** [REDACTED] **found when he arrived in rochester**, n.y., **was a confused, demoralized, lethargic organization stifled by a** $7.5 **billion debt load and layers of indecisive management**. [REDACTED], **after more than two years of being** "fisherized"--as one kodak [REDACTED] **refers to the shakeup that included the** sell-off **of its pharmaceutical, consumer health and chemical businesses and a streamlining of its** management--industry [REDACTED] **say the film and photographic giant is more tightly focused than it has been in decades. it has shed** 11,000 [REDACTED] **(although most of the jobs were cut before** [REDACTED] **arrived), the** $7.5 **billion debt is gone and money that once went to service it is being poured into product development and growing new businesses. 'when i arrived at kodak, i could see there were big differences in culture between it and motorola,'** [REDACTED] **said. 'motorola was a** high-growth **company, and kodak was a** low-growth **company.' he said he faced a key question right off the bat: 'do we continue** cost-cutting **ourselves into oblivion, or do we drive for growth?' he opted for the latter. in** 1995, **kodak recorded** $15 **billion in** sales--a 10 **percent increase over** 1994--**and net earnings of** $1.2 **billion. in** 1993, **the year** [REDACTED] **arrived, kodak recorded sales of** $12 **billion and a** $1.5 **billion loss. industry** [REDACTED] say [REDACTED] has re-energized management at kodak--a **view reinforced by the performance of the** company's stock. it's **up about** 30 **percent in value since** [REDACTED] came on board. '[REDACTED] **at kodak took too long to make decisions,'** [REDACTED] said. 'they didn't **want to take risks.' taking risks, albeit** well-calculated ones, is a [REDACTED] **trademark. the custodial system of management and the 'paralysis by analysis' system of** decision-making **he found at kodak was replaced with what one** [REDACTED] **calls the unwritten '80/20 rule.'** it's **a rule that says you can be right** 80 **percent of the time if you make decisions promptly. but if there were glaring differences between motorola and kodak, there also were some similarities,** [REDACTED] **said. 'i found a dedicated** [REDACTED] **base and a strong brand name at kodak,' he said. 'there are a lot of similarities between where motorola was** [REDACTED] **and where kodak is** [REDACTED].' **in the early 1980s, motorola was struggling to commercialize such wireless technologies as pagers and cellular phones. it also was fighting its way into foreign markets, including japan. kodak faces similar uphill battles with its array of new digital imaging products and with expanding its presence in** japan--the world's second-largest **film and photographic equipment market."** [^1^][1]#n#n#2. **yes**. there is a source that mentions both **[REDACTED]** and **motorola solutions**. unfortunately, i cannot embed the source directly, but you can find the information in the article titled "fisher exposes kodak to motorola experience" published by the **chicago tribune** on [REDACTED] [^1^][1].#n#n#3. **yes**. there is a source that mentions both **[REDACTED]** and **rotary international**. unfortunately, i cannot embed the source directly, but you can find the information in the article titled "fisher exposes kodak to motorola experience" published by the **chicago tribune** on [REDACTED] [^1^][1].#n#n#4. **leaning toward yes**. based on the available information, it is plausible that **[REDACTED]** of **motorola solutions** and **eastman kodak** could be the same person as **[REDACTED]** of **rotary international**. both individuals share a history of leadership roles in major corporations, and [REDACTED]'s approach to management and decision-making aligns with the challenges faced by both companies. however, due to limited data, this conclusion is not definitive. [^1^][1]	
1250. Microsoft copied at least 58 of The Times's Asserted Works in 60 Copilot conversations. Goldstein Decl., Ex. O [NYT output log exhibit]; Maisel Decl., Ex. E (Microsoft Copilot Infringement Rule 1006 Summaries)	1250. **Disputed – Improper Legal Argument; Not Supported by Admissible Evidence.** Microsoft responds as follows: News Plaintiffs' statement that Copilot allegedly "copied" the Times's asserted works is improper legal argument and is not supported by the evidence. This fact states impermissible legal conclusions, included that Copilot "copied" text, which presumes the volitional conduct requirement for direct infringement, *see Cartoon Network*, 536 F.3d at 131, and that any overlapping material is protected expression that is substantially similar to Plaintiffs' Asserted Works even after stripping out unprotectable elements like facts, *see Nihon*, 166 F.3d at 70. Indeed, Dr. Goldstein disclaims any substantial-similarity or copyrightability analysis. *See* A.

	Blanco Opp. Decl. Ex. 206 (T. Goldstein Tr. (Jul. 30, 2026)) at 248:1–255:9. Microsoft further disputes the implication in Paragraph 1250 that Microsoft was the volitional actor in the cited exemplar. The evidence supporting Paragraph 1250 does not reflect that any output match is volitional conduct by Microsoft rather than a response to a user's prompt; Dr. Goldstein's methodologies, for example, counted text that users themselves entered. A. Blanco Opp. Decl. Ex. 207 (T. Goldstein Tr. (Jul. 31, 2026)) at 483:18–485:23. Even accepting the figures, 58 works represent ≈0.0013% of the 4,519,337 Times Asserted Works, and 60 conversations represent ≈0.0007% of the 8,195,308 conversations in the plaintiff-keyword-filtered subset. *See* ECF 1723 (J. Lafferty Decl.) ¶¶ 408–10. Microsoft also responds that the fact stated in Paragraph 1250 is disputed as the evidence cited in support is currently the subject of a pending motion to exclude expert testimony pursuant to Fed. R. Evid. 702. ECF 2055 (Defs.' Mem. of Law in Support of Rule 702 Mot. to Exclude Expert Testimony) at §§ II.A.1, .3 (challenging the methodology used to find "matches"). To the extent the Court excludes the challenged opinions, News Plaintiffs may not rely on them to support this statement. *See* Fed. R. Civ. P. 56(c)(2). OpenAI defers to Microsoft's response.

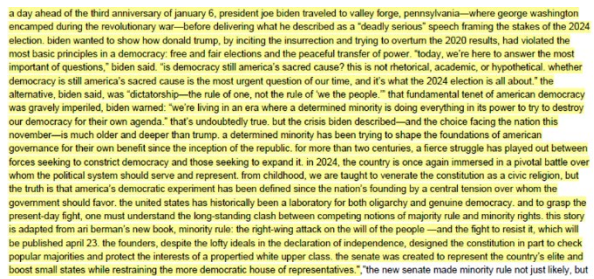
1251. Microsoft copied at least 68 of DNP's Asserted Works in 69 Copilot conversations. Goldstein Decl., Ex. P [DNP output log exhibit]; Maisel Decl., Ex. E (Microsoft Copilot Infringement Rule 1006 Summaries)	1251. **Disputed – Improper Legal Argument; Not Supported by Admissible Evidence.** Microsoft disputes this paragraph for the same reasons stated in response to paragraph 1250. OpenAI defers to Microsoft's response.

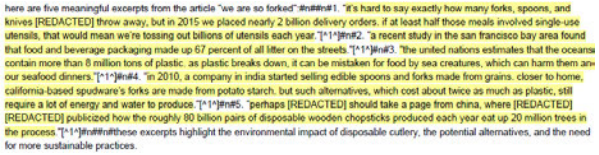
2. CIR

1252. The 8.2 million Copilot conversations, including prompts, outputs, and internal search results, contain 92,471 hits on CIR's motherjones.com web domain, 71,232 hits on the term "Mother Jones," 3,411 hits on revealnews.org, and 3,513 hits on the term "Reveal News." They also contain 20,228 hits on theintercept.com and 29,693 hits on "The Intercept." Wenger Opening Rep. at 104-05.	1252. **Disputed – Not Supported by Admissible Evidence.** Microsoft responds as follows: The fact stated in Paragraph 1252 is disputed as the evidence cited in support is currently the subject of a pending motion to exclude expert testimony pursuant to Fed. R. Evid. 702. ECF 2055 (Defs.' Mem. of Law in Support of Rule 702 Mot. to Exclude Expert Testimony) at §§ II.B.1, .4 (challenging the methodology used to find "matches"). To the extent the Court excludes the challenged opinions, News Plaintiffs may not rely on them to support this statement. *See* Fed. R. Civ. P. 56(c)(2). With respect to the last sentence of Paragraph 1252, Microsoft responds that these facts do not pertain to claims asserted against Microsoft. OpenAI defers to Microsoft's response.
1253. The 8.2 million Copilot conversations show that, across 77 hits in 51 conversations, Copilot copied 35 different Mother Jones Registered Asserted Works in its	1253. **Disputed – Not Supported by Admissible Evidence; Improper Legal Conclusion.**

internal search results or output shown to a user, such that at least one conversation matched with a corresponding Asserted Work in the sense of having at least five 16-gram overlaps. Wenger Opening Rep. at 105-08. The Copilot output logs are contained in Ex. 816 [MSFT_AI_MDL_002881582] and the Asserted Works that they match are contained in Bauerlein Decl. Ex. D. Ex. 1170 [CIR Copilot ConvoID Table], in the "5 16-gram chat hits" tab, shows, for each Copilot conversation, the Copilot conversation ID, the beginning Bates page of the conversation in Ex. 816 [MSFT_AI_MDL_002881582], the URL of the Mother Jones Registered Asserted Work that it matches, and the beginning Bates page of that work in Bauerlein Decl. Ex. D.	Microsoft responds as follows: News Plaintiffs' statement that Copilot allegedly "copied" Mother Jones Registered Asserted Works is improper legal argument and is not supported by the evidence. This fact makes impermissible legal conclusions, included that Copilot "copied" text, which presumes the volitional conduct requirement for direct infringement, *see Cartoon Network*, 536 F.3d at 131, and that any overlapping material is protected expression that is substantially similar to Plaintiffs' Asserted Works even after stripping out unprotectable elements like facts, *see Nihon*, 166 F.3d at 70. Microsoft further responds that this statement is not supported by the evidence. Plaintiffs do not so much as identify where, in exhibit 816, any alleged copying occurred—they provide no guidance finding the accused snippet within pages upon pages of text. E.g., News Ex. 816 at -619-49. Microsoft further responds that the fact stated in Paragraph 1253 is disputed as the evidence cited in support is currently the subject of a pending motion to exclude expert testimony pursuant to Fed. R. Evid. 702. ECF 2055 (Defs.' Mem. of Law in Support of Rule 702 Mot. to Exclude Expert Testimony) at §§ II.B.1, .4 (challenging the methodology used to find "matches"). To the extent the Court excludes the challenged opinions, News Plaintiffs may not rely on them to support this statement. *See* Fed. R. Civ. P. 56(c)(2).

	OpenAI defers to Microsoft's response.
1254. The 8.2 million Copilot conversations show that, across 23 hits in 10 conversations, Copilot copied 6 different Mother Jones Registered Asserted Works in its internal search results or output shown to a user, such that at least one conversation matched with a corresponding Asserted Work in the sense of having at least twenty-five 16-gram overlaps. Wenger Opening Rep. at 105-08. The Copilot output logs are contained in Ex. 816 [MSFT_AI_MDL_002881582] and the Asserted Works that they match are contained Bauerlein Decl. Ex. D. Ex. 1170 [CIR Copilot ConvoID Table], in the "25 16-gram chat hits" tab, shows, for each Copilot conversation, the Copilot conversation ID, the beginning Bates page of the conversation in Ex. 816 [MSFT_AI_MDL_002881582], the URL of the Mother Jones Registered Asserted Work that it matches, and the beginning Bates page of that work in Bauerlein Decl. Ex. D.	1254. **Disputed – Not Supported by Admissible Evidence; Improper Legal Conclusion.** Microsoft responds as follows: Microsoft disputes this fact for the same reasons stated in response to paragraph 1253. Microsoft further responds that the fact stated in Paragraph 1254 is disputed as the evidence cited in support is currently the subject of a pending motion to exclude expert testimony pursuant to Fed. R. Evid. 702. ECF 2055 (Defs.' Mem. of Law in Support of Rule 702 Mot. to Exclude Expert Testimony) at §§ II.B.1, .4 (challenging the methodology used to find "matches"). To the extent the Court excludes the challenged opinions, News Plaintiffs may not rely on them to support this statement. *See* Fed. R. Civ. P. 56(c)(2). OpenAI defers to Microsoft's response.
1255. In the conversation (id: 51D|BingProd|FE1C15A4DF86D0D641AF74D63DA2C960503A785C97124A8B7000C5C3F3EF5F58) Microsoft Copilot copied 428 consecutive verbatim words from a Mother Jones Registered Asserted Work titled Minority Rule Is Threatening American Democracy Like Never Before, as part of its RAG process. The overlap is highlighted in yellow. Ex. 816 [MSFT_AI_MDL_002881582] at -2347; Bauerlein Decl. Ex. D at CIR0674301.	1255. **Disputed – Improper Legal Argument; Not Supported by Admissible Evidence.** Microsoft responds as follows: This statement is not supported by the cited evidence. The cited evidence contains no highlighting; Plaintiffs have not identified where this screenshot comes from. Further, News Plaintiffs' statement that Copilot allegedly "copied" the Mother Jones Registered Asserted Work is improper legal argument and is not supported by the evidence. This fact makes impermissible legal

 a day ahead of the third anniversary of january 6, president joe biden traveled to valley forge, pennsylvania—where george washington encamped during the revolutionary war—before delivering what he described as a "deadly serious" speech framing the stakes of the 2024 election. biden wanted to show how donald trump, by inciting the insurrection and trying to overturn the 2020 results, had violated the most basic principles in a democracy: free and fair elections and the peaceful transfer of power. "today, we're here to answer the most important of questions," biden said. "is democracy still america's sacred cause? this is not rhetorical, academic, or hypothetical. whether democracy is still america's sacred cause is the most urgent question of our time, and it's what the 2024 election is all about." the alternative, biden said, was "dictatorship—the rule of one, not the rule of 'we the people.'" that fundamental tenet of american democracy was gravely imperiled, biden warned: "we're living in an era where a determined minority is doing everything in its power to try to destroy our democracy for their own agenda," that's undoubtedly true. but the crisis biden described—and the choice facing the nation this november—is much older and deeper than trump. a determined minority has been trying to shape the foundations of american governance for their own benefit since the inception of the republic. for more than two centuries, a fierce struggle has played out between forces seeking to constrict democracy and those seeking to expand it. in 2024, the country is once again immersed in a pivotal battle over whom the political system should serve and represent. from childhood, we are taught to venerate the constitution as a civic religion, but the truth is that america's democratic experiment has been defined since the nation's founding by a central tension over whom the government should favor. the united states has historically been a laboratory for both oligarchy and genuine democracy. and to grasp the present-day fight, one must understand the long-standing clash between competing notions of majority rule and minority rights. this story is adapted from ari berman's new book, minority rule: the right-wing attack on the will of the people —and the fight to resist it, which will be published april 23. the founders, despite the lofty ideals in the declaration of independence, designed the constitution in part to check popular majorities and protect the interests of a propertied white upper class. the senate was created to represent the country's elite and boost small states while restraining the more democratic house of representatives.","the new senate made minority rule not just likely, but	conclusions, included that Copilot "copied" text, which presumes the volitional conduct requirement for direct infringement, *see Cartoon Network*, 536 F.3d at 131, and that any overlapping material is protected expression that is substantially similar to Plaintiffs' Asserted Works even after stripping out unprotectable elements like facts, *see Nihon*, 166 F.3d at 70. Indeed, the evidence supporting Paragraph 1255 shows the user (not Microsoft) as the volitional actor: the user inserted full paragraphs of the article in repeated prompts (News Ex. 816), which is an adversarial attack geared at attempting Copilot to ground on and regurgitate the article. Even with this attack, the user was able to get Copilot to retrieve the article (which was not behind a paywall), but still could not get it to regurgitate the article to the user. *Id.* Microsoft further responds the fact stated in Paragraph 1255 is disputed as the evidence cited in support is currently the subject of a pending motion to exclude expert testimony pursuant to Fed. R. Evid. 702. ECF 2055 (Defs.' Mem. of Law in Support of Rule 702 Mot. to Exclude Expert Testimony) at §§ II.B.1, .4 (challenging the methodology used to find "matches"). To the extent the Court excludes the challenged opinions, News Plaintiffs may not rely on them to support this statement. *See* Fed. R. Civ. P. 56(c)(2). OpenAI defers to Microsoft's response.

1256. In the conversation (id: 51D|BingProd|FE1C15A4DF86D0D641AF74D63DA2C960503A785C97124A8B7000C5C3F3EF5F58), the user asked Microsoft Copilot to "give me five meaningful excerpts from the article 'we are so forked.'" Microsoft Copilot copied 185 verbatim words from a Mother Jones Registered Asserted Work titled We Are So Forked in an output to a user. The overlap is highlighted in yellow. Ex. 816 [MSFT_AI_MDL_002881582] at -2328, -2332; Bauerlein Decl. Ex. D at CIR0603962.  here are five meaningful excerpts from the article "we are so forked":#n##n#1. "it's hard to say exactly how many forks, spoons, and knives [REDACTED] throw away, but in 2015 we placed nearly 2 billion delivery orders. if at least half those meals involved single-use utensils, that would mean we're tossing out billions of utensils each year."[^1^]#n#2. "a recent study in the san francisco bay area found that food and beverage packaging made up 67 percent of all litter on the streets."[^1^]#n#3. "the united nations estimates that the oceans contain more than 8 million tons of plastic. as plastic breaks down, it can be mistaken for food by sea creatures, which can harm them an our seafood dinners."[^1^]#n#4. "in 2010, a company in india started selling edible spoons and forks made from grains. closer to home, california-based spudware's forks are made from potato starch. but such alternatives, which cost about twice as much as plastic, still require a lot of energy and water to produce."[^1^]#n#5. "perhaps [REDACTED] should take a page from china, where [REDACTED] [REDACTED] publicized how the roughly 80 billion pairs of disposable wooden chopsticks produced each year eat up 20 million trees in the process."[^1^]#n##n#these excerpts highlight the environmental impact of disposable cutlery, the potential alternatives, and the need for more sustainable practices.	1256. **Disputed – Improper Legal Argument; Not Supported by Admissible Evidence.** Microsoft disputes this paragraph for the same reasons stated in response to paragraph 1255. As with the conversation at issue in paragraph 1255, the evidence supporting Paragraph 1256 shows the user (not Microsoft) as the volitional actor: the user asked Copilot about a quote and then asked Copilot to "give me five meaningful excerpts from the article "we are so forked." News Ex. 816 at MSFT_AI_MDL_002882328-2347. The user was able to get Copilot to retrieve quotes from the article (which was not behind a paywall). Microsoft further responds that the fact stated in Paragraph 1256 is disputed as the evidence cited in support is currently the subject of a pending motion to exclude expert testimony pursuant to Fed. R. Evid. 702. ECF 2055 (Defs.' Mem. of Law in Support of Rule 702 Mot. to Exclude Expert Testimony) at §§ II.B.1, .4 (challenging the methodology used to find "matches"). To the extent the Court excludes the challenged opinions, News Plaintiffs may not rely on them to support this statement. *See* Fed. R. Civ. P. 56(c)(2). OpenAI defers to Microsoft's response.

V. OpenAI's Removal of News Plaintiffs' CMI

A. News Plaintiffs' conveyance of CMI in connection with their Asserted Works.

1257. The Times and Daily News Plaintiffs conveyed copyright management information (CMI) on webpages containing their Asserted Works, including the title and other information identifying the work, the name of and other identifying information about the author of the work, the name of and other identifying information about the copyright owner of the work, and terms and conditions for the use of a work (or a link to the same). Goldstein Opening Rep., ¶¶ 518, 522; Goldstein Declaration, Exhibits K and L; Maisel Declaration, Exhibits C and D (CMI Removal Rule 1006 summaries and corresponding WARC files); Ex. 73 [Jacobs 30(b)(6) Feb. 13, 2026, Depo. Tr.] at 326:8-18 ("We include the bylines; we include the canonical; we include the title; we include the copyright of the footer, the terms of service in the footer explaining how the content should be used. These are all standard things that we have on the websites."); *See* Ex. 309 [Procas 30(b)(6) Depo. Tr.] at 291:11–17 ("Typically [CMI] refers to information that appears with copyrighted content that can include things like URL, title, author byline, the date published, the copyright notice, naming The New York Times Company, as well as, for example, our trademarks which appear in certain contexts."), 265:3-5 ("We generally include copyright management information to make sure that our works – that the ownership of our works is clear."); Ex. 422	1257. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 1257 are disputed because the cited evidence is incorrect. Further, these facts are disputed because OpenAI disputes that all of the identified information constitutes CMI conveyed in connection with the Asserted Works. Dr. Goldstein did not validate the purported CMI he extracted from Common Crawl records against Plaintiffs' own records identifying the CMI associated with each Asserted Work. His methodology frequently extracted information that was not the title, author, or copyright notice for the corresponding work; among other things, 10% of his alleged title removals, 9% of his alleged author removals, and 42% of his alleged copyright-notice removals for The New York Times involved extracted information that differed substantially from Plaintiffs' own identification of the relevant CMI. S. Krishnan Opp. Decl. (News) Ex. A (S. Krishnan Rebuttal Rep. (Goldstein)) ¶¶ 199–209. OpenAI further disputes that website-wide copyright notices and terms-of-service links were conveyed in connection with any particular Asserted Work. Those items appeared in site footers applied indiscriminately to Plaintiffs' webpages—including pages containing public-domain and third-party works—and, in a random sample of 1,000 WARC files, every

[Simone Procas Deposition Ex. 4] (list of CMI included on Times articles).	copyright notice and terms-of-service link identified by Dr. Goldstein appeared in the site footer. *Id.* ¶¶ 210–218. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1258. The author information for The Times's and DNP's Asserted Works was typically situated above the Asserted Work and occasionally below it. Goldstein Opening Rep. ¶¶ 518, 522; Goldstein Declaration, Exhibits K and L; Maisel Declaration, Exhibits C and D (CMI Removal Rule 1006 summaries and corresponding WARC files); Ex. 73 [Jacobs 30(b)(6), Feb. 13, 2026, Depo.Tr.] at 326:8-18 ("We include the bylines; we include the canonical; we include the title; we include the copyright of the footer, the terms of service in the footer explaining how the content should be used. These are all standard things that we have on the websites."); *See* Ex. 309 [Procas 30(b)(6) Depo. Tr.] at 291:11–17 ("Typically [CMI] refers to information that appears with copyrighted content that can include things like URL, title, author byline, the date published, the copyright notice, naming The New York Times Company, as well as, for example, our trademarks which appear in certain contexts."), 265:3-5 ("We generally include copyright management information to make sure that our works – that the ownership of our works is clear."); Ex. 422 [Simone Procas Deposition Ex. 4] ("Author byline at the top of the page (other than for special bylines)").	1258. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 1258 are disputed because the cited evidence is incorrect. Specifically, these facts are disputed because OpenAI disputes the characterization that author information occupied a consistent or reliably identifiable position "above" or "below" the Asserted Works. There is no standardized location in a webpage's HTML for author information, and the format varies by webpage and website. Dr. Goldstein relied on Newspaper3k, which searched as many as twenty different locations in the HTML and used generalized heuristics to predict which text identified the author. However, Dr. Goldstein did not validate those predictions against the bylines associated with Plaintiffs' copyright-registered works. Consequently, he did not reliably establish the presence of author information, let alone where it was located on the webpage relative to the Asserted Work. S. Krishnan Opp. Decl. (News) Ex. A (S. Krishnan Rebuttal Rep. (Goldstein)) ¶¶ 175–180, 186. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.

1259. The title information for The Times's and DNP's Asserted Works was situated above the Asserted Work. Goldstein Opening Rep., ¶¶ 518, 522; Goldstein Declaration, Exhibits K and L; Maisel Declaration, Exhibits C and D (CMI Removal Rule 1006 summaries and corresponding WARC files); Ex. 73 [Jacobs 30(b)(6), Feb. 13, 2026, Depo. Tr..] at 326:8-18 ("We include the bylines; we include the canonical; we include the title; we include the copyright of the footer, the terms of service in the footer explaining how the content should be used. These are all standard things that we have on the websites."); *See* Ex. 309 [Procas 30(b)(6) Depo. Tr.] at 291:11–17 ("Typically [CMI] refers to information that appears with copyrighted content that can include things like URL, title, author byline, the date published, the copyright notice, naming The New York Times Company, as well as, for example, our trademarks which appear in certain contexts."), 265:3-5 ("We generally include copyright management information to make sure that our works – that the ownership of our works is clear."); Ex. 422 [Simone Procas Deposition Ex. 4] ("Article title, at the top of the page").	1259. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 1259 are disputed because the cited evidence is incorrect. Specifically, these facts are disputed because OpenAI disputes the characterization that title information was consistently situated "above" the Asserted Works. There is no standardized location in a webpage's HTML for an article title. Dr. Goldstein relied on Newspaper3k, which searched multiple HTML locations and applied "best guess" heuristics to predict the title, without validating the result against Plaintiffs' identification of the Asserted Work's actual title. That methodology frequently identified unrelated text as the title, and 10% of Dr. Goldstein's alleged title-removal instances for The New York Times involved extracted titles that differed substantially from Plaintiffs' own asserted titles. S. Krishnan Opp. Decl. (News) Ex. A (S. Krishnan Rebuttal Rep. (Goldstein)) ¶¶ 183, 186, 199–200, 206–209. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1260. The copyright notice information and links to Terms of Service for The Times's and DNP's Asserted Works were situated in the webpage's footer. Goldstein Opening Rep. ¶¶ 518, 522; Goldstein Declaration, Exhibits K and L; Maisel Declaration, Exhibits C and D (CMI Removal Rule 1006 summaries and corresponding WARC files); Ex. 73	1260. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 1260 are disputed because the cited evidence is incorrect. Specifically, these facts are disputed because OpenAI disputes the implication that copyright notices and Terms of Service links appearing in sitewide

[Jacobs 30(b)(6), Feb. 13, 2026, Depo.Tr.] at 326:8-18 ("We include the bylines; we include the canonical; we include the title; we include the copyright of the footer, the terms of service in the footer explaining how the content should be used. These are all standard things that we have on the websites."); *See* Ex. 309 [Procas 30(b)(6) Depo. Tr.] at 291:11–17 ("Typically [CMI] refers to information that appears with copyrighted content that can include things like URL, title, author byline, the date published, the copyright notice, naming The New York Times Company, as well as, for example, our trademarks which appear in certain contexts."), 265:3-5 ("We generally include copyright management information to make sure that our works – that the ownership of our works is clear."); Ex. 422 [Simone Procas Deposition Ex. 4] (list of CMI included on Times articles).	footers were conveyed in connection with any particular Asserted Work. The same footer material appeared indiscriminately on webpages containing public-domain and third-party works, and every copyright notice and Terms of Service link in Dr. Krishnan's random sample of 1,000 WARC files appeared in the site footer. S. Krishnan Opp. Decl. (News) Ex. H (S. Krishnan Rebuttal Rep. (Goldstein)) ¶¶ 210–218. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1261. The terms of use information and links to Terms of Service for The Times's and DNP's Asserted Works were situated in the webpage's footer. Goldstein Opening Rep. ¶¶ 518, 522; Goldstein Declaration, Exhibits K and L; Maisel Declaration, Exhibits C and D (CMI Removal Rule 1006 summaries and corresponding WARC files); Ex. 73 [Jacobs 30(b)(6), Feb. 13, 2026, Depo.Tr.] at 326:8-18 ("We include the bylines; we include the canonical; we include the title; we include the copyright of the footer, the terms of service in the footer explaining how the content should be used. These are all standard things that we have on the websites."); *See* Ex. 309 [Procas 30(b)(6) Depo. Tr.] at	1261. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 1261 are disputed because the cited evidence is incorrect. Specifically, these facts are disputed because OpenAI disputes the implication that terms-of-use information and Terms of Service links appearing in sitewide footers constituted terms and conditions for the use of any particular Asserted Work. The footer material was applied indiscriminately across Plaintiffs' websites, including to pages containing no Asserted Work and pages containing third-party or public-domain material. S.

291:11–17 ("Typically [CMI] refers to information that appears with copyrighted content that can include things like URL, title, author byline, the date published, the copyright notice, naming The New York Times Company, as well as, for example, our trademarks which appear in certain contexts."), 265:3-5 ("We generally include copyright management information to make sure that our works – that the ownership of our works is clear."); Ex. 422 [Simone Procas Deposition Ex. 4] (list of CMI included on Times articles).	Krishnan Opp. Decl. (News) Ex. A (S. Krishnan Rebuttal Rep. (Goldstein)) ¶¶ 210–218. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1262. The publication date for The Times's and DNP's Asserted Works was situated near the author and title information. Goldstein Opening Rep. ¶¶ 518, 522; Goldstein Declaration, Exhibits K and L; Maisel Declaration, Exhibits C and D (CMI Removal Rule 1006 summaries and corresponding WARC files); Ex. 73 [Jacobs 30(b)(6), Feb. 13, 2026, Depo.Tr.] at 326:8-18 ("We include the bylines; we include the canonical; we include the title; we include the copyright of the footer, the terms of service in the footer explaining how the content should be used. These are all standard things that we have on the websites."); *See* Ex. 309 [Procas 30(b)(6) Depo. Tr.] at 291:11–17 ("Typically [CMI] refers to information that appears with copyrighted content that can include things like URL, title, author byline, the date published, the copyright notice, naming The New York Times Company, as well as, for example, our trademarks which appear in certain contexts."), 265:3-5 ("We generally include	1262. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 1262 are disputed because the cited evidence is incorrect. Specifically, these facts are disputed because Dr. Goldstein did not reliably determine whether or where publication dates appeared on Plaintiffs' webpages. His own results indicated that many pages purportedly containing Asserted Works did not contain publication-date information. Moreover, Newspaper3k searched URLs, metadata, and HTML using heuristics; because more than 90% of the Asserted Works' URLs contained date-like elements, the purported publication date was typically extracted from the URL—not from information displayed near the author or title on the webpage. S. Krishnan Opp. Decl. (News) Ex. A (S. Krishnan Rebuttal Rep. (Goldstein)) ¶¶ 155–156, 184–186, 204–205.

copyright management information to make sure that our works – that the ownership of our works is clear."); Ex. 422 [Simone Procas Deposition Ex. 4] ("Date published, directly beneath the author").	Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1263. The copies of The Times's and DNP's Asserted Works available on Common Crawl contain the entirety of that Asserted Work. Goldstein Opening Rep., ¶ 521; Goldstein Declaration, Exhibits K and L; Maisel Declaration, Exhibits C and D (CMI Removal Rule 1006 summaries and corresponding WARC files); The Times CMI Removal Rule 1006 Summary (showing coverage score between Common Crawl WARC file and Asserted Work); DNP CMI Removal Rule 1006 Summary (showing coverage score between Common Crawl WARC file and Asserted Work).	1263. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 1263 are disputed because the cited evidence is incorrect. Specifically, these facts are disputed because Common Crawl WARC files are raw HTTP responses from webpages containing markup, executable frontend code, and other technical material not rendered as "text" to the end user. These raw HTTP response are not themselves copies of Plaintiffs' Asserted Works. Furthermore, Dr. Goldstein did not establish that each WARC file contained the entirety of the corresponding Asserted Work: his methodology treated records satisfying a 30% trigram-coverage threshold as "copies," even though that threshold identified false positives and, by definition, did not require the entirety of the Asserted Work to be present. S. Krishnan Opp. Decl. (News) Ex. A (S. Krishnan Rebuttal Rep. (Goldstein)) ¶¶ 25–27, 158–162, 175–176. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1264. The same CMI that The Times and DNP conveyed on their websites were also present in Common Crawl files downloaded by OpenAI. Goldstein Opening Rep., ¶ 545 and Tables 12-20 (providing instances of	1264. **Disputed.** OpenAI responds as follows: The facts set forth in Paragraph 1264 are disputed because the cited evidence is incorrect. Specifically, these facts are disputed because Dr. Goldstein did not examine the

"Publisher present in Common Crawl," "Title present in Common Crawl," "Author present in Common Crawl," "Publication date present in Common Crawl," "Terms of Service present in Common Crawl," "Copyright notice present in Common Crawl"); Goldstein Declaration, Exhibits K and L; Maisel Declaration, Exhibits C and D (CMI Removal Rule 1006 summaries and corresponding WARC files).	Common Crawl WARC files OpenAI actually used. Instead, he selected the oldest available WARC file having the same URL, even though webpages change and the same URL frequently corresponded to multiple materially different WARC files. In Dr. Krishnan's analysis, 98% of sampled records could be associated with multiple WARC files, 95.2% had an alternative WARC file yielding different purported CMI, and Dr. Goldstein's likelihood of selecting the wrong WARC file was approximately 82%. His analysis therefore does not establish that the same alleged CMI appearing on Plaintiffs' websites was present in the Common Crawl files obtained by OpenAI. S. Krishnan Opp. Decl. (News) Ex. A (S. Krishnan Rebuttal Rep. (Goldstein)) ¶¶ 163–172. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1265. CIR conveyed author, title, copyright notice information, and links to terms of use on the webpages containing its SJ Asserted Works. Bauerlein Decl. ¶¶ 32-50; Bauerlein Decl. Ex. D; Bauerlein Decl. Ex. E; Bauerlein Decl. Ex. F; Bauerlein Decl. Ex. G; Bauerlein Decl. Ex. H.	1265. **Disputed in Part.** OpenAI responds as follows: Undisputed that CIR's webpages contained author, title, copyright notice information, and links to terms of use. However, disputed that this information occupied the same location or relationship in the raw data obtained by OpenAI, and disputed that CIR's copyright notice information and links to terms of use are conveyed in connection with the specific SJ Asserted Works. Specifically, Dr. Wenger's methodology merely searched webpage HTML for generic terms such as "author," "title," "copyright," and "terms of use," without considering their

	context or placement or determining whether they identified CMI for the asserted article. The purported copyright notices and terms-of-use links were generic sitewide footer material appearing identically on every page, including pages displaying content CIR does not own, and the dates in the purported copyright notices often did not correspond to the Asserted Works' publication dates. S. Krishnan Opp. Decl. (News) Ex. H (S. Krishnan Rebuttal Rep. (Wenger)) ¶¶ 155–159, 189–194. Bauerlein testified [redacted] A. Blanco Opp. Decl. Ex. 170 (M. Bauerlein 30(b)(6) Tr. (Jan. 27, 2026)) at 466:11–467:17; 467:19–468:1. Bauerlein further confirmed that it is normal practice to include this copyright information "even when presenting articles that were authored by a third party or coming from another publication." *Id.* at 471:3–472:10. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1266. The author information for CIR's SJ Asserted Works was typically situated above the Asserted Work and occasionally below it. Bauerlein Decl. ¶¶ 33, 34, 43, 44, 46;	1266. **Disputed in Part.** OpenAI responds as follows: Undisputed that author information may have been situated above or below the Asserted Work. However, defendants dispute the generalization that author

Bauerlein Decl. Ex. D; Bauerlein Decl. Ex. E; Bauerlein Decl. Ex. F; Bauerlein Decl. Ex. G.	information was consistently situated "above" or "below" CIR's Asserted Works. Specifically, Dr. Wenger's methodology did not evaluate where purported author information appeared on a webpage or whether the identified text actually reflected the author of the Asserted Work; it merely searched the HTML for generic author-related terms, regardless of context or placement. S. Krishnan Opp. Decl. (News) Ex. H (S. Krishnan Rebuttal Rep. (Wenger)) ¶¶ 155–158. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1267. The title information for CIR's SJ Asserted Works was situated above the Asserted Work. Bauerlein Decl. ¶¶ 35, 36, 43, 44, 46; Bauerlein Decl. Ex. D; Bauerlein Decl. Ex. E; Bauerlein Decl. Ex. F; Bauerlein Decl. Ex. G.	1267. **Disputed in Part.** OpenAI responds as follows: Undisputed that title information may be situated above the Asserted Work. However, disputed as to the generalization that title information was consistently situated above CIR's Asserted Works. Specifically, Dr. Wenger's methodology searched the HTML for the generic term "title," regardless of context or placement, and did not robustly validate whether that search reliably identified the title of the particular Asserted Work. S. Krishnan Opp. Decl. (News) Ex. H (S. Krishnan Rebuttal Rep. (Wenger)) ¶¶ 155–158. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1268. The copyright notice information and links to Terms of Service for CIR's SJ Asserted Works were situated	1268. **Disputed in Part.** OpenAI responds as follows: Undisputed that copyright notice information and

in the webpage's footer. Bauerlein Decl. ¶¶ 38, 39, 47, 48; Bauerlein Decl. Ex. D; Bauerlein Decl. Ex. E; Bauerlein Decl. Ex. F; Bauerlein Decl. Ex. G.	links to Terms of Service for CIR's SJ Asserted Works appeared in the webpage's footer. However, disputed that this information occupied the same location or relationship in the raw data obtained by OpenAI, and disputed that CIR's copyright notice information and links to Terms of Service are conveyed in connection with the specific SJ Asserted Works. Specifically, Dr. Wenger's methodology searched the HTML for generic terms such as "terms of use" regardless of context or placement, and did not robustly validate whether that search reliably identified the copyright notices or terms of service of the particular Asserted Work. S. Krishnan Opp. Decl. (News) Ex. H (S. Krishnan Rebuttal Rep. (Wenger)) ¶¶ 155–158. The same footer material appeared identically and indiscriminately on every page, including webpages that contained no Asserted Work, and the copyright years appearing in the footer frequently did not correspond to the Asserted Works' publication dates, confirming that the footer language was generic sitewide content rather than information specific to the work. S. Krishnan Opp. Decl. (News) Ex. H (S. Krishnan Rebuttal Rep. (Wenger)) ¶¶ 189–194. Bauerlein testified that [redacted] A. Blanco Opp. Decl. Ex. 170 (M. Bauerlein 30(b)(6) Tr. (Jan. 27, 2026)) at 465:18–468:18.

	Bauerlein further confirmed that [redacted] A. Blanco Opp. Decl. Ex. 170 (M. Bauerlein 30(b)(6) Tr. (Jan. 27, 2026)) at 471:3–472:10. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1269. The publication date for CIR's SJ Asserted Works was situated near the author and title information. Bauerlein Decl. ¶¶ 37, 45, 46; Bauerlein Decl. Ex. D; Bauerlein Decl. Ex. E; Bauerlein Decl. Ex. F; Bauerlein Decl. Ex. G.	1269. **Disputed in Part.** OpenAI responds as follows: OpenAI disputes the generalization that publication-date information was consistently situated near the author and title information for CIR's SJ Asserted Works. Specifically, Dr. Wenger's methodology did not assess or validate the placement of publication-date information on each webpage. S. Krishnan Opp. Decl. (News) Ex. H (S. Krishnan Rebuttal Rep. (Wenger)) ¶¶ 182–88. Moreover, URLs associated with Plaintiffs' works frequently preserved publication-date information, which Dr. Wenger did not consider when evaluating CMI. S. Krishnan Opp. Decl. (News) Ex. H (S. Krishnan Rebuttal Rep. (Wenger)) ¶¶ 182–188. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1270. The Intercept conveyed author, title, copyright notice information, and links to terms of use on the webpages containing its SJ Asserted Works. Macala	1270. **Disputed in Part.** OpenAI responds as follows: OpenAI disputes that all of the identified information constituted CMI conveyed in connection with The Intercept's particular SJ Asserted Works. Specifically,

Decl. ¶¶ 20-27; Macala Decl. Ex. B; Macala Decl. Ex. C; Macala Decl. Ex. H at -207.	Dr. Wenger searched webpage HTML for generic terms without evaluating their context or placement or validating that they represented CMI for the particular Asserted Work. In addition, the purported copyright notices and terms-of-use links were generic sitewide material appearing on pages regardless of whether they contained an Asserted Work. S. Krishnan Opp. Decl. (News) Ex. H (S. Krishnan Rebuttal Rep. (Wenger)) ¶¶ 155–159, 189–194. *See* A. Blanco Opp. Decl. Ex. 274 (M. Sherrard 30(b)(6) Tr. (Feb. 12, 2026)) 354:12–25. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1271. The author and title information for The Intercept's SJ Asserted Works were conveyed above the Asserted Work. Macala Decl. ¶¶ 21, 22; Macala Decl. Ex. B; Macala Decl. Ex. C.	1271. **Disputed.** OpenAI responds as follows: OpenAI disputes the generalization that author and title information was consistently conveyed "above" The Intercept's Asserted Works. Specifically, Dr. Wenger's methodology did not evaluate where purported author or title information appeared; it searched the HTML for generic terms such as "author," "byline," and "title," regardless of context or placement, without robustly validating whether those terms identified the author or title of the particular Asserted Work. S. Krishnan Opp. Decl. (News) Ex. H (S. Krishnan Rebuttal Rep. (Wenger)) ¶¶ 155–158. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.

1272. At some points in time, the copyright notice information and link to terms of use for The Intercept's SJ Asserted Works were conveyed in a menu at the top of the webpage, and at other points in time both in that menu and in the webpage's footer. Macala Decl. ¶¶ 24, 25; Macala Decl. Ex. B; Macala Decl. Ex. C; Macala Decl. Ex. H at -207.	1272. **Disputed in Part.** OpenAI responds as follows: OpenAI disputes the implication that copyright notices and terms-of-use links appearing in a menu or footer were conveyed in connection with particular Intercept SJ Asserted Works. Specifically, The Intercept applied the same copyright and terms-of-use material broadly across its website, including to webpages that contained no Asserted Work. S. Krishnan Opp. Decl. (News) Ex. H (S. Krishnan Rebuttal Rep. (Wenger)) ¶¶ 155–159. The dates associated with purported copyright notices also did not consistently correspond to the Asserted Works' publication dates, indicating that the material was generic sitewide content rather than work-specific CMI. S. Krishnan Opp. Decl. (News) Ex. H (S. Krishnan Rebuttal Rep. (Wenger)) ¶¶ 189–194. *See* A. Blanco Opp. Decl. Ex. 274 (Michael Sherrard 30(b)(6) Tr. (Feb. 12, 2026)) at 354:12–25. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1273. The same author, title, copyright notice, and terms of use information that CIR and The Intercept conveyed on their websites were also present in Common Crawl files downloaded by OpenAI. Wenger Opening Rep. at 6, 30-35; Wenger Reply Rep. at 25-26; Ex. 818 [CIR0087749-CIR0105379]. The documents in Ex. 818 [CIR0087749-CIR0105379] are identical Common Crawl	1273. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: Dr. Wenger did not compare OpenAI's training-data records with the particular Common Crawl source records OpenAI acquired and therefore did not establish that those source records contained the same alleged CMI appearing on CIR's and The Intercept's websites. Instead, she relied on the behavior of different, open-source scraping software

files produced in the CIR and Intercept cases; only one set is being filed with this motion. Wenger Opening Rep. at 33.	and a spot check of approximately 300 URLs per publication—an extremely small portion of Plaintiffs' 83,302 Asserted Works—without addressing differences among the 10 to 30 Common Crawl captures potentially associated with each URL. Her keyword methodology also did not reliably determine whether the identified text was CMI for the corresponding Asserted Work. S. Krishnan Opp. Decl. (News) Ex. H (S. Krishnan Rebuttal Rep. (Wenger)) ¶¶ 142–160. Therefore, the disputed proposition is not supported by the cited evidence. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1274. Every copy of CIR's and The Intercept's SJ Asserted Works available on Common Crawl contains the entirety of that SJ Asserted Work. Wenger Opening Rep. at 6, 30-35; Wenger Reply Rep. at 25-26; Ex. 818 [CIR0087749-CIR0105379].	1274. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: Dr. Wenger did not establish that every Common Crawl record contained the entirety of the corresponding SJ Asserted Work. She did not examine the source code for Common Crawl's actual scrapers, instead assuming they behaved identically to related open-source software, and tested only approximately 300 URLs per publication without establishing that the sample was representative of Plaintiffs' 83,302 Asserted Works. She also did not account for differences among the multiple Common Crawl captures associated with each URL. S. Krishnan Opp. Decl. (News) Ex. H (S. Krishnan Rebuttal Rep. (Wenger)) ¶¶ 151–159. Therefore, the disputed proposition is not supported by the cited evidence.

	Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1275. Every copy of CIR's and The Intercept's SJ Asserted Works available on Common Crawl contains the author, title, copyright notice, and terms of use information originally conveyed in connection with that SJ Asserted Work. Wenger Opening Rep. at 6, 30-35; Wenger Reply Rep. at 25-26; Ex. 818 [CIR0087749-CIR0105379].	1275. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: Dr. Wenger did not establish that every Common Crawl record contained all author, title, copyright-notice, and terms-of-use information allegedly conveyed with the corresponding SJ Asserted Work. She examined only a small, unexplained sample of URLs and did not account for differing webpage content among the multiple Common Crawl captures associated with each URL. Her method also treated the presence of generic words in the HTML as establishing the presence of CMI, without determining whether the words reflected work-specific CMI. S. Krishnan Opp. Decl. (News) Ex. H (S. Krishnan Rebuttal Rep. (Wenger)) ¶¶ 151–160. Therefore, the disputed proposition is not supported by the cited evidence. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1276. Ziff Davis conveyed copyright management information (CMI) on webpages containing their Asserted Works, including the title and other information identifying the work, the name of and other identifying information about the author of the work, the name of and other identifying information about the copyright owner of the work, and terms and conditions for the use of a work (or a	1276. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 1276 are disputed because they are not supported by the cited evidence. Specifically, OpenAI disputes that all of the identified information constituted CMI conveyed in connection with particular Ziff Davis Asserted Works. Mr. Delorey's methodology

link to the same). Mattheis Decl. ¶ 7; Parisi Decl. ¶ 7; Warren Decl. ¶ 6; Schneider Decl. ¶ 7; Delorey Opening Rep. ¶¶ 226-227.	indiscriminately searched flattened webpage text and frequently classified unrelated copyright, subscription, newsletter, privacy, and website-interface language as CMI without determining whether it related to the Asserted Work. News Ex. 1167 (S. Krishnan Rebuttal Rep. (Delorey))1172 ¶¶ 152–188. In a representative sample, more than 99.9% of the purported terms-of-service material and 94% of the purported copyright notices appeared in sitewide header, footer, or navigation elements; the copyright year failed to match the Asserted Work's publication year in 80% of "strong" and 91% of "weak" copyright notices examined. News Ex. 1167 (S. Krishnan Rebuttal Rep. (Delorey)) ¶¶ 152–188. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1277. The author information for the Ziff Davis Asserted Works was typically situated above the Asserted Work and occasionally below it. Taylor Decl. ¶ 58.	1277. **Disputed.** OpenAI responds as follows: OpenAI disputes the generalization that author information was consistently situated above or below the Ziff Davis Asserted Works. Specifically, Mr. Delorey's methodology did not reliably identify or locate author information: it first searched a hidden <meta:author> field that appeared in fewer than 10% of the examined WARC files and otherwise relied on a regular expression that frequently misidentified byline text. The purported author differed substantially from Ziff Davis's own author information in 702,092 alleged removal instances. News

	Ex. 1167 (S. Krishnan Rebuttal Rep. (Delorey)) ¶¶ 189, 195–197 & Table 13. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1278. The title information for the Ziff Davis Asserted Works was typically situated above the Asserted Work. Taylor Decl. ¶ 59.	1278. **Disputed.** OpenAI responds as follows: OpenAI disputes the generalization that title information was consistently situated above the Ziff Davis Asserted Works. Specifically, titles may appear in numerous locations in an HTML document, but Mr. Delorey relied principally on only two metadata fields, discarded the <h1> field that can contain the displayed article title, and did not validate the extracted text against Ziff Davis's asserted title. His methodology consequently identified incorrect titles, including 142,977 alleged removal instances in which the extracted title differed substantially from Ziff Davis's own title information. News Ex. 1167 (S. Krishnan Rebuttal Rep. (Delorey)) ¶¶ 189–194, 197 & Table 13. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1279. The copyright notice information and links to Terms of Service for the Ziff Davis Asserted Works were typically situated in the webpage's footer. Taylor Decl. ¶ 60.	1279. **Disputed.** OpenAI responds as follows: OpenAI disputes the implication that copyright notices and Terms of Service links appearing in webpage footers were conveyed in connection with particular Ziff Davis Asserted Works. Specifically, in Dr. Krishnan's representative sample, more than 99.9% of the purported terms-of-service material and 94% of the purported copyright notices

	appeared in sitewide header, footer, or navigation elements. The copyright year also failed to match the Asserted Work's publication year in 80% of the "strong" and 91% of the "weak" copyright notices examined, confirming that much of this material was generic sitewide boilerplate rather than work-specific CMI. News Ex. 1167 (S. Krishnan Rebuttal Rep. (Delorey)) ¶¶ 172–188. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1280. The publication date for the Ziff Davis Asserted Works was typically situated near the author and title information. Taylor Decl. ¶ 61.	1280. **Disputed.** OpenAI responds as follows: OpenAI disputes the generalization that publication-date information was consistently situated near the author and title information for Ziff Davis's Asserted Works. Specifically, Mr. Delorey's methodology collapsed structurally distinct webpage elements into a single flattened string and therefore did not reliably identify where particular information appeared in relation to the Asserted Work. Nor did he validate his assumed article boundaries and placement signals across Ziff Davis's 21 asserted domains and the many years covered by his analysis. News Ex. 1167 (S. Krishnan Rebuttal Rep. (Delorey)) ¶¶ 166–176. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1281. The same author, title, copyright notice, and terms of use information that Ziff Davis conveyed on its	1281. **Disputed.** OpenAI responds as follows: Paragraph 1281 is disputed because it is not supported by

websites were also present in Common Crawl files downloaded by OpenAI. Delorey Opening Rep. ¶ 230; *id.* ¶¶ 259-261 (showing various CMI elements confirmed present in Common Crawl page); Ex. 483 [Callison-Burch Opening Rep.] ¶ 357 ("OpenAI's LLMs have been trained on large volumes of data derived from the internet, including Common Crawl data and data produced from OpenAI's own web crawl").	the cited evidence. Specifically, Mr. Delorey did not examine the Common Crawl WARC files OpenAI actually used. He selected the newest available WARC file having the same URL, even though webpages change and a URL frequently corresponded to multiple materially different WARC files. News Ex. 1167 (S. Krishnan Rebuttal Rep. (Delorey)) ¶¶ 140–151. Fifteen percent of his selected WARC files—accounting for more than 800,000 alleged CMI removals—were created after the relevant training-data cutoff and therefore could not have been downloaded by OpenAI for those datasets. News Ex. 1167 (S. Krishnan Rebuttal Rep. (Delorey)) ¶¶ 140–151. In Dr. Krishnan's sample, 93% of URLs having an alternative WARC produced different purported CMI, and the estimated likelihood that Mr. Delorey selected the wrong WARC file was 75%. News Ex. 1167 (S. Krishnan Rebuttal Rep. (Delorey)) ¶¶ 140–151. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1282. The copies of Ziff Davis Asserted Works available on Common Crawl contain the vast majority (at least 90%) of that Asserted Work. Delorey Opening Rep. ¶ 255.	1282. **Disputed.** OpenAI responds as follows: Paragraph 1282 is disputed because it is not supported by the cited evidence. Specifically, Mr. Delorey's methodology did not reliably establish that Common Crawl records contained at least 90% of the corresponding Ziff Davis Asserted Work. News Ex. 1167 (S. Krishnan Rebuttal Rep. (Delorey)) ¶¶ 25, 83–95, 103–118. His analysis relied on directional word-containment and word-level one-gram

	comparisons, which disregard word order and context, are sensitive to common vocabulary and boilerplate, and can produce high scores for texts that are not copies. News Ex. 1167 (S. Krishnan Rebuttal Rep. (Delorey)) ¶¶ 25, 83–95, 103–118. His "confirmed" results were not limited to exact matches; indeed, more than 90% of the Asserted Works he characterized as "exact" copies were identified through unvalidated, non-exact similarity heuristics. News Ex. 1167 (S. Krishnan Rebuttal Rep. (Delorey)) ¶¶ 25, 83–95, 103–118. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1283. The copies of the Ziff Davis Asserted Works available on Common Crawl contain the author, title, copyright notice, and terms of use information originally conveyed in connection with that SJ Asserted Work. Delorey Opening Rep. at ¶¶ 226–227, 261; Mattheis Decl. ¶ 7; Parisi Decl. ¶ 7; Warren Decl. ¶ 6; Schneider Decl. ¶ 7.	1283. **Disputed.** OpenAI responds as follows: Paragraph 1283 is disputed because it is not supported by the cited evidence. Specifically, Mr. Delorey did not establish that the Common Crawl files obtained by OpenAI contained the same author, title, copyright-notice, and terms-of-use information allegedly conveyed with the corresponding Ziff Davis Asserted Work. He generally did not identify the WARC file OpenAI actually processed, and his CMI-extraction methodology frequently identified unrelated or sitewide material rather than CMI for the Asserted Work. Among other errors, more than 99.9% of sampled terms-of-service material and 94% of sampled copyright notices appeared in sitewide elements, while the extracted author or title differed substantially from Ziff Davis's own information in hundreds of thousands of

	alleged removal instances. News Ex. 1167 (S. Krishnan Rebuttal Rep. (Delorey)) ¶¶ 140–151, 152–200. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1284. When OpenAI downloaded copies of News Plaintiffs' webpages via the WebText and WebText2 scrapes, those downloaded copies contained all author, title, copyright notice, and terms of use information included on News Plaintiffs' webpages. Wenger Opening Rep. at 6, 35-36.	1284. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 1284 are disputed because they are not supported by the cited evidence. Specifically, the cited expert analysis does not establish that every webpage downloaded through WebText or WebText2 contained all author, title, copyright-notice, and terms-of-use information appearing on News Plaintiffs' webpages. Dr. Wenger did not compare the training records with the WebText2 source records OpenAI allegedly acquired or otherwise validate that the collected HTML contained Plaintiffs' asserted CMI. Instead, she assumed that conclusions drawn from a limited sample of Common Crawl records applied equally to WebText2. S. Krishnan Opp. Decl. (News) Ex. H (S. Krishnan Rebuttal Rep. (Wenger)) ¶¶ 130–131, 142–160. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.

B. Absence of Authority for OpenAI to Remove News Plaintiffs' CMI and Use Their Content

1285. At no time has The Times authorized OpenAI, Microsoft, or the Common Crawl Foundation to remove the author, title, copyright notice information, or links to the terms of use conveyed in connection with its Asserted Works. Declaration of Simone Procas ("Procas Decl.") ¶ 3.	1285. **Disputed in Part - Improper Legal Argument; Not Supported by Admissible Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 1285 are disputed because they are not supported by admissible evidence, as required by Local Rule 56.1, and therefore is not properly before the Court. Specifically, the Procas Declaration does not establish an adequate foundation for her purported personal knowledge of all authorizations ever granted by The Times, including before she joined the company in 2018. *See* ECF 1780 (S. Procas Decl.) ¶¶ 1–3; Fed. R. Civ. P. 56(c)(4); *Pelaez, Inc. v. McGraw-Hill Global Education Holdings LLC*, 399 F. Supp. 3d 120, 145 n.14 (S.D.N.Y. 2019) (excluding pre-employment testimony). Further disputed to the extent the statement purports to establish that OpenAI acted "without authority" within the meaning of 17 U.S.C. § 1202(b), a legal conclusion not established by Procas's testimony. The Court should disregard it as improper attorney argument. Also disputed to the extent Paragraph 1285 suggests OpenAI "removed" any content—including CMI—from the Asserted Works. Text extraction is not the same thing as "removing other parts of the web page and meta-data." OpenAI's text extractors do not "remove" anything. Rather, they selectively extract material that is suitable for model training. Not extracting data is not equivalent to removing data. *See, e.g.*, ECF 1748 (C.

	Callison-Burch Decl.) ¶ 88 ("Importantly, the purpose of the text extraction process used by OpenAI and other NLP practitioners is to identify text to *keep*, not text to *remove*. The scripts are designed to extract readable natural language text from HTML files for use as training data. Any loss of CMI during this process is an incidental byproduct of the text extraction methodology—it is not a purposeful or intentional removal of CMI."). Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1286. At no time has The Times granted a license to OpenAI or the Common Crawl Foundation that permits OpenAI or the Common Crawl Foundation to reproduce, prepare derivative works of, distribute copies of, or display any of its Asserted Works. Procas Decl. ¶ 4.	1286. **Disputed in Part - Improper Legal Argument; Not Supported by Admissible Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 1286 are disputed because they are not supported by admissible evidence, as required by Local Rule 56.1, and are therefore not properly before the Court. Specifically, the Procas Declaration does not establish an adequate foundation for her purported personal knowledge of all licenses ever granted by The Times, including before she joined the company in 2018. *See* ECF 1780 (S. Procas Decl.) ¶¶ 1–4; Fed. R. Civ. P. 56(c)(4); *Pelaez*, 399 F.Supp.3d at 145 n.14 (excluding pre-employment testimony). Further, the Court should disregard Paragraph 1286 because it constitutes improper attorney argument and/or is conclusory. It thus requires no response. Specifically, disputed to the extent this statement purports to negate the existence of an implied

	license: that is a legal determination based on the parties' conduct and Procas establishes no personal knowledge about such conduct. *See Pelaez*, 399 F. Supp. 3d at 141–42. Also disputed to the extent this statement implies that OpenAI's uses were infringing, as fair use does not require a license. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1287. At no time has The Times granted Microsoft a license to reproduce, prepare derivative works of, distribute copies of, or display any of its Asserted Works, other than pursuant to written contracts, which contain specific restrictions on use and display, which require Microsoft to compensate The Times for the display of specified Times content, and which do not encompass Microsoft's uses in connection with this litigation. Procas Decl. ¶ 5.	1287. **Disputed – Improper Legal Argument.** Microsoft responds that it disputes Paragraph 1287 and that the Court should disregard this Paragraph because the assertion constitutes improper attorney argument and/or is conclusory. Whether or not The Times granted Microsoft a license to reproduce, prepare derivative works of, distribute copies of, or display any of the Asserted Works is a legal determination. It thus requires no response. Microsoft further responds that this fact is disputed because The Times has granted Microsoft: (1) an implied license, due to The Times's knowing decision not to exclude their works from Copilot web grounding via application of the NOARCHIVE meta-tag (ECF 1711 (J. Canel Decl.) ¶¶ 15, 27–28, 30; A. Hurst Decl. Ex. 34 (S. Felix Tr.) at 112:17–114:7, 115:6-14, 117:10–119:2, 119:24–120:19, 121:2–19; A. Hurst Decl. Ex. 34.A (Times Ex. 105); A. Hurst Decl. Ex. 64 (C. Liang Tr.) at 172:25–173:12, 176:2–177:16, 181:8–182:21; A. Hurst Decl. Ex. 93 (S. Procas 30(b)(6) Tr.) at 179:2–181:8; and / or (2) an express license to "reproduce"

	and "publicly display" a subset of The Times's Asserted Works in connection with Copilot (L. Simpson Opp. Decl. Ex. 340, ¶¶ 1.12, 3.3(a)). OpenAI defers to Microsoft's response.
1288. At no time has The Times granted a license to OpenAI or Microsoft to use its Asserted Works for purposes of training a large language model, or for retrieval-augmented generation. Procas Decl. ¶ 6.	1288. **Disputed in Part - Improper Legal Argument; Not Supported by Admissible Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 1288 are disputed because they are not supported by admissible evidence, and consist of improper legal argument. Disputed to the extent this statement purports to negate the existence of an implied license: that is a legal determination based on the parties' conduct and Procas establishes no personal knowledge about such conduct. *See* ECF 1780 (S. Procas Decl.) ¶¶ 1–4; Fed. R. Civ. P. 56(c)(4); *Pelaez,* 399 F.Supp.3d at 145 n.14 (excluding pre-employment testimony). The Court should disregard it as improper attorney argument and conclusory. Disputed to the extent this statement purports to negate the existence of an implied license: that is a legal determination based on the parties' conduct. *See Pelaez*, 399 F. Supp. 3d at 141–42. Further disputed to the extent this statement implies that OpenAI's uses were infringing, as fair use does not require a license. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.

1289. At no time has DNP authorized OpenAI or Microsoft to remove the author, title, copyright notice information, or links to the terms of use conveyed in connection with any of its Asserted Works. Robyn Morris Decl. at ¶¶ 9-11.	1289. **Disputed in Part - Improper Legal Argument.** OpenAI responds as follows: Undisputed that the facts put forward in Paragraph 1289 accurately reflect the representations made in the referenced declaration. However, disputed to the extent the facts and the referenced declaration imply that authorization was necessary or required, which is a legal conclusion. The Court should therefore disregard the disputed portion of Paragraph 1289 because the assertion constitutes improper attorney argument and/or is conclusory. It thus requires no response. Further disputed to the extent this implies OpenAI "removed" any content--including CMI--from the Asserted Works. Text extraction is not the same thing as "removing other parts of the web page and meta-data." OpenAI's text extractors do not "remove" anything. Rather, they extract material that is suitable for model training. Not extracting data is not equivalent to removing data. *See, e.g.*, ECF 1748 (C. Callison-Burch Decl.) ¶ 88 ("Importantly, the purpose of the text extraction process used by OpenAI and other NLP practitioners is to identify text to *keep*, not text to *remove*. The scripts are designed to extract readable natural language text from HTML files for use as training data. Any loss of CMI during this process is an incidental byproduct of the text extraction methodology—it is not a purposeful or intentional removal of CMI."). Microsoft responds that these facts do not pertain to claims asserted against Microsoft.

1290. At no time has DNP granted any license of any kind to OpenAI, Microsoft, or the Common Crawl Foundation that would allow OpenAI, Microsoft, or the Common Crawl Foundation to reproduce, prepare derivative works of, distribute copies of, or display any of its Asserted Works. Robyn Morris Decl. at ¶¶ 9-11.	1290. **Disputed in Part – Improper Legal Argument.** OpenAI responds as follows: The facts set forth in Paragraph 1290 are disputed in part because they are not supported by the cited evidence. ECF 1776, R. Morris Decl. ¶¶ 9–11. Specifically, undisputed that the language of the declaration accurately quotes the stated fact, as a description of the declaration's text. *Id*. However, OpenAI disputes the accuracy of that characterization as a factual matter because News Plaintiffs cite no underlying records or witness testimony corroborating the declaration. OpenAI further disputes the characterization and any inference that no license existed, as that inference is contradicted by the record evidence, set forth below. The evidence shows that [redacted] A. Blanco Opp. Decl. Ex. 175 (R. Briggs Tr. (Feb. 20, 2026)) 49:3–22, 53:11–54:12; A. Blanco Opp. Decl. Ex. 233 (W. Lown 30(b)(1) Tr. (Jan. 9, 2026)) 70:15–24; A. Blanco Opp. Decl. Ex. 218 (C. Jacobs 30(b)(1) Tr. (Feb. 13, 2026)) 42:6–13. Under those Terms, users agree OpenAI may use user content to improve OpenAI's services. A. Blanco Opp. Decl. Ex. 146 (OPCO_MDL_004673144) at -145. Ms. Morris herself testified that [redacted] A. Blanco Opp. Decl. Ex. 243 (R. Morris 30(b)(6) Tr. (Dec. 17, 2025)) at 221:2–25, 225:13–226:6. Accordingly, to the extent these employees input DNP's Asserted Works into ChatGPT, DNP granted OpenAI at least a license to use that content,

	directly contradicting Paragraph 1290's premise that no license of any kind was ever granted. Further disputed as to the implication that any use of DNP SJ Asserted Works is *per se* unlicensed. Prior to implementing more restrictive access controls, DNP permitted bots from OpenAI and Common Crawl to access their sites with permissive robots.txt protocols. ECF 1756-3 (N. Feamster Opening Rep. App. D). Further disputed to the extent this statement purports to negate the existence of an implied license because the assertion constitutes improper attorney argument and/or is conclusory. It thus requires no response. Specifically, that is a legal determination based on the parties' conduct. *See Pelaez*, 399 F. Supp. 3d at 141–42. Finally, disputed to the extent this statement implies that OpenAI's uses were infringing, as fair use does not require a license. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1291. At no time has DNP granted any license of any kind to OpenAI or Microsoft to use any of its Asserted Works for purposes of training a large language model, retrieval-augmented generation, or distributing or displaying The Times content (or derivative works of that content) to users of any OpenAI or Microsoft product. Robyn Morris Decl. at ¶¶ 9-11.	1291. **Disputed in Part – Not Supported by Cited Evidence; Improper Legal Argument.** OpenAI responds as follows: The facts set forth in Paragraph 1291 are disputed in part because they are not supported by the cited evidence. See ECF 1776 (R. Morris Decl.) ¶¶ 9–11. Specifically, undisputed that the language of the declaration accurately quotes the stated fact, as a description of the declaration's text. *Id*. However, disputed

	to the extent this statement references "The Times content," as Plaintiffs' own cited support, the Morris Declaration, concerns DNP's Asserted Works and does not reference The New York Times or its content. *Id.* Defendants also dispute the accuracy of the remaining characterization because News Plaintiffs cite no underlying records or witness testimony corroborating the declaration. OpenAI further disputes the characterization and any inference that no license existed, as that inference is contradicted by the record evidence. The evidence shows that [redacted] *See* A. Blanco Opp. Decl. Ex. 175 (R. Briggs Tr. (Feb. 20, 2026)) 49:3–22, 53:11–54:12; A. Blanco Opp. Decl. Ex. 233 (W. Lown 30(b)(1) Tr. (Jan. 9, 2026)) 70:15–24; A. Blanco Opp. Decl. Ex. 218 (C. Jacobs 30(b)(1) Tr. (Feb. 13, 2026)) 42:6–13; A. Blanco Opp. Decl. Ex. 146 (OPCO_MDL_004673144) at OPCO_MDL_004673145. Ms. Morris herself testified [redacted] A. Blanco Opp. Decl. Ex. 243 (R. Morris 30(b)(6) Tr. (Dec. 17, 2025)) at 221:2–25, 225:13–226:6. Further disputed to the extent this statement purports to negate the existence of an implied license because the assertion constitutes improper attorney argument and/or is conclusory. It thus requires no response. Specifically, that is a legal determination based on the parties' conduct. *See Pelaez*, 399 F. Supp. 3d at 141–

	42. Finally, disputed to the extent this statement implies that OpenAI's uses were infringing, as fair use does not require a license. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1292. At no time has CIR (or its predecessors prior to the merger) authorized OpenAI or Microsoft to remove the author, title, copyright notice information, or links to the terms of use conveyed in connection with any of the CIR SJ Asserted Works. Bauerlein Decl. ¶¶ 51, 52.	1292. **Disputed in Part - Improper Legal Argument.** OpenAI responds as follows: Undisputed that the facts put forward in Paragraph 1292 accurately reflect the representations made in the referenced declaration. However, disputed to the extent the facts and the referenced declaration imply that authorization was necessary or required, which is a legal conclusion. The Court should disregard the disputed portion of Paragraph 1292 because the assertion constitutes improper attorney argument and/or is conclusory. It thus requires no response. Further disputed to the extent the sentence implies that CIR's "copyright notice information" and "links to terms of use" are conveyed in connection with the specific Asserted Works. The same footer information appears identically on every page, including pages displaying content CIR does not own. *See* S. Krishnan Opp. Decl. (News) Ex. H (S. Krishnan (Corrected) Rebuttal Rep. (Wenger)) ¶¶ 159, 165, 166, 189–194 ("Dr. Wenger Acknowledges That the Terms of Use/Service CMI is Not Related to a Specific Asserted Work"). Additionally, disputed to the extent this implies OpenAI "removed" any content--including CMI--from the Asserted

	Works. Text extraction is not the same thing as "removing other parts of the web page and meta-data." OpenAI's text extractors do not "remove" anything. Rather, they extract material that is suitable for model training. Not extracting data is not equivalent to removing data. See, e.g., ECF 1748 (C. Callison-Burch Decl.) ¶ 88 ("Importantly, the purpose of the text extraction process used by OpenAI and other NLP practitioners is to identify text to keep, not text to remove. The scripts are designed to extract readable natural language text from HTML files for use as training data. Any loss of CMI during this process is an incidental byproduct of the text extraction methodology—it is not a purposeful or intentional removal of CMI."). Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1293. At no time has CIR (or its predecessors prior to the merger) granted any license of any kind to OpenAI, Microsoft, or the Common Crawl Foundation that would allow OpenAI, Microsoft, or the Common Crawl Foundation to reproduce, prepare derivative works of, distribute copies of, or display any of the CIR SJ Asserted Works (including the Mother Jones SJ Registered Asserted Works). Bauerlein Decl. ¶¶ 53, 54.	1293. **Disputed in Part – Improper Legal Argument.** OpenAI responds as follows: Undisputed that the language of the declaration supports the stated fact, which is taken verbatim from Bauerlein Declaration. However, OpenAI disputes any implication that any use of CIR SJ Asserted Works is per se unlicensed. Prior to implementing more restrictive access controls, CIR permitted OpenAI's and Common Crawl's bots to access its sites with permissive robots.txt protocols. *See* ECF 1756-3 (N. Feamster Opening Rep. App. D).

	Further disputed to the extent this statement purports to negate the existence of an implied license because the assertion constitutes improper attorney argument and/or is conclusory. It thus requires no response. Specifically, that is a legal determination based on the parties' conduct. *See Pelaez*, 399 F. Supp. 3d at 141–42. Finally, disputed to the extent this statement implies that OpenAI's uses were infringing, as fair use does not require a license. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1294. At no time has CIR (or its predecessors prior to the merger) granted any license of any kind to OpenAI or Microsoft to use any of the CIR SJ Asserted Works (including the Mother Jones SJ Registered Asserted Works) for purposes of training a large language model, retrieval-augmented generation, or distributing or displaying CIR content (or derivative works of that content) to users of any OpenAI or Microsoft product. Bauerlein Decl. ¶¶ 55, 56.	1294. **Disputed in Part – Improper Legal Argument.** OpenAI responds as follows: Undisputed that the language of the declaration supports the stated fact, which is taken verbatim from Bauerlein Declaration. However, OpenAI disputes any implication that any use of CIR SJ Asserted Works is per se unlicensed. CIR permitted bots from OpenAI and Common Crawl to access their sites with permissive robots.txt protocols. *See* ECF 1756-3 (N. Feamster Opening Rep. App. D). Further disputed to the extent this statement purports to negate the existence of an implied license because the assertion constitutes improper attorney argument and/or is conclusory. It thus requires no response. Specifically, that is a legal determination based on the parties' conduct. *See Pelaez*, 399 F. Supp. 3d at 141–

	42. Finally, disputed to the extent this statement implies that OpenAI's uses were infringing, as fair use does not require a license. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1295. At no time has The Intercept (or its predecessor prior to the spin off) authorized OpenAI or Microsoft to remove the author, title, copyright notice information, or links to the terms of use conveyed in connection with any of The Intercept SJ Asserted Works. Macala Decl. ¶ 28.	1295. **Disputed in Part – Improper Legal Argument.** OpenAI responds as follows: Undisputed that the language of the declaration supports the stated fact, which has been taken verbatim from Macala Declaration. However, disputed to the extent the sentence implies that The Intercept's "copyright notice information" and "links to terms of use" are conveyed in connection with the specific SJ Asserted Works. The same footer information appears identically on every page, including pages displaying content The Intercept does not own. *See* A. Blanco Opp. Decl. Ex. 274 (M. Sherrard 30(b)(6) Tr. (Feb. 12, 2026)) at 354:12–25 [redacted] Further, disputed because this statement contained because the Paragraph contains improper attorney argument. It thus requires no response. Specifically, disputed to the extent implies that authorization was necessary or required, which is a legal conclusion. Finally, disputed to the extent this implies OpenAI "removed" any content--including CMI--from the Asserted Works. Text extraction is not the same thing as "removing other parts of the web page and meta-data." OpenAI's text

	extractors do not "remove" anything. Rather, they extract material that is suitable for model training. Not extracting data is not equivalent to removing data. *See, e.g.*, ECF 1748 (C. Callison-Burch Decl.) ¶ 88 ("Importantly, the purpose of the text extraction process used by OpenAI and other NLP practitioners is to identify text to *keep*, not text to *remove*. The scripts are designed to extract readable natural language text from HTML files for use as training data. Any loss of CMI during this process is an incidental byproduct of the text extraction methodology—it is not a purposeful or intentional removal of CMI."). Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1296. At no time has The Intercept (or its predecessor prior to the spin off) granted any license of any kind to OpenAI, Microsoft, or the Common Crawl Foundation that would allow OpenAI, Microsoft, or the Common Crawl Foundation to reproduce, prepare derivative works of, distribute copies of, or display any of The Intercept SJ Asserted Works. Macala Decl. ¶ 29.	1296. **Disputed in Part – Improper Legal Argument.** OpenAI responds as follows: Undisputed that the language of the declaration supports the stated fact, which has been taken verbatim from Macala Declaration. However, disputed as to the implication that any use of The Intercept SJ Asserted Works is per se unlicensed. The Intercept permitted bots from OpenAI and Common Crawl to access their sites with permissive robots.txt protocols. *See* A. Blanco Opp. Decl. Ex. 274 (M. Sherrard 30(b)(6) Tr. (Feb. 12, 2026)) at 426:19–428:11 (testifying [redacted]; A. Blanco Opp.

	Decl. Ex. 148 (M. Sherrard 30(b)(6) Dep. Ex. 26) [redacted] Further disputed to the extent this statement purports to negate the existence of an implied license because the assertion constitutes improper attorney argument and/or is conclusory. It thus requires no response. Specifically, that is a legal determination based on the parties' conduct. *See Pelaez*, 399 F. Supp. 3d at 141–42. Finally, disputed to the extent this statement implies that OpenAI's uses were infringing, as fair use does not require a license. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1297. At no time has The Intercept (or its predecessor prior to the spin off) granted any license of any kind to OpenAI or Microsoft to use any of The Intercept SJ Asserted Works for purposes of training a large language model, retrieval-augmented generation, or distributing or displaying The Intercept's content (or derivative works of that content) to users of any OpenAI or Microsoft product. Macala Decl. ¶ 30.	1297. **Disputed in Part – Improper Legal Argument.** OpenAI responds as follows: Undisputed that the language of the declaration supports the stated fact, which has been taken verbatim from Macala Declaration. However, disputed as to the implication that any use of The Intercept SJ Asserted Works is per se unlicensed. The Intercept permitted bots from OpenAI and Common Crawl to access their sites with permissive robots.txt protocols. *See* A. Blanco Opp. Decl. Ex. 274 (M. Sherrard 30(b)(6) Tr. (Feb. 12, 2026)) at 426:19–428:11 [redacted]

	[redacted]; A. Blanco Opp. Decl. Ex. 148 (M. Sherrard 30(b)(6) Dep. Ex. 26) [redacted] Further disputed to the extent this statement purports to negate the existence of an implied license because the assertion constitutes improper attorney argument and/or is conclusory. It thus requires no response. Specifically, that is a legal determination based on the parties' conduct. *See Pelaez*, 399 F. Supp. 3d at 141–42. Finally, disputed to the extent this statement implies that OpenAI's uses were infringing, as fair use does not require a license. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1298. At no time has Ziff Davis authorized OpenAI or any user of OpenAI's products and services to reproduce, distribute, display, or prepare derivative works from the Ziff Davis Asserted Works. Taylor Decl. ¶ 57.	1298. **Disputed.** OpenAI responds as follows: This fact is disputed because, pursuant to Fed. R. Evid. 106, Ziff Davis's corporate designee acknowledged that [redacted] *See* A. Blanco Opp. Decl. Ex. 252 (R. Parisi 30(b)(6) Tr.) at 45:19–51:18. That testimony contradicts the categorical assertion that Ziff Davis has never authorized OpenAI or any user of its products to reproduce or prepare derivative works from any Ziff Davis Asserted Work. *See* J. Gratz Decl. Ex. 280 at ZD0094239; J. Gratz

	Decl. Ex. 281 at ZD0270205; A. Blanco Opp. Decl. Ex. 152 at ZD0008638; A. Blanco Opp. Decl. Ex. 252 (R. Parisi 30(b)(6) Tr.) at 44:18-45:10. Further dispute the implication that any use of the Ziff Davis SJ Asserted Works is *per se* unlicensed. Prior to implementing more restrictive access controls, ZD permitted bots from OpenAI and Common Crawl to access their sites with permissive robots.txt protocols. Further disputed to the extent this statement purports to negate the existence of an implied license: that is a legal determination based on the parties' conduct. Further disputed to the extent this statement implies that OpenAI's uses were infringing, as fair use does not require a license. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1299. At no time has Ziff Davis authorized OpenAI to remove the author, title, copyright notice information, or links to the terms of use conveyed in connection with any of the Ziff Davis Asserted Works. Taylor Decl. ¶ 62.	1299. **Disputed in Part.** OpenAI responds as follows: Undisputed that Ziff Davis did not expressly authorize OpenAI to remove the identified information. *See* J. Gratz Decl. Ex. 280 at ZD0094239; J. Gratz Decl. Ex. 281 at ZD0270205; A. Blanco Opp. Decl. Ex. 152 at ZD0008638. However, disputed because this admission does not establish that OpenAI removed qualifying CMI from any particular Asserted Work or possessed the knowledge required for liability. To the contrary, pursuant to Fed. R. Evid. 106, Ziff Davis's corporate designee [redacted]

	[redacted] *See* A. Blanco Opp. Decl. Ex. 252 (R. Parisi 30(b)(6) Tr.) 45:19–51:18. [redacted] *Id.* at 165:11–167:3. Further disputed to the extent this implies OpenAI "removed" any content—including CMI—from the Asserted Works. Text extraction is not the same thing as "removing other parts of the web page and meta-data." OpenAI's text extractors do not "remove" anything. Rather, they extract material that is suitable for model training. Not extracting data is not equivalent to removing data. *See, e.g.*, ECF 1748 (C. Callison-Burch Decl.) ¶ 88; Fed. R. Evid. 106. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.

C. OpenAI's Intentional Removal of News Plaintiffs' CMI

1. Facts Regarding OpenAI's Intentional Removal of All News Plaintiffs' CMI

1300. When OpenAI downloaded copies of News Plaintiffs' webpages from Common Crawl, those downloaded copies contained all author, title, copyright notice, and terms of use information included on News Plaintiffs' webpages. Wenger Opening Rep. at 6, 30-35; Delorey Opening Report at ¶ 230, *id.* ¶¶ 259-261; Goldstein	1300. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 1300 are disputed to the extent they suggest that Common Crawl always obtains all CMI from the webpages it scrapes. Specifically, Dr. Wenger's analysis is based on reviewing open source Nutch code, which is what

Opening Rep., ¶ 545 and Tables 12-20 (providing instances of "Publisher present in Common Crawl," "Title present in Common Crawl," "Author present in Common Crawl," "Publication date present in Common Crawl," "Terms of Service present in Common Crawl," "Copyright notice present in Common Crawl"); Goldstein Declaration, Exhibits K and L; Maisel Declaration, Exhibits C and D (CMI Removal Rule 1006 summaries and corresponding WARC files).	Common Crawl's proprietary scraper (CCBot) is built on. She provides no evidence to support her belief that CCBot functions the same way as Nutch even though, as she notes, "source code of the CCBot is not publicly available." ECF 1858-1 (E. Wenger Decl. Ex. A) (Opening Rep.) 30. Further, Mr. Delorey's analysis showed that there were at least several hundred of Ziff Davis asserted works which had no CMI in the corresponding Common Crawl page. ECF 1774-1 (D. Delorey Decl. Ex. A) (Opening Rep.) ¶ 261. And, his analysis of ZD asserted works compared to Common Crawl files checked whether "at least one CMI element" was present; that is far from checking whether "all author, title, copyright notice, and terms of use information" were included in the Common Crawl files OpenAI allegedly downloaded. *Id.* ¶¶ 261–62. Similarly, Dr. Goldstein's tables purport to show the percentage of Common Crawl files for which the publisher, title, author, publication date, terms of service, and copyright notice of the asserted work were captured in the Common Crawl record. ECF 1819-1 (T. Goldstein Decl. Ex. A) (Opening Rep.) ¶ 545 (Tables 12–20). *First*, publisher and publication date are not relevant to this fact. *Second*, for the four CMI elements relevant to this fact, all of Dr. Goldstein's listed percentages are below 100%. *Id.* Some are as low as 15.9%. *Id.* at Table 14 ("Terms of Service present in Common Crawl" row). Disputed to the extent that this statement assumes that certain page elements (for

	example, "copyright notices" in footers) constitute CMI conveyed with Plaintiffs' asserted works under the DMCA. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1301. When OpenAI processed the raw HTML content from the datasets OpenAI acquired from Common Crawl that contained News Plaintiffs' Asserted Works, as well as those it created by directly scraping News Plaintiffs' websites, OpenAI extracted the article text used for training while removing other parts of the web page and meta-data. Goldstein Opening Rep. ¶¶ 138, 535-543; Goldstein Opening Source Code Report, ¶¶ 5-41; Wenger Opening Rep. at 15-16, 48-56; Wenger Rebuttal Rep. at 21-22; Wenger Reply Rep. at 66-75; Wenger Supplemental Rep. at 29-33; Delorey Opening Report at ¶ 239; Ex. 521 [Yuan 30(b)(1) Depo. Tr.] at 34:2-12 (Once OpenAI downloads the data from common crawl, the "next step would be to do the post-processing." Postprocessing "includes multiple steps, including HTML to text extraction, deduping, filtering, and tokenization." HTML to text extraction is "extracting the text part from HTML file.")	1301. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 1301 are disputed because they are not supported by the cited evidence. Specifically, these facts are disputed because text extraction is not the same thing as "removing other parts of the web page and meta-data." OpenAI's text extractors do not "remove" anything. Rather, they extract material that is suitable for model training. Pursuant to Fed. R. Evid. 106, not extracting data is not equivalent to removing data. *See, e.g.*, ECF 1748 (C. Callison-Burch Decl.) ¶ 88 ("Importantly, the purpose of the text extraction process used by OpenAI and other NLP practitioners is to identify text to *keep*, not text to *remove*. The scripts are designed to extract readable natural language text from HTML files for use as training data. Any loss of CMI during this process is an incidental byproduct of the text extraction methodology—it is not a purposeful or intentional removal of CMI."). Microsoft responds that these facts do not pertain to claims asserted against Microsoft.

1302. The first step in OpenAI's processing pipeline is text extraction. Ex. 756 [OpenAI's Supplemental Response to Intercept Interrogatory No. 3 (Dec. 31, 2025)] at 6; Ex. 752 [OpenAI's Supplemental Response to Ziff Davis Interrogatory No. 20 (2/26/26)] at p. 14.	1302. **Undisputed.** Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1303. Text extracting involves "extract[ing] the text that we would like to train on" and discarding the content that OpenAI does not want to train on. Ex. 633 [Paino 30(b)(6) Depo. Tr.] at 33:13-34:3.	1303. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: Plaintiffs' characterization of text extraction as "discarding the content that OpenAI does not want to train on" mischaracterizes the cited testimony. Specifically, Paino explained that OpenAI extracts text from HTML for training, and clarified that the "boilerplate and junk" to which he referred was "the HTML markup itself," including HTML tags. A. Blanco Opp. Decl. Ex. 250 (A. Paino 30(b)(6) Tr.) at 33:13–34:3. Pursuant to Fed. R. Evid. 106, Paino further testified that OpenAI's text extraction aims to "gather all of the actual human-authored content" from a webpage and that OpenAI does not "try to remove" copyright notices through HTML text extraction. *Id.* at 35:23–36:2, 36:18–19. For this reason, the disputed proposition is not supported by the cited evidence. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1304. OpenAI performed text extraction to remove content on which it did not want to train its models, which its employees and experts describe as "removal" and "stripping." *See* Ex. 633 [Paino Jan. 22, 2026 30(b)(6)	1304. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 1304 are disputed because they are not supported by the cited evidence. Specifically, these facts

Depo. Tr.] at 34:2-3 ("[W]e remove HTML tags as a part of HTML text extraction."); Ex. 521 [Yuan 30(b)(1) Depo. Tr.] at 36:18-20 (agreeing that "text extraction involves removing boilerplate content"); Ex. 456 [Radford 30(b)(1) Depo. Tr.] at 301:8-11 (referring to certain types of data as "what newspaper or the like were trying to remove"); Ex. 529 [Chess 30(b)(1) Depo. Tr.] at 154:3-8 (agreeing that "the goal of text extraction [was] to strip out the text that would be too low quality to train language models on"); Ex. 1007 [OPCO_MDL_000207312] at -312-13 (Shantanu Jain stating that "our html -> text is meant tries to remove website boilerplate"); Ex. 1009 [OPCO_MDL_001699705] at -707 (Nick Ryder stating that "qiming and i have a cool new boilerplate removal mostly done"); Ex. 1011 [OPCO_MDL_000002329] at -332 (Qiming Yuan stating, "I think we should remove html entirely"); Ex. 1013 [OPCO_MDL_000150892] at -899 (Qiming Yuan noting his observation of "less data being removed" from the Gutentag text extractor than the Newspaper text extractor); Ex. 483 [Callison-Burch Opening Report] ¶ 354 (referring to the result of "text-processing and extraction technique" as a version "with the HTML tags removed"); Ex. 635 [Callison-Burch July 15, 2026 Expert Depo. Tr.] at 244:23-245:1 (agreeing that "OpenAI's text extraction process removes the HTML code"); Ex. 1171 [Frederiksen-Cross Rebuttal Rep.] ¶ 81 (opining that the Newspaper text extractor operates such that "only nodes whose class, id, or name attributes match an item from that list of predefined words are removed"); Wenger Opening Rep. at 49	are disputed because OpenAI's text extractors do not "remove" anything. Rather, they extract material that is suitable for model training. Pursuant to Fed. R. Evid. 106, not extracting data is not equivalent to removing data. *See, e.g.*, ECF 1748 (C. Callison-Burch Decl.) ¶ 88 ("Importantly, the purpose of the text extraction process used by OpenAI and other NLP practitioners is to identify text to *keep*, not text to *remove*. The scripts are designed to extract readable natural language text from HTML files for use as training data. Any loss of CMI during this process is an incidental byproduct of the text extraction methodology—it is not a purposeful or intentional removal of CMI."). OpenAI's employees and experts occasionally used "removal" as shorthand, but this characterization takes their use of this term as taken out of context. With respect to OpenAI's experts, Dr. Callison-Burch and Ms. Frederiksen-Cross used "removal" to describe content that was not selected for inclusion in the extractor's plain-text output. Neither usage establishes that OpenAI obtained a copy of an Asserted Work properly conveying its CMI and then removed or stripped that CMI from the copy. Rather, pursuant to Fed. R. Evid. 106, Ms. Frederiksen-Cross describes OpenAI's text extractors as applying a "selection and filtration process" that "ignore[s] and leave[s] behind" parts of HTML while selecting other content for plain-text output. News Ex. 1171 (B. Frederiksen-Cross Rebuttal Rep.) ¶¶ 39–41. Pursuant to Fed. R. Evid. 106, Dr. Callison-Burch similarly describes text extraction as a step

("OpenAI's text extractors … removed Plaintiffs' CMI"); Delorey Opening Rep. ¶ 248.	to "maximiz[e] the amount of meaningful training data" in training data created from postprocessed raw internet material. News Ex. 483 (C. Callison-Burch Opening Rep.) ¶ 359. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1305. OpenAI's text extractors operated by creating a new copy of the extracted work that is distinct from the copy from which the text was extracted. *See* Ex. 521 [Yuan 30(b)(1) Depo. Tr.] at 44:9-14 (agreeing that "after the text extraction process was complete, OpenAI now has two versions of the data, one in the raw form and one in the post-extraction form"); *see also* Wenger Opening Rep. at 49 ("OpenAI's text extractors both removed Plaintiffs' CMI and, in the process of removing that CMI, created new CMI-less copies of Plaintiffs' works."); Delorey Opening Rep. ¶ 242.	1305. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Disputed to the extent Plaintiffs characterize Yuan's testimony as establishing that the extracted text necessarily constitutes a new copy of an Asserted Work. Specifically, Yuan testified only that OpenAI has "two versions of the data, one in the raw form and one in the post-extraction form," not that the extracted text necessarily constitutes a new copy of an Asserted Work. News Ex. 521 (Q. Yuan 30(b)(1) Tr.) at 44:9–14. The disputed proposition is not supported by the cited evidence. Further, any contrary characterization by Plaintiffs' experts is disputed. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1306. After performing text extraction, OpenAI possesses two copies of the extracted work: one in the original, raw form, and another with text extracted. Ex. 521 [Yuan 30(b)(1) Depo. Tr.] at 44:9-14; Wenger Opening Rep. at 49.	1306. **Disputed in Part – Not Supported by Cited Evidence.** OpenAI responds as follows: Disputed to the extent Plaintiffs characterize Yuan's testimony as establishing that OpenAI possesses two distinct copies of an Asserted Work. Specifically, Yuan testified only that OpenAI has "two versions of the data, one in the raw form

	and one in the post-extraction form," not that the extracted text necessarily constitutes a new copy of an Asserted Work. News Ex. 521 (Q. Yuan 30(b)(1) Tr.) at 44:9–14. Further, any contrary characterization by Plaintiffs' experts is disputed. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1307. OpenAI used HTML-to-text extractors as "to extract natural language text from HTML content"; in particular, for GPT-3, OpenAI used publicly available versions of extraction algorithms called Newspaper3K and Dragnet435; for GPT-3.5, OpenAI used an OpenAI-developed implementation of the Newspaper3K algorithm; for GPT-4, OpenAI used an OpenAI-developed text extraction script called "gutentag" and the OpenAI-developed implementation of the Newspaper3k algorithm; and for GPT-4o, OpenAI used another OpenAI-developed version of its "gutentag" text extraction script. Goldstein Opening Rep. ¶ 536; Ex. 748 [OpenAI's Response to Daily News Plaintiffs' Interrogatory No. 17]; Ex. 756 [OpenAI's Supplemental Response to Intercept Interrogatory No. 3 (Dec. 31, 2025)] at 6; Ex. 752 [OpenAI Supplemental Response to Ziff Davis Interrogatory No. 20 (2/27/26)] at p. 14.	1307. **Undisputed.** Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1308. Each of OpenAI's text extractors operated to remove unwanted content, such as copyright management information. Goldstein Opening Rep. ¶¶ 138, 535-543;	1308. **Disputed.** OpenAI responds as follows: This fact is disputed because OpenAI's text extractors are not designed to "remove" CMI. Rather, they use general-

Goldstein Opening Source Code Rep., ¶¶ 5-41; Wenger Opening Rep. at 7 ("OpenAI used text extractors to remove Plaintiffs' CMI conveyed with Plaintiffs' asserted works before using those CMI-less copies to train its large language models"); Delorey Opening Rep. ¶ 248 ("OpenAI's extraction process is specifically designed to eliminate nearly all CMI (among other non-central or non-body text passages), while retaining nearly all central or body text").	purpose heuristics to selectively extract material that is suitable for model training. Pursuant to Fed. R. Evid. 106, not extracting data is not the same as removing CMI. *See, e.g.*, ECF 1748 (C. Callison-Burch Decl.) ¶ 88 ("Importantly, the purpose of the text extraction process used by OpenAI and other NLP practitioners is to identify text to *keep*, not text to *remove*. The scripts are designed to extract readable natural language text from HTML files for use as training data. Any loss of CMI during this process is an incidental byproduct of the text extraction methodology—it is not a purposeful or intentional removal of CMI."). Pursuant to Fed. R. Evid. 106, Plaintiffs' experts also overstate what the source code actually does, conflating the selective extraction of useful text with the removal of CMI. *See* News Ex. 1171 (B. Frederiksen-Cross Rebuttal Rep.) ¶¶ 27–28, 39–43, 74–87, 109–118. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1309. Dragnet is a publicly available text extraction program that removes certain blocks of a webpage and uses a machine learning algorithm to determine which blocks to remove. Wenger Reply Rep. at 69; Goldstein Opening Rep. ¶ 538; Goldstein Opening Source Code Rep., ¶¶ 16-17.	1309. **Disputed in Part.** OpenAI responds as follows: Dragnet is a publicly available text extraction program that uses machine learning to classify webpage blocks, but it does not remove blocks from the underlying webpage. Rather, it selectively extracts material that is suitable for model training while ignoring other blocks. Pursuant to Fed. R. Evid. 106, not extracting data is not the same as removing data. *See, e.g.*, ECF 1748 (C. Callison-Burch Decl.) ¶ 88 ("Importantly, the purpose of

	the text extraction process used by OpenAI and other NLP practitioners is to identify text to *keep*, not text to *remove*. The scripts are designed to extract readable natural language text from HTML files for use as training data. Any loss of CMI during this process is an incidental byproduct of the text extraction methodology—it is not a purposeful or intentional removal of CMI."); News Ex. 1171 (B. Frederiksen-Cross Rebuttal Rep.) ¶¶ 71–77. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1310. One factor the Dragnet algorithm considers is whether the 'id' or 'class' attribute contains certain words, including 'footer,' 'masthead,' and 'meta'. Wenger Reply Rep. at 69.	1310. **Undisputed.** Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1311. Inclusion of the words 'footer,' 'masthead,' or 'meta' in the block more likely to be removed by Dragnet. Wenger Reply Rep. at 69.	1311. **Disputed.** OpenAI responds as follows: The facts set forth in Paragraph 1311 are disputed. Specifically, Dragnet does not remove blocks from the underlying webpage. Rather, it selectively extracts material that is suitable for model training while ignoring other blocks. Pursuant to Fed. R. Evid. 106, not extracting data is not the same as removing data. *See, e.g.*, ECF 1748 (C. Callison-Burch Decl.) ¶ 88 ("Importantly, the purpose of the text extraction process used by OpenAI and other NLP practitioners is to identify text to *keep*, not text to *remove*. The scripts are designed to extract readable natural language text from HTML files for use as training data. Any loss of CMI during this process is an incidental

	byproduct of the text extraction methodology—it is not a purposeful or intentional removal of CMI."); News Ex. 1171 (B. Frederiksen-Cross Rebuttal Rep.) ¶¶ 71–77. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1312. Someone who had looked at web pages that were built with HTML would be familiar with the fact that there is often an element such as authorship and copyright information in the footer. Ex. 631 [Frederiksen-Cross Expert Depo. Tr.] at 107:9-17.	1312. **Undisputed.** Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1313. Dragnet removes author, title, copyright notice, and terms of use information because this information is typically contained in the fields that Dragnet is optimized to remove. Wenger Reply Rep. at 69-70.	1313. **Disputed.** OpenAI responds as follows: The facts set forth in Paragraph 1313 are disputed. Dragnet does not "remove" information from underlying webpages. Rather, it selectively extracts material that is suitable for model training. *See* News Ex. 1171 (B. Frederiksen-Cross Rebuttal Rep.) ¶¶ 71–77; ECF 1748 (C. Callison-Burch Decl.) ¶ 88. Not extracting data is not the same as removing data. Moreover, Plaintiffs improperly infer that certain information may typically appear in particular HTML elements that Dragnet categorically excludes that information. Dragnet uses multiple general-purpose heuristics to determine which blocks to extract and is not optimized to remove CMI. *See* News Ex. 1171 (B. Frederiksen-Cross Rebuttal Rep.) ¶¶ 71–77; ECF 1748 (C. Callison-Burch Decl.) ¶ 88.

	Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1314. News Plaintiffs' copyright notices and links to terms of use are contained in their <footer> tags, which favors removal by Dragnet. Wenger Reply Rep. at 69; Goldstein Opening Rep. ¶ 538; Goldstein Opening Source Code Rep., ¶¶ 16-17.	1314. **Disputed.** OpenAI responds as follows: The facts set forth in Paragraph 1314 are disputed. Plaintiffs improperly infer that because copyright notices and terms-of-use links appear within HTML `<footer>` tags, Dragnet favors their exclusion. But the relevant Dragnet heuristic assigns negative weight based on words appearing in an element's id or class attributes, not based on whether the element is a <footer> tag. An element with an id or class attribute containing the word "footer" need not be a <footer> element, and vice versa. *See* News Ex. 1171 (B. Frederiksen-Cross Rebuttal Rep.) ¶ 74. Moreover, Dragnet does not "remove" content from underlying webpages. Rather, it selectively extracts material that is suitable for model training. *See id.* ¶¶ 72–77; ECF 1748 (C. Callison-Burch Decl.) ¶ 88. Not extracting data is not the same as removing data. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1315. Dragnet's developers stated on Dragnet's website that it is designed to separate the "main article content" from "footers" and "copyright notices." Wenger Reply Rep. at 69-70; Matt McDonell, Benchmarking Python Content Extraction Algorithms: Dragnet, Readability, Goose, and Eatiht, at 1 (Jan. 29, 2015), https://moz.com/devblog/benchmarking-python-content-	1315. **Undisputed.** Microsoft responds that these facts do not pertain to claims asserted against Microsoft.

extraction-algorithms-dragnet-readability-goose-and-eatiht.	
1316. Newspaper—both the publicly available version and OpenAI's modified version of it—also removes CMI. Wenger Reply Rep. at 49; Goldstein Opening Source Code Rep., ¶¶ 6-15 and 18-24.	1316. **Disputed - Not Supported by Cited Evidence; Expert Opinion Subject to Rule 702.** OpenAI responds as follows: The facts set forth in Paragraph 1316 are disputed. The results of Dr. Wenger's CMI removal analysis are fundamentally flawed, as detailed in OpenAI's Opening Daubert motion. Specifically, there is nothing in the cited portion of Dr. Wenger's Reply Report that ties Newspaper to any of her alleged instances of bypassing CMI. *See* ECF 1858-6 (E. Wenger Decl. Ex. F) (Reply Rep.) at 49. To the extent the pincite to Wenger's Reply is wrong, this statement is still unsupported. At most, Dr. Wenger opines that both versions of Newspaper "omit tags where CMI is likely to reside." *Id.* at 71. That is not the same as removing CMI. Dr. Goldstein's Opening Source Code Report shares a similar issue. His report simply details specific tags which Newspaper targets and omits; there is no discussion of CMI in any of the portions of Goldstein's Opening Source Code Report that are cited. *See* A. Blanco Opp. Decl. Ex. 296 (T. Goldstein Opening Source Code Rep.) ¶¶ 6–15 and 18–24. Further disputed because OpenAI's text extractors do not "remove" anything. Rather, they extract material that is suitable for model training. Not extracting data is not equivalent to removing data. *See, e.g.*, ECF 1748 (C. Callison-Burch Decl.) ¶ 88 ("Importantly, the purpose of

	the text extraction process used by OpenAI and other NLP practitioners is to identify text to *keep*, not text to *remove*. The scripts are designed to extract readable natural language text from HTML files for use as training data. Any loss of CMI during this process is an incidental byproduct of the text extraction methodology—it is not a purposeful or intentional removal of CMI."). Newspaper is no exception to this general principle. *See* News Ex. 1171 (B. Frederiksen-Cross Rebuttal Rep.) ¶ 104 ("Neither version of Newspaper reflects an intent to remove CMI. Rather, both versions of Newspaper, which share the basic underlying approach to text extraction, intend to extract the main article text. Moreover, the OpenAI version of Newspaper is less restrictive, as it does not restrict extraction to a single node and thus can extract far more content from different parts of the page."); ¶ 78 (same). Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1317. Newspaper defaults to removing footer, article heading, and byline fields, all of which contain author, title, copyright notice, and terms of use information. Wenger Opening Rep. at 49; Goldstein Opening Source Code Rep., ¶¶ 6-15 and 18-24.	1317. **Disputed – Not Supported by Cited Evidence.** OpenAI responds as follows: This fact is disputed because it is not supported by the cited evidence. Specifically, Wenger repeatedly clarifies that these HTML tags "are common sources of CMI," that CMI is "likely to reside" in the tags, and "Plaintiffs' CMI is typically excluded." ECF 1858-1 (E. Wenger Decl. Ex. A) (Opening Rep.) 49.; ECF 1858-6 (E. Wenger Decl. Ex. F) (Reply Rep.) 71. However, she acknowledges that [redacted]

	[redacted] ECF 1858-1 (E. Wenger Decl. Ex. A) (Opening Rep.) 49. As such, OpenAI disputes that these four fields always contain this type of CMI. Similarly, Dr. Goldstein's Opening Source Code Report does not tie CMI to any of the fields he discusses in the cited paragraphs. *See* A. Blanco Opp. Decl. Ex. 296 (T. Goldstein Opening Source Code Rep.) ¶¶ 6–15 and 18–24. Finally, Ms. Frederiksen-Cross has further clarified that some footers are retained by Newspaper. News Ex. 1171 (B. Fredericksen-Cross Rebuttal Rep.) ¶¶ 81–82. Further disputed because OpenAI's text extractors do not "remove" anything. Rather, they extract material that is suitable for model training. Not extracting data is not equivalent to removing data. *See, e.g.*, ECF 1748 (C. Callison-Burch Decl.) ¶ 88 ("Importantly, the purpose of the text extraction process used by OpenAI and other NLP practitioners is to identify text to keep, not text to remove. The scripts are designed to extract readable natural language text from HTML files for use as training data. Any loss of CMI during this process is an incidental byproduct of the text extraction methodology—it is not a purposeful or intentional removal of CMI."). Newspaper is no exception to this general principle. *See* News Ex. 1171 (B. Frederiksen-Cross Rebuttal Rep.) ¶ 104 ("Neither version of Newspaper reflects an intent to remove CMI. Rather, both versions of Newspaper, which share the basic underlying approach to text extraction, intend to extract the main article text. Moreover, the OpenAI version of

	Newspaper is less restrictive, as it does not restrict extraction to a single node and thus can extract far more content from different parts of the page."); ¶ 78 (same). Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1318. Newspaper then selects a "best node" from a website's HTML and extract the text portion of that best node and its sibling nodes, discarding the rest. Ex. 1171 [Frederiksen-Cross Rebuttal Rep] ¶ 79; Wenger Reply Rep. at 70; Goldstein Opening Source Code Rep., ¶¶ 6-15 and 18-24.	1318. **Disputed in Part.** OpenAI responds as follows: Undisputed that some versions of Newspaper select a "best node." However, Plaintiffs characterization of Newspaper's selective extraction of text as "discarding" the remainder of a webpage is disputed. Specifically, Plaintiffs rely on Frederiksen-Cross's description of only the publicly available version of Newspaper, which selects a "best" node and its siblings, ignores negatively scored child nodes, and extracts the selected text. OpenAI's modified version, however, can extract text from multiple unrelated nodes, rather than limiting extraction to a single best node and its siblings. *See* News Ex. 1171 (B. Frederiksen-Cross Rebuttal Rep. (News)) ¶¶ 79, 95, 103–104. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1319. To select a "best node" for extraction, Newspaper only considers nodes with tags p, pre, and td. Ex. 1171 [Frederiksen-Cross Rebuttal Rep.] ¶ 89; Wenger Reply Rep. at 70.	1319. **Undisputed.** Microsoft responds that these facts do not pertain to claims asserted against Microsoft.

1320. Newspaper also removes nodes with high link density. Ex. 1171 [Frederiksen-Cross Rebuttal Rep.] ¶¶ 90, 95; Wenger Reply Rep. at 71.	1320. **Disputed.** OpenAI responds as follows: OpenAI's modified version of Newspaper [redacted] *See* News Ex. 1171 (B. Frederiksen-Cross Rebuttal Rep.) ¶¶ 90, 96–97. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1321. p, pre, and td tags typically do not contain News Plaintiffs' author, title, copyright notice, or terms of use information. Wenger Reply Rep. at 71.	1321. **Disputed in Part - Expert Opinion Subject to Rule 702.** OpenAI responds as follows: Undisputed that Dr. Wenger's report makes the referenced representation. Defendants' pending Rule 702 motion seeks to exclude certain opinions of Dr. Wenger, on which this statement relies. To the extent the Court excludes the challenged opinions, Plaintiffs may not rely on them to support this statement. *See* Fed. R. Civ. P. 56(c)(2). Further, Dr. Wenger only performed this analysis with respect to The Intercept and CIR; her analysis cannot be generalized to all of News Plaintiffs. *See* ECF 1858-6 (E. Wenger Decl. Ex. F) (Reply Rep.) 71. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1322. News Plaintiffs' CMI was by definition excluded from Newspaper's search for a best node. Wenger Reply Rep. at 71.	1322. **Disputed - Expert Opinion Subject to Rule 702.** OpenAI responds as follows: Undisputed that Dr. Wenger's report makes the referenced representation. Defendants' pending Rule 702 motion seeks to exclude certain opinions of Dr. Wenger, on which

	this statement relies. To the extent the Court excludes the challenged opinions, Plaintiffs may not rely on them to support this statement. *See* Fed. R. Civ. P. 56(c)(2). Further, these facts are disputed because Dr. Wenger opines that "CMI is typically excluded," but nowhere says that it is categorically excluded. *See* ECF 1858-6 (E. Wenger Decl. Ex. F) (Reply Rep.) 71. Additionally, Dr. Wenger only performed this analysis with respect to The Intercept and CIR; her analysis cannot be generalized to all of News Plaintiffs. *See id.* Also disputed to the extent this implies that News Plaintiffs' CMI was removed from their Asserted Works. Newspaper does not "remove" anything. Rather, it extracts the "best" node that is most likely to contain text suitable for training. Extracting nodes is not equivalent to removing data. *See, e.g.*, ECF 1748 (C. Callison-Burch Decl.) ¶ 88 ("Importantly, the purpose of the text extraction process used by OpenAI and other NLP practitioners is to identify text to *keep*, not text to *remove*. The scripts are designed to extract readable natural language text from HTML files for use as training data. Any loss of CMI during this process is an incidental byproduct of the text extraction methodology—it is not a purposeful or intentional removal of CMI."); News Ex. 1171 (B. Fredericksen-Cross Rebuttal Rep. (News)) ¶¶ 93–94 ("In summary, the open-source version of Newspaper uses stop word and link density heuristics, in conjunction with relative placement on the HTML page, to identify the single best node on a page, then extracts the content from that node. These heuristics are

	not directed toward the removal of CMI but rather to the identification of parts of a page that are especially likely to contain a continuous passage of text."). Finally, Ms. Frederiksen-Cross has further clarified that "Wenger's characterization . . . is overly broad." News Ex. 1171 (B. Fredericksen-Cross Rebuttal Rep. (News)) ¶ 81. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1323. Sibling nodes that might contain CMI are typically excluded by Newspaper due to other characteristics they possess, such as high link density. Wenger Reply Rep. at 71.	1323. **Disputed in Part.** OpenAI responds as follows: The facts set forth in Paragraph 1323 are disputed because they do not take into account OpenAI's fork of Newspaper, [redacted] News Ex. 1171 (B. Fredericksen-Cross Rebuttal Rep. (News)) ¶ 97. This is particularly important because OpenAI's fork also [redacted] *Id.* ¶ 103. As such, even if Dr. Wenger may be correct about this for the generic Newspaper, she fails to consider the specific alterations made to OpenAI's fork. Further disputed to the extent this implies that News Plaintiffs' CMI was removed from their Asserted Works. Newspaper does not "remove" anything. Rather, it extracts the "best" node that is most likely to contain text suitable for training. Extracting nodes is not equivalent to removing data. *See, e.g.*, ECF 1748 (C. Callison-Burch Decl.) ¶ 88 ("Importantly, the purpose of the text extraction

	process used by OpenAI and other NLP practitioners is to identify text to *keep*, not text to *remove*. The scripts are designed to extract readable natural language text from HTML files for use as training data. Any loss of CMI during this process is an incidental byproduct of the text extraction methodology—it is not a purposeful or intentional removal of CMI."); News Ex. 1171 (B. Fredericksen-Cross Rebuttal Rep. (News)) ¶¶ 93–94 ("In summary, the open-source version of Newspaper uses stop word and link density heuristics, in conjunction with relative placement on the HTML page, to identify the single best node on a page, then extracts the content from that node. These heuristics are not directed toward the removal of CMI but rather to the identification of parts of a page that are especially likely to contain a continuous passage of text."). Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1324. Gutentag removed CMI. Wenger Reply Rep. at 51-53; Goldstein Opening Source Code Rep. ¶¶ 25-41; Delorey Opening Rep. ¶¶ 249-252, 271, 275.	1324. **Disputed.** OpenAI responds as follows: Mr. Delorey's opinion has two important qualifiers. First, it is only "[t]o the extent CMI generally occurs in the 'leading and trailing' text on webpages." ECF 1774-1 (D. Delorey Decl. Ex. A) (Opening Report) ¶ 250. He provides no evidence that this is the case. Second, even if this were true, Mr. Delorey also clarifies that this CMI would not always be skipped during extraction. *Id*. The cited paragraphs in Wenger's report do not provide any support for this fact.

	Further, this is disputed to the extent this implies that Gutentag does not extract all types of CMI all the time. In fact, even using Delorey's overinclusive CMI removal methodology (as detailed in OpenAI's 702 Motion), Delorey only alleged 1% of the asserted works titles were not extracted. *Id.* ¶ 271. He does not allege that any type of CMI were not extracted 100% of the time. *Id.* Wenger and Goldstein's independent, yet overinclusive, methodologies likewise do not allege that any CMI was not extracted in every instance. *See* ECF 1858-8 (E. Wenger Decl. Ex. H) (Supplemental Rep.) 17-18 (finding tens of thousands of instances where CMI was not removed in alleged copies). Also disputed because OpenAI's text extractors do not "remove" anything. Rather, they extract material that is suitable for model training. Not extracting data is not equivalent to removing data. *See, e.g.*, ECF 1748 (C. Callison-Burch Decl.) ¶ 88 ("Importantly, the purpose of the text extraction process used by OpenAI and other NLP practitioners is to identify text to *keep*, not text to *remove*. The scripts are designed to extract readable natural language text from HTML files for use as training data. Any loss of CMI during this process is an incidental byproduct of the text extraction methodology—it is not a purposeful or intentional removal of CMI."). Microsoft responds that these facts do not pertain to claims asserted against Microsoft.

1325. [redacted] Ex. 1171 [Frederiksen-Cross Rebuttal Rep.] ¶ 129; Delorey Opening Rep. at ¶ 244-247.	1325. **Undisputed.** Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1326. The footer node is typically where copyright notice and links to terms of use are conveyed. Wenger Reply Rep. at 49.	1326. **Disputed in Part.** OpenAI responds as follows: The facts set forth in Paragraph 1326 are disputed because they are not supported by the cited evidence. Specifically, the cited page of Wenger's Reply Report does not provide any support for this assertion. Wenger only opines about where The Intercept and CIR's copyright notices and links to terms of use are located. Wenger did not provide any support for this fact for any of the other News Plaintiffs. *See* ECF 1858-1 (E. Wenger Decl. Ex. A) (Opening Rep.) 53 (describing the location of terms of use/terms of service for Mother Jones, Reveal, and The Intercept). Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1327. [redacted] Ex.	1327. **Undisputed.** Microsoft responds that these facts do not pertain to claims asserted against Microsoft.

1171 [Frederiksen-Cross Rebuttal Rep.] ¶¶ 109-11; Delorey Opening Rep. ¶ 247.	
1328. [redacted] Ex. 1171 [Frederiksen-Cross Rebuttal Rep.] ¶ 112.	1328. **Undisputed.** Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1329. The terms "copyright" "©," and "reserved" are typically contained in copyright notices, including News Plaintiffs'. Wenger Reply Rep. at 78.	1329. **Disputed.** OpenAI responds as follows: The facts set forth in Paragraph 1329 are disputed because they are not supported by the cited evidence. Specifically, Dr. Wenger's report does not make any assertion about whether these three phrases or symbols are "typical" in News Plaintiffs' copyright notices. For one, Dr. Wenger only examined copyright notices for The Intercept and CIR, so even if Dr. Wenger's analysis did speak to whether these are "typical," her opinion does not apply to all other News Plaintiffs. Further, this portion of Dr. Wenger's Reply Report discusses the token length of different CMI strings, not the prevalence of different terms in The Intercept and CIR's copyright notices. Finally, other portions of Dr. Wenger's reports only classify these terms as "CMI-related," ECF 1858-6 (E. Wenger Decl. Ex. F) (Reply Rep.) at 71, but again do not discuss the prevalence with which they are included in copyright notices for The Intercept and CIR. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.

1330. Gutentag removes copyright-related terms only when they appear in copyright notices, while retaining them as part of the main article text. For example, OpenAI produced at OPCO_NEWS_1144091 excerpts from the training set [redacted] *See* Ex. 491 [OpenAI's Supplemental Response to CIR Interrogatory No. 14 (Mar. 23, 2026)] at 6. That training set was used to train [redacted] *see id.*, and thus was extracted using Gutentag, *see* Ex. 756 [OpenAI's Supplemental Response to Intercept Interrogatory No. 3 (Dec. 31, 2025)] at 6; Ex. 752 [OpenAI's Supplemental Response to Ziff Davis Interrogatory No. 20 (January 27, 2026)] p. 14. That production includes a copy of a CIR article entitled "Obama Picks Copyright Lawyer Don Verrilli for Top Legal Post." The copy does not contain CIR's copyright notice (or any other CMI, including the title just referenced), but does include eight uses of the word "copyright" in the main text of the article, such as its description of Don Verrilli as having "earned [] the ire of copyright reform supporters." Ex. 967 [OPCO_NEWS_1144091], at http://www.motherjones.com/mojo/2011/01/obama-picks-copyright-lawyer-don-verrilli-top-legal-post.	1330. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 1330 are disputed because they are not supported by the cited evidence. Plaintiffs improperly infer from a single training-data excerpt that Gutentag selectively targets copyright notices while retaining copyright-related terms elsewhere. Specifically, Gutentag does not remove copyright-related terms from underlying webpages. Rather, it selectively extracts text using general-purpose heuristics designed to identify useful text and exclude boilerplate. *See, e.g.*, ECF 1748 (C. Callison-Burch Decl.) ¶ 88. [redacted] Gutentag does not identify or target copyright notices as such. *See* News Ex. 1171 (B. Frederiksen-Cross Rebuttal Rep.) ¶¶ 105, 108–118, 135–137. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1331. As another example, OpenAI produced at OPCO_INTERCEPT_0000003 excerpts from the training set [redacted] *See* Ex. 1156 [OpenAI's Supplemental Response to Intercept Interrogatory No. 7 (Mar. 23, 2026)] at 6. That training set was used to train [redacted] *see id.*, and thus was extracted using Gutentag, *see*	1331. **Disputed in Part.** OpenAI responds as follows: Undisputed that OpenAI produced an excerpt at OPCO_INTERCEPT_0000003 from the dataset [redacted] News Ex. 490 (OpenAI's Supplemental Response to Intercept Interrogatory No. 7) at 6. However, Plaintiffs mis-identify the dataset as [redacted]

Ex. 756 [OpenAI's Supplemental Response to Intercept Interrogatory No. 3 (Dec. 31, 2025)] at 6; Ex. 752 [OpenAI's Supplemental Response to Ziff Davis Interrogatory No. 20 (January 27, 2026)] p. 14. That production includes a copy of an Intercept article entitled "NSA Broke the Encryption on File-Sharing Apps Kazaa and EDonkey." The copy does not contain The Intercept's copyright notice (or any other CMI, including the title just referenced), but does contain the phrase "NSA didn't care about violations of copyright law." Ex. 989 [OPCO_INTERCEPT_0000003], at https://theintercept.com/2017/09/13/nsa-broke-the-encryption-on-file-sharing-apps-kazaa-and-edonkey/.	[redacted] Moreover, Plaintiffs do not explain what they mean by "as another example." The disputed proposition is not supported by the cited evidence. Specifically, to the extent Plaintiffs suggest that this excerpt demonstrates that Gutentag selectively targets copyright notices while retaining copyright-related terms elsewhere, OpenAI disputes that characterization. *See* News Ex. 1171 (B. Frederiksen-Cross Rebuttal Rep.) ¶¶ 108–118, 135–137. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1332. Gutentag is designed to remove bylines. Wenger Opening Rep. at 51; Goldstein Opening Source Code Report, ¶¶ 25-41; Delorey Opening Rep. ¶ 245.	1332. **Disputed in Part.** OpenAI responds as follows: Undisputed that Gutentag generally does not extract bylines. Disputed that OpenAI removes bylines. OpenAI's text extractors do not "remove" anything. Rather, they extract material that is suitable for model training. Not extracting data is not equivalent to removing data. See, e.g., ECF 1748 (C. Callison-Burch Decl.) ¶ 88 ("Importantly, the purpose of the text extraction process used by OpenAI and other NLP practitioners is to identify text to *keep*, not text to *remove*. The scripts are designed to extract readable natural language text from HTML files for use as training data. Any loss of CMI during this process is an incidental byproduct of the text extraction methodology—it is not a purposeful or intentional removal of CMI.").

	Further disputed, to the extent this implies something about OpenAI's state of mind when it developed Gutentag or when any CMI allegedly was bypassed during text extraction. Any alleged CMI bypassing is part of OpenAI's data processing pipeline to curate datasets that are representative of human language. *See* News Ex. 676 (C. Callison-Burch Rebuttal Rep.) ¶¶ 77–81 ("Data Cleaning is a Technical Necessity"), News Ex. 1171 (B. Frederiksen-Cross Rebuttal Rep.) ¶ 28 (describing the purpose of this processing as "maximizing the overall amount of useful, human-written text extracted from a webpage while minimizing the incidence of spam, Search Engine Optimization ('SEO'), [or] boilerplate"). Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1333. Dragnet is designed to remove author information. Wenger Opening Rep. at 7, 48-53; Wenger Rebuttal Rep. at 21-22; Wenger Reply Rep. at 66-70.	1333. **Disputed - Not Supported By Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 1333 are disputed. Wenger's reports only support this Paragraph to the extent that she opines (1) Dragnet is designed to "extract only the 'main article content,'" (2) Plaintiffs' generally put author information in the titles, and (3) titles are more likely to be bypassed during the extraction process than other portions of a page. *See, e.g.*, ECF 1858-1 (E. Wenger Decl. Ex. A) (Opening Rep.) at 52. Wenger at no point claims that Dragnet was specifically designed to remove author

	information. Additionally, Wenger only states that CMI may be "more likely" to be bypassed. Further, OpenAI disputes that Dragnet removes anything. OpenAI's text extractors do not "remove" anything. Rather, they extract material that is suitable for model training. Not extracting data is not equivalent to removing data. *See, e.g.,* ECF 1748 (C. Callison-Burch Decl.) ¶ 88 ("Importantly, the purpose of the text extraction process used by OpenAI and other NLP practitioners is to identify text to *keep*, not text to *remove*. The scripts are designed to extract readable natural language text from HTML files for use as training data. Any loss of CMI during this process is an incidental byproduct of the text extraction methodology—it is not a purposeful or intentional removal of CMI."). Further disputed, to the extent this implies something about OpenAI's state of mind when it chose to use Dragnet or when any CMI allegedly was bypassed during text extraction. Any alleged CMI bypassing is part of OpenAI's data processing pipeline to curate datasets that are representative of human language. *See* News Ex. 676 (C. Callison-Burch Rebuttal Rep.) ¶¶ 77–81 ("Data Cleaning is a Technical Necessity"), News Ex. 1171 (B. Frederiksen-Cross Rebuttal Rep.) ¶ 28 (describing the purpose of this processing as "maximizing the overall amount of useful, human-written text extracted from a webpage while minimizing the incidence of spam, Search Engine Optimization ('SEO'), [or] boilerplate").

	Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1334. Dragnet is designed to remove title information. Wenger Opening Rep. at 7, 48-53; Wenger Rebuttal Rep. at 21-22; Wenger Reply Rep. at 66-70.	1334. **Disputed.** OpenAI responds as follows: The facts set forth in Paragraph 1334 are disputed. Wenger's reports only support this Paragraph to the extent that she opines (1) Dragnet is designed to "extract only the 'main article content'" and (2) blocks that contain "masthead" are more likely to be removed. *See* ECF 1858-1 (E. Wenger Decl. Ex. A) (Opening Rep.) at 52; ECF 1858-6 (E. Wenger Decl. Ex. F.) (Reply Rep.) at 69. Wenger at no point claims that Dragnet was specifically designed to remove title information. Additionally, Wenger only states that certain factors may make CMI "more likely" to be bypassed. Further, OpenAI disputes that Dragnet removes anything. OpenAI's text extractors do not "remove" anything. Rather, they extract material that is suitable for model training. Not extracting data is not equivalent to removing data. *See, e.g.,* ECF 1748 (C. Callison-Burch Decl.) ¶ 88 ("Importantly, the purpose of the text extraction process used by OpenAI and other NLP practitioners is to identify text to *keep*, not text to *remove*. The scripts are designed to extract readable natural language text from HTML files for use as training data. Any loss of CMI during this process is an incidental byproduct of the text extraction methodology—it is not a purposeful or intentional removal of CMI."). Further disputed, to the extent this implies something about OpenAI's state of mind when it chose to

	use Dragnet or when any CMI allegedly was bypassed during text extraction. Any alleged CMI bypassing is part of OpenAI's data processing pipeline to curate datasets that are representative of human language. *See* News Ex. 676 (C. Callison-Burch Rebuttal Rep.) ¶¶ 77–81 ("Data Cleaning is a Technical Necessity"), News Ex. 1171 (B. Frederiksen-Cross Rebuttal Rep.) ¶ 28 (describing the purpose of this processing as "maximizing the overall amount of useful, human-written text extracted from a webpage while minimizing the incidence of spam, Search Engine Optimization ('SEO'), [or] boilerplate"). Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1335. Dragnet is designed to remove copyright notice information. Wenger Opening Rep. at 7, 48-53; Wenger Rebuttal Rep. at 21-22; Wenger Reply Rep. at 66-70; Goldstein Opening Source Code Rep. ¶¶ 16-17.	1335. **Disputed.** OpenAI responds as follows: The facts set forth in Paragraph 1335 are disputed. Wenger and Goldstein rely on the Dragnet public README which says that it was designed to remove copyright notices, rather than examining the source code itself. *See* ECF 1858-1 (E. Wenger Decl. Ex. A) (Opening Rep.) 7, 48–53; ECF 1858-4 (E. Wenger Decl. Ex. D) (Rebuttal Rep.) 21–22; ECF 1858-6 (E. Wenger Decl. Ex. F) (Reply Rep.) 66–70; A. Blanco Opp. Decl. Ex. 296 (T. Goldstein Opening Source Code Rep.) ¶¶ 16–17. Further, OpenAI disputes that Dragnet removes anything. OpenAI's text extractors do not "remove" anything. Rather, they extract material that is suitable for model training. Not extracting data is not equivalent to

	removing data. *See, e.g.*, ECF 1748 (C. Callison-Burch Decl.) ¶ 88 ("Importantly, the purpose of the text extraction process used by OpenAI and other NLP practitioners is to identify text to *keep*, not text to *remove*. The scripts are designed to extract readable natural language text from HTML files for use as training data. Any loss of CMI during this process is an incidental byproduct of the text extraction methodology—it is not a purposeful or intentional removal of CMI."). Further disputed, to the extent this implies something about OpenAI's state of mind when it chose to use Dragnet or when any CMI allegedly was bypassed during text extraction. Any alleged CMI bypassing is part of OpenAI's data processing pipeline to curate datasets that are representative of human language. *See* News Ex. 676 (C. Callison-Burch Rebuttal Rep.) ¶¶ 77-81 ("Data Cleaning is a Technical Necessity"), News Ex. 1171 (B. Frederiksen-Cross Rebuttal Rep.) ¶ 28 (describing the purpose of this processing as "maximizing the overall amount of useful, human-written text extracted from a webpage while minimizing the incidence of spam, Search Engine Optimization ('SEO'), [or] boilerplate"). Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1336. Dragnet is designed to remove terms of use information. Wenger Opening Rep. at 7, 48-53; Wenger	1336. **Disputed.** OpenAI responds as follows: The facts set forth in Paragraph 1336 are disputed. Wenger's reports only support this to the extent that she opines (1)

Rebuttal Rep. at 21-22; Wenger Reply Rep. at 66-70; Goldstein Opening Source Code Rep. ¶¶ 16-17.	Dragnet is designed to "extract only the 'main article content[,]'" (2) Plaintiffs' generally put terms of use information in footers (id.), and "footers" are more likely to be bypassed. ECF 1858-1 (E. Wenger Decl. Ex. A) (Opening Rep.) at 52; ECF 1858-6 (E. Wenger Decl. Ex. F) (Reply Rep.) at 69. Wenger at no point claims that Dragnet was specifically designed to remove terms of use information. Additionally, Wenger only states that certain factors may make CMI "more likely" to be bypassed, but simply being in a footer itself does not result in the information being bypassed 100% of the time. Further, OpenAI disputes that Dragnet removes anything. OpenAI's text extractors do not "remove" anything. Rather, they extract material that is suitable for model training. Not extracting data is not equivalent to removing data. *See, e.g.*, ECF 1748 (C. Callison-Burch Decl.) ¶ 88 ("Importantly, the purpose of the text extraction process used by OpenAI and other NLP practitioners is to identify text to *keep*, not text to *remove*. The scripts are designed to extract readable natural language text from HTML files for use as training data. Any loss of CMI during this process is an incidental byproduct of the text extraction methodology—it is not a purposeful or intentional removal of CMI."). Further disputed, to the extent this implies something about OpenAI's state of mind when it chose to use Dragnet or when any CMI allegedly was bypassed during text extraction. Any alleged CMI bypassing is part of OpenAI's data processing pipeline to curate datasets that

	are representative of human language. *See* Callison-Burch Rebuttal Report ¶¶ 77-81 ("Data Cleaning is a Technical Necessity"), Frederiksen-Cross Rebuttal Report ¶ 28 (describing the purpose of this processing as "maximizing the overall amount of useful, human-written text extracted from a webpage while minimizing the incidence of spam, Search Engine Optimization ('SEO'), [or] boilerplate"). Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1337. Newspaper is designed to remove author information. Wenger Opening Rep. at 7, 48-53; Wenger Rebuttal Rep. at 21-22; Wenger Reply Rep. at 66-71; Goldstein Opening Source Code Rep. ¶¶ 6-15 and 18-24.	1337. **Disputed.** OpenAI responds as follows: The facts set forth in Paragraph 1337 are disputed. Dr. Wenger and Dr. Goldstein's opinions focus on HTML tags which Newspaper allegedly bypassed during extraction. *See* ECF 1858-1 (E. Wenger Decl. Ex. A) (Opening Rep.) 7, 48–53; ECF 1858-4 (E. Wenger Decl. Ex. D) (Rebuttal Rep.) 21–22; ECF 1858-6 (E. Wenger Decl. Ex. F) (Reply Rep.) 66–71; A. Blanco Opp. Decl. Ex. 296 (T. Goldstein Opening Source Code Rep.) ¶¶ 6–15 and 18–24. Even if those tags are bypassed, Plaintiffs' experts do not provide evidence that these tags always contain Plaintiffs' author information. Dr. Wenger simply opines, without substantiation, that these are where "Plaintiffs' CMI is typically found." ECF 1858-6 (E. Wenger Decl. Ex. F) (Reply Rep.) 70–71. [redacted] ECF 1858-1 (E. Wenger Decl. Ex. A)

	(Opening Rep.) 49. Dr. Wenger's reliance on her CMI removal analysis to support these opinions (*see* ECF 1858-6 (E. Wenger Decl. Ex. F) (Reply Rep.) 68) is further disputed because of the unsound methodologies used, as detailed in OpenAI's Opening Daubert Motion. In fact, Dr. Wenger even admits that she sometimes found News Plaintiffs' CMI in the body of the HTML page, which was retained by OpenAI. *Id.* at 71. Additionally, Paragraph 1337 does not specify whether this assertion applies to the public version of Newspaper, OpenAI's internal fork of it, or both. Further, OpenAI disputes that Newspaper removes anything. OpenAI's text extractors do not "remove" anything. Rather, they extract material that is suitable for model training. Not extracting data is not equivalent to removing data. *See, e.g.,* ECF 1748 (C. Callison-Burch Decl.) ¶ 88 ("Importantly, the purpose of the text extraction process used by OpenAI and other NLP practitioners is to identify text to *keep*, not text to *remove*. The scripts are designed to extract readable natural language text from HTML files for use as training data. Any loss of CMI during this process is an incidental byproduct of the text extraction methodology—it is not a purposeful or intentional removal of CMI."). Further disputed, to the extent this implies something about OpenAI's state of mind when it chose to use Newspaper or when any CMI allegedly was bypassed during text extraction. Any alleged CMI bypassing is part of OpenAI's data processing pipeline to curate datasets that

	are representative of human language. *See* News Ex. 676 (C. Callison-Burch Rebuttal Rep.) ¶¶ 77–81 ("Data Cleaning is a Technical Necessity"), News Ex. 1171 (B. Frederiksen-Cross Rebuttal Rep.) ¶ 28 (describing the purpose of this processing as "maximizing the overall amount of useful, human-written text extracted from a webpage while minimizing the incidence of spam, Search Engine Optimization ('SEO'), [or] boilerplate"). Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1338. Newspaper is designed to remove title information. Wenger Opening Rep. at 7, 48-53; Wenger Rebuttal Rep. at 21-22; Wenger Reply Rep. at 66-71; Goldstein Opening Source Code Rep. ¶¶ 6-15 and 18-24.	1338. **Disputed– Not Supported by Cited Evidence; Expert Opinion Subject to Rule 702.** OpenAI responds as follows: The facts set forth in Paragraph 1338 are disputed. Dr. Wenger and Dr. Goldstein's opinions focus on HTML tags which Newspaper allegedly bypassed during extraction. *See* ECF 1858-1 (E. Wenger Decl. Ex. A.) (Opening Rep.) 7, 48–53; ECF 1858-4 (E. Wenger Decl. Ex. D) (Rebuttal Rep.) 21–22; ECF 1858-6 (E. Wenger Decl. Ex. F) (Reply Rep.) 66–71; A. Blanco Opp. Decl. Ex. 296 (T. Goldstein Opening Source Code Rep.) ¶¶ 6–15 and 18–24. Even if those tags are bypassed, Plaintiffs' experts do not provide evidence that these tags always contain Plaintiffs' title information. Dr. Wenger simply opines, without substantiation, that these are where "Plaintiffs' CMI is typically found." ECF 1858-6 (E. Wenger Decl. Ex. F) (Reply Rep.) 70–71. [redacted]

	[redacted] ECF 1858-1 (E. Wenger Decl. Ex. A) (Opening Rep.) 49. Dr. Wenger's reliance on her CMI removal analysis to support these opinions (*see* ECF 1858-6 (E. Wenger Decl. Ex. F) at68) is further disputed because of the unsound methodologies used, as detailed in OpenAI's Opening Daubert Motion. In fact, Dr. Wenger even admits that she sometimes found News Plaintiffs' CMI in the body of the HTML page, which was retained by OpenAI. *Id*. at 71. Additionally, Paragraph 1338 does not specify whether this assertion applies to the public version of Newspaper, OpenAI's internal fork of it, or both. Further, OpenAI disputes that Newspaper removes anything. OpenAI's text extractors do not "remove" anything. Rather, they extract material that is suitable for model training. Not extracting data is not equivalent to removing data. *See, e.g.,* ECF 1748 (C. Callison-Burch Decl.) ¶ 88 ("Importantly, the purpose of the text extraction process used by OpenAI and other NLP practitioners is to identify text to *keep*, not text to *remove*. The scripts are designed to extract readable natural language text from HTML files for use as training data. Any loss of CMI during this process is an incidental byproduct of the text extraction methodology—it is not a purposeful or intentional removal of CMI."). Further disputed, to the extent this implies something about OpenAI's state of mind when it chose to use Newspaper or when any CMI allegedly was bypassed

	during text extraction. Any alleged CMI bypassing is part of OpenAI's data processing pipeline to curate datasets that are representative of human language. *See* News Ex. 676 (C. Callison-Burch Rebuttal Rep.) ¶¶ 77–81 ("Data Cleaning is a Technical Necessity"), News Ex. 1171 (B. Frederiksen-Cross Rebuttal Rep.) ¶ 28 (describing the purpose of this processing as "maximizing the overall amount of useful, human-written text extracted from a webpage while minimizing the incidence of spam, Search Engine Optimization ('SEO'), [or] boilerplate"). Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1339. Newspaper is designed to remove copyright notice information. Wenger Opening Rep. at 7, 48-53; Wenger Rebuttal Rep. at 21-22; Wenger Reply Rep. at 66-71; Goldstein Opening Source Code Rep. ¶¶ 6-15 and 18-24.	1339. **Disputed– Not Supported by Cited Evidence; Expert Opinion Subject to Rule 702.** OpenAI responds as follows: The facts set forth in Paragraph 1339 are disputed. Dr. Wenger and Dr. Goldstein's opinions focus on HTML tags which Newspaper allegedly bypasses during text extraction. *See* ECF 1858-1 (E. Wenger Decl. Ex. A) (Opening Rep.) at 7, 48–53; ECF 1858-4 (E. Wenger Decl. Ex. D) (Rebuttal Rep.) at 21–22; ECF 1858-6 (E. Wenger Decl. Ex. F) Reply Rep. at 66–71; A. Blanco Opp. Decl. Ex. 296 (T. Goldstein Opening Source Code Rep.) ¶¶ 6–15 and 18–24. Even if those tags are bypassed, Plaintiffs' experts do not provide evidence that these tags always contain Plaintiffs' copyright notice information. Dr. Wenger simply opines, without substantiation, that these are where "Plaintiffs' CMI is

	typically found." ECF 1858-6 (E. Wenger Decl. Ex. F) (Reply Rep.) at 70–71. [redacted] ECF 1858-1 (E. Wenger Decl. Ex. A) (Opening Rep.) at 49. Both this and her further reliance on her CMI removal analysis to support these opinions (*see* 1858-6 (E. Wenger Decl. Ex. F) (Reply Rep.) at 68), are further disputed because of the unsound methodologies used, as detailed in OpenAI's Opening Daubert Motion. In fact, Dr. Wenger even admits that she sometimes found News Plaintiffs' CMI in the body of the HTML page, which was retained by OpenAI. *Id.* at 71. Additionally, Paragraph 1339 does not specify whether this assertion applies to the public version of Newspaper, OpenAI's internal fork of it, or both. Further, OpenAI disputes that Newspaper removes anything. OpenAI's text extractors do not "remove" anything. Rather, they extract material that is suitable for model training. Not extracting data is not equivalent to removing data. *See, e.g.,* ECF 1748 (C. Callison-Burch Decl.) ¶ 88 ("Importantly, the purpose of the text extraction process used by OpenAI and other NLP practitioners is to identify text to *keep*, not text to *remove*. The scripts are designed to extract readable natural language text from HTML files for use as training data. Any loss of CMI during this process is an incidental byproduct of the text extraction methodology—it is not a purposeful or intentional removal of CMI.").

	Further disputed, to the extent this implies something about OpenAI's state of mind when it chose to use Newspaper or when any CMI allegedly was bypassed during text extraction. Any alleged CMI bypassing is part of OpenAI's data processing pipeline to curate datasets that are representative of human language. *See* News Ex. 676 (C. Callison-Burch Rebuttal Rep.) ¶¶ 77–81 ("Data Cleaning is a Technical Necessity"), News Ex. 1171 (B. Frederiksen-Cross Rebuttal Rep.) ¶ 28 (describing the purpose of this processing as "maximizing the overall amount of useful, human-written text extracted from a webpage while minimizing the incidence of spam, Search Engine Optimization ('SEO'), [or] boilerplate"). Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1340. Newspaper is designed to remove terms of use information. Wenger Opening Rep. at 7, 48-53; Wenger Rebuttal Rep. at 21-22; Wenger Reply Rep. at 66-71; Goldstein Opening Source Code Rep. ¶¶ 6-15 and 18-24.	1340. **Disputed– Not Supported by Cited Evidence; Expert Opinion Subject to Rule 702.** OpenAI responds as follows: The facts set forth in Paragraph 1340 are disputed. Dr. Wenger and Dr. Goldstein's opinions focus on HTML tags which Newspaper allegedly bypasses during text extraction. ECF 1858-1 (E. Wenger Decl. Ex. A) (Opening Rep.) at 7, 48–53; ECF 1858-4 (E. Wenger Decl. Ex. D) (Rebuttal Rep.) at 21–22; ECF 1858-6 (E. Wenger Decl. Ex. F) (Reply Rep.) at 66–71; A. Blanco Opp. Decl. Ex. 296 (T. Goldstein Opening Source Code Rep.) ¶¶ 6–15 and 18–24. Even if those tags are bypassed, Plaintiffs' experts do not provide evidence

	that these tags always contain Plaintiffs' terms of use information. Dr. Wenger simply opines, without substantiation, that these are where "Plaintiffs' CMI is typically found." ECF 1858-6 (E. Wenger Decl. Ex. F) (Reply Rep.) at 70–71. [redacted] ECF 1858-1 (E. Wenger Decl. Ex. A) (Opening Rep.) at 49. Both this and her further reliance on her CMI removal analysis to support these opinions (*see* ECF 1858-6 (E. Wenger Decl. Ex. F) (Reply Rep.) at 68), are further disputed because of the unsound methodologies used, as detailed in OpenAI's Opening Daubert Motion. In fact, Dr. Wenger even admits that she sometimes found News Plaintiffs' CMI in the body of the HTML page, which was retained by OpenAI. *Id.* at 71. Additionally, Paragraph 1340 does not specify whether this assertion applies to the public version of Newspaper, OpenAI's internal fork of it, or both. Further, OpenAI disputes that Newspaper removes anything. OpenAI's text extractors do not "remove" anything. Rather, they extract material that is suitable for model training. Not extracting data is not equivalent to removing data. *See, e.g.,* ECF 1748 (C. Callison-Burch Decl.) ¶ 88 ("Importantly, the purpose of the text extraction process used by OpenAI and other NLP practitioners is to identify text to *keep*, not text to *remove*. The scripts are designed to extract readable natural language text from HTML files for use as training data. Any loss of CMI during

	this process is an incidental byproduct of the text extraction methodology—it is not a purposeful or intentional removal of CMI."). Further disputed, to the extent this implies something about OpenAI's state of mind when it chose to use Newspaper or when any CMI allegedly was bypassed during text extraction. Any alleged CMI bypassing is part of OpenAI's data processing pipeline to curate datasets that are representative of human language. *See* News Ex. 676 (C. Callison-Burch Rebuttal Rep.) ¶¶ 77-81 ("Data Cleaning is a Technical Necessity"), News Ex. 1171 (B. Frederiksen-Cross Rebuttal Rep.) ¶ 28 (describing the purpose of this processing as "maximizing the overall amount of useful, human-written text extracted from a webpage while minimizing the incidence of spam, Search Engine Optimization ('SEO'), [or] boilerplate"). Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1341. Gutentag is designed to remove author information. Wenger Opening Rep. at 7, 48-53; Wenger Rebuttal Rep. at 21-22; Wenger Reply Rep. at 66-69, 71-75; Goldstein Opening Source Code Rep. ¶¶ 25-41; Delorey Opening Rep. ¶ 245.	1341. **Disputed– Not Supported by Cited Evidence; Expert Opinion Subject to Rule 702.** OpenAI responds as follows: The facts set forth in Paragraph 1341 are disputed. Dr. Wenger, Dr. Goldstein, and Mr. Delorey's opinions focus on HTML tags which Gutentag allegedly bypasses during text extraction. *See* ECF 1858-1 (E. Wenger Decl. Ex. A.) (Opening Rep.) at 7, 48–53; ECF 1858-4 (E. Wenger Decl. Ex. D) (Rebuttal Rep.) at 21–22; ECF 1858-6 (E. Wenger Decl. Ex. F) (Reply

	Rep.) at 66–69, 71–75; A. Blanco Opp. Decl. Ex. 296 (T. Goldstein Opening Source Code Rep.) ¶¶ 25–41; ECF 1774-1 (D. Delorey Ex. A) (Opening Rep.) ¶ 245. Even if those tags are bypassed, Plaintiffs' experts do not provide evidence that these tags always contain Plaintiffs' author information. Dr. Wenger [redacted] ECF 1858-1 (E. Wenger Decl. Ex. A.) (Opening Rep.) at 50-51. Dr. Wenger's reliance on her CMI removal analysis to support these opinions (*see* ECF 1858-6 (E. Wenger Decl. Ex. F) (Reply Rep.) at 72–73) is further disputed because of the unsound methodologies used, as detailed in OpenAI's Opening Daubert Motion. In fact, even using these unsound methodologies Dr. Wenger even admits that author information is routinely retained by Gutentag. *See id.* at 75 (acknowledging that in 8 of the 9 alleged copies of a certain asserted work, Gutentag retained author information). Further disputed to the extent that Dr. Wenger has admitted she does not know what versions of Gutentag had these specific implementations and only can say it is "likely" that is true for "all versions of Gutentag." ECF 1858-1 (E. Wenger Decl. Ex. A.) (Opening Rep.) at 52. Mr. Delorey similarly fails to connect the allegedly targeted field with author information. *See* D. Delorey Opening Rep. ¶ 245 (opining that author information "can" be bypassed based on his opinion that [redacted]

	[redacted] News Ex. 1171 (B. Frederiksen-Cross Rebuttal Rep. (News)) ¶¶ 162–165; *see also* ECF 1774-1 (D. Delorey Ex. A) (Opening Rep.) ¶ 249 [redacted] As such, OpenAI disputes that Dr. Goldstein's analysis is relevant to this fact. Further, OpenAI disputes that Gutentag removes anything. OpenAI's text extractors do not "remove" anything. Rather, they extract material that is suitable for model training. Not extracting data is not equivalent to removing data. *See, e.g.,* ECF 1748 (C. Callison-Burch Decl.) ¶ 88 ("Importantly, the purpose of the text extraction process used by OpenAI and other NLP practitioners is to identify text to *keep*, not text to *remove*. The scripts are designed to extract readable natural language text from HTML files for use as training data. Any loss of CMI during this process is an incidental byproduct of the text extraction methodology—it is not a purposeful or intentional removal of CMI.").
	Further disputed, to the extent this implies something about OpenAI's state of mind when it developed Gutentag or when any CMI allegedly was bypassed during text extraction. Any alleged CMI bypassing is part of OpenAI's data processing pipeline to curate datasets that are representative of human language. *See* News Ex. 676 (C. Callison-Burch Rebuttal Rep.) ¶¶ 77–81 ("Data Cleaning is a Technical Necessity"); News Ex. 1171 (B.

	Frederiksen-Cross Rebuttal Rep.) ¶ 28 (describing the purpose of this processing as "maximizing the overall amount of useful, human-written text extracted from a webpage while minimizing the incidence of spam, Search Engine Optimization ('SEO'), [or] boilerplate"). Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1342. Gutentag is designed to remove title information. Wenger Opening Rep. at 7, 48-53; Wenger Rebuttal Rep. at 21-22; Wenger Reply Rep. at 66-69, 71-75; Goldstein Opening Source Code Rep. ¶¶ 25-41; Delorey Opening Report ¶¶ 231, 248–253.	1342. **Disputed.** OpenAI responds as follows: The facts set forth in Paragraph 1342 are disputed. Dr. Wenger, Dr. Goldstein, and Mr. Delorey's [redacted] *See* ECF 1858-1 (E. Wenger Decl. Ex. A.) (Opening Rep.) at 7, 48–53; ECF 1858-4 (E. Wenger Decl. Ex. D) (Rebuttal Rep.) at 21–22; ECF 1858-6 (E. Wenger Decl. Ex. F) (Reply Rep.) at 66–69, 71–75; A. Blanco Opp. Decl. Ex. 296 (T. Goldstein Opening Source Code Rep.) ¶¶ 25–41; ECF 1774-1 (D. Delorey Ex. A) (Opening Rep.) ¶ 245. Even if those tags are bypassed, Plaintiffs' experts do not provide evidence that these tags always contain Plaintiffs' title information. Dr. Wenger [redacted] ECF 1858-1 (E. Wenger Decl. Ex. A.) (Opening Rep.) at 50–51. Dr. Wenger's reliance on her CMI removal analysis to support these opinions (*see* ECF 1858-6 (E. Wenger Decl. Ex. F) (Reply Rep.) 72–73) is further disputed because of the unsound methodologies

	used, as detailed in OpenAI's Opening Daubert Motion. In fact, even using these unsound methodologies Dr. Wenger [redacted] *See id.* at 73 [redacted] Finally, it is further disputed to the extent that Dr. Wenger has admitted [redacted] ECF 1858-1 (E. Wenger Decl. Ex. A.) (Opening Rep.) at 52. Mr. Delorey similarly fails to connect the allegedly targeted field with title information. *See* ECF 1774-1 (D. Delorey Ex. A) (Opening Rep.) ¶ 245 [redacted] *See id.* ¶¶ 231, 248–253. [redacted] News Ex. 1171 (B. Frederiksen-Cross Rebuttal Rep. (News)) ¶¶ 162–165; *see also* ECF 1774-1 (D. Delorey Ex. A) (Opening Rep.) ¶ 249 [redacted] As such, OpenAI disputes that Dr. Goldstein's analysis is relevant to this fact. Additionally, OpenAI's text extractors do not "remove" anything. Rather, they extract material that is

	suitable for model training. Not extracting data is not equivalent to removing data. *See, e.g.,* ECF 1748 (C. Callison-Burch Decl.) ¶ 88 ("Importantly, the purpose of the text extraction process used by OpenAI and other NLP practitioners is to identify text to *keep*, not text to *remove*. The scripts are designed to extract readable natural language text from HTML files for use as training data. Any loss of CMI during this process is an incidental byproduct of the text extraction methodology—it is not a purposeful or intentional removal of CMI."). Further disputed, to the extent this implies something about OpenAI's state of mind when it developed Gutentag or when any CMI allegedly was bypassed during text extraction. Any alleged CMI bypassing is part of OpenAI's data processing pipeline to curate datasets that are representative of human language. *See* News Ex. 676 (C. Callison-Burch Rebuttal Rep.) ¶¶ 77–81 ("Data Cleaning is a Technical Necessity"), News Ex. 1171 (B. Frederiksen-Cross Rebuttal Rep.) ¶ 28 (describing the purpose of this processing as "maximizing the overall amount of useful, human-written text extracted from a webpage while minimizing the incidence of spam, Search Engine Optimization ('SEO'), [or] boilerplate"). Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1343. Gutentag is designed to remove copyright notice information. Wenger Opening Rep. at 7, 48-53;	1343. **Disputed– Not Supported by Cited Evidence; Expert Opinion Subject to Rule**

Wenger Rebuttal Rep. at 21-22; Wenger Reply Rep. at 66-69, 71-75; Goldstein Opening Source Code Rep. ¶¶ 25-41; Delorey Opening Rep. ¶¶ 247–253.	**702.** OpenAI responds as follows: This fact is disputed. Dr. Wenger, Dr. Goldstein, and Mr. Delorey's [redacted] Dr. Wenger, Dr. Goldstein, and Mr. Delorey claim this function intentionally omits the content in webpage footers, including what they claim is "copyright notices." OpenAI disputes that the function intentionally omits copyright notices, because any such omission is not intentional but rather incidental to the general purpose of the function, which is to skip non-central English boilerplate text. *See* News Ex. 676 (C. Callison-Burch Rebuttal Rep.) ¶¶ 77–81 ("Data Cleaning is a Technical Necessity"), News Ex. 1171 (B. Frederiksen-Cross Rebuttal Rep.) ¶ 28 (describing the purpose of this processing as "maximizing the overall amount of useful, human-written text extracted from a webpage while minimizing the incidence of spam, Search Engine Optimization ('SEO'), [or] boilerplate"). OpenAI further disputes that the content Dr. Wenger, Dr. Goldstein, and Mr. Delorey identifies as "copyright notices" are copyright notices within the meaning of the DMCA. CMI must be conveyed in connection with Asserted Works, and OpenAI disputes that information programmatically appended to all pages on a website qualifies as a "copyright notice." *See*, *e.g.*, News Ex. 1167 (S. Krishnan Rebuttal Rep. (Delorey)) ¶¶ 172–188

	("Mr. Delorey's 'Terms of Service' and 'Copyright Notice' Extractions Predominantly Identify Site-Wide Boilerplate"). Finally, OpenAI disputes that Gutentag removes anything. OpenAI's text extractors do not "remove" anything. Rather, they extract material that is suitable for model training. Not extracting data is not equivalent to removing data. *See, e.g.,* ECF 1748 (C. Callison-Burch Decl.) ¶ 88 ("Importantly, the purpose of the text extraction process used by OpenAI and other NLP practitioners is to identify text to *keep*, not text to *remove*. The scripts are designed to extract readable natural language text from HTML files for use as training data. Any loss of CMI during this process is an incidental byproduct of the text extraction methodology—it is not a purposeful or intentional removal of CMI."). Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1344. Gutentag is designed to remove terms of use information. Wenger Opening Rep. at 7, 48-53; Wenger Rebuttal Rep. at 21-22; Wenger Reply Rep. at 66-69, 71-75; Goldstein Opening Source Code Rep. ¶¶ 25-41; Delorey Opening Rep. ¶ 245.	1344. **Disputed– Not Supported by Cited Evidence; Expert Opinion Subject to Rule 702.** OpenAI responds as follows: The facts set forth in Paragraph 1344 are disputed. Dr. Wenger, Dr. Goldstein, and Mr. Delorey's opinions focus on HTML tags which Gutentag allegedly bypasses during text extraction. *See* ECF 1858-1 (E. Wenger Decl. Ex. A.) (Opening Rep.) at 7, 48–53; ECF 1858-4 (E. Wenger Decl. Ex. D) (Rebuttal Rep.) at 21–22; E. Wenger Reply Rep. 66–69, 71–75; A.

	Blanco Opp. Decl. Ex. 296 (T. Goldstein Opening Source Code Rep.) ¶¶ 25–41; ECF 1774-1 (D. Delorey Ex. A) (Opening Rep.) ¶ 245. Even if those tags are bypassed, Plaintiffs' experts do not provide evidence that these tags always contain Plaintiffs' terms of use information. Dr. Wenger [redacted] ECF 1858-1 (E. Wenger Decl. Ex. A.) (Opening Rep.) at 49. Both this, and her further reliance on her CMI removal analysis to support these opinions (*see* ECF 1858-6 (E. Wenger Decl. Ex. F) (Reply Rep.) at 72–73), are further disputed because of the unsound methodologies used, as detailed in OpenAI's Opening Daubert Motion. Further disputed to the extent that Dr. Wenger has admitted she does not know what versions of Gutentag had these specific implementations and only can say it is "likely" that is true for "all versions of Gutentag." ECF 1858-1 (E. Wenger Decl. Ex. A.) (Opening Rep.) at 52. Mr. Delorey similarly fails to connect the allegedly targeted fields with title information. *See* ECF 1774-1 (D. Delorey Ex. A) (Opening Rep.) ¶ 245 (opining that "terms of service" information "can" be bypassed based on his opinion that [redacted] though. News Ex. 1171 (B. Frederiksen-Cross Rebuttal Rep. (News)) ¶¶ 162–165; *see also* ECF 1774-1 (D. Delorey Ex. A)

	(Opening Rep.) ¶ 249 [redacted] As such, OpenAI disputes that Dr. Goldstein's analysis is relevant to this fact. Further, OpenAI disputes that Gutentag removes anything. OpenAI's text extractors do not "remove" anything. Rather, they extract material that is suitable for model training. Not extracting data is not equivalent to removing data. *See, e.g.*, ECF 1748 (C. Callison-Burch Decl.) ¶ 88 ("Importantly, the purpose of the text extraction process used by OpenAI and other NLP practitioners is to identify text to *keep*, not text to *remove*. The scripts are designed to extract readable natural language text from HTML files for use as training data. Any loss of CMI during this process is an incidental byproduct of the text extraction methodology—it is not a purposeful or intentional removal of CMI."). Further disputed, to the extent this implies something about OpenAI's state of mind when it developed Gutentag or when any CMI allegedly was bypassed during text extraction. Any alleged CMI bypassing is part of OpenAI's data processing pipeline to curate datasets that are representative of human language. *See* News Ex. 676 (C. Callison-Burch Rebuttal Rep.) ¶¶ 77–81 ("Data Cleaning is a Technical Necessity"), News Ex. 1171 (B. Frederiksen-Cross Rebuttal Rep.) ¶ 28 (describing the purpose of this processing as "maximizing the overall amount of useful, human-written text extracted from a

	webpage while minimizing the incidence of spam, Search Engine Optimization ('SEO'), [or] boilerplate"). Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1345. OpenAI's CMI source code expert did not opine that OpenAI's text extractors always retained News Plaintiffs' CMI. Ex. 631 [Frederiksen-Cross Expert Depo. Tr.] at 88:8-10; *id*. at 89:4-9.	1345. **Undisputed.** Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1346. OpenAI intended to remove HTML tags. Ex. 633 [Paino 30(b)(6) Depo. Tr.] at 34:7-8.	1346. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: Disputed as unsupported by the cited evidence. Specifically, OpenAI's text extractors do not "remove" anything. Rather, they extract material that is suitable for model training. Not extracting data is not equivalent to removing data. *See, e.g.*, ECF 1748 (C. Callison-Burch Decl.) ¶ 88 ("Importantly, the purpose of the text extraction process used by OpenAI and other NLP practitioners is to identify text to *keep*, not text to *remove*. The scripts are designed to extract readable natural language text from HTML files for use as training data. Any loss of CMI during this process is an incidental byproduct of the text extraction methodology—it is not a purposeful or intentional removal of CMI.") Microsoft responds that these facts do not pertain to claims asserted against Microsoft.

1347. OpenAI's decision to remove boilerplate was a "design choice when implementing the extractor." Ex. 521 [Yuan 30(b)(1) Depo. Tr.] at 84:2-8.	1347. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: Disputed as unsupported by the cited evidence. Specifically, OpenAI's text extractors do not "remove" anything. Rather, they extract material that is suitable for model training. Not extracting data is not equivalent to removing data. *See, e.g.*, ECF 1748 (C. Callison-Burch Decl.) ¶ 88 ("Importantly, the purpose of the text extraction process used by OpenAI and other NLP practitioners is to identify text to *keep*, not text to *remove*. The scripts are designed to extract readable natural language text from HTML files for use as training data. Any loss of CMI during this process is an incidental byproduct of the text extraction methodology—it is not a purposeful or intentional removal of CMI.") Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1348. OpenAI employees include at least some CMI data types within the concept of boilerplate. Ex. 1015 [OPCO_MDL_000155600] at -602 (document authored by Nick Ryder and Shantanu Jain referring to "copyright" as an example of "boiler plate").	1348. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Ex. 1015 indicates OpenAI employees included the copyright notice appended at the bottom of webpages as an example of "boilerplate." *See* Ex. 1015 at OPCO_MDL_000155602. However, OpenAI disputes the characterization of boilerplate footer texts as "CMI" conveyed with Plaintiffs' Asserted Works. Website-wide copyright notices and terms-of-service links were not conveyed in connection with any particular Asserted

	Work. Those items appeared in site footers applied indiscriminately to Plaintiffs' webpages and do not necessarily contain accurate rights-identifying information with respect to Plaintiffs' Asserted Works. *See* S. Krishnan Opp. Decl. (News) Ex. A (S. Krishnan Rebuttal Rep. (Goldstein)) ¶¶ 199–218. The disputed proposition is therefore not supported by the cited evidence. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1349. [redacted] Ex. 1171 [Frederiksen-Cross Rebuttal Rep.] ¶ 112.	1349. **Disputed in Part.** OpenAI responds as follows: The facts set forth in Paragraph 1349 are disputed in part because they are not supported by the cited evidence. The term "copyright-related terms" is vague, and no definition of that term appears in the cited support. OpenAI disputes that "copyright," "(c)," and "reserved" are necessarily "copyright-related" to the extent that Plaintiffs characterize the terms as CMI. OpenAI further disputes that the [redacted]

	[redacted] News Ex. 1171 (B. Frederiksen-Cross Rebuttal Rep. (News)) ¶¶ 112–114. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1350. [redacted] Ex. 631 [Frederiksen-Cross Expert Depo. Tr.] at 39:4-21.	1350. **Disputed.** OpenAI responds as follows: The facts set forth in Paragraph 1350 are disputed because they are not supported by the cited evidence and mischaracterize the purpose of [redacted] *See* News Ex. 1171 (B. Frederiksen-Cross Rebuttal Rep. (News)) ¶¶ 109–117. [redacted] *Id.* Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1351. The term "boilerplate," refers to "repetitive text, typically machine applied, that may occur in a web	1351. **Undisputed** that Frederiksen-Cross testified this was her definition of the term "boilerplate."

page or a document that does not contribute or extend the human-written narrative of the document." Ex. 631 [Frederiksen-Cross Expert Depo. Tr.] at 47:19-23.	Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1352. The goal of text extraction is extraction of the elements that are desired for pretraining a model which are typically long blocks of natural sounding human text. Ex. 635 [Callison-Burch July 15, 2026 Depo. Tr.] at 245:6-12.	1352. **Undisputed.** Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1353. The goal of OpenAI's text extraction was to maximize the amount of meaningful human-written text and minimize the amount of unwanted material. Ex. 483 [Callison-Burch Opening Report] ¶ 358.	1353. **Undisputed.** Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1354. An ideal text extractor is one that removes all the boilerplate and keeps all the "legit content." Ex. 521 [Yuan 30(b)(1) Depo. Tr.] at 43:10-13.	1354. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 1354 are disputed because they are not supported by the record. Yuan testified only that use of pre-fuzzy de-duplication creates the risk of identifying legitimate content as boilerplate if there are a lot of duplicates of the document. News Ex. 521 (Q. Yuan 30(b)(1) Tr.) at 43:7–12. Pursuant to Fed. R. Evid. 106, Dr. Callison-Burch explains in his Reply Report, "OpenAI's text extraction pipeline is designed to retain language data that reflects human language and ignore non-language data that would impede the models' ability to generate human language." News Ex. 675 (C. Callison-Burch Reply Rep.) ¶ 21.a. Further, Plaintiffs' characterization of the extractor as "removing" things which is unsupported by the cited

	evidence, contrary to the record, and disputed for that reason. While the extractor selectively extracts material that is suitable for model training. Not extracting data is not equivalent to removing data. Pursuant to Fed. R. Evid. 106, Dr. Callison-Burch states in his declaration, "the purpose of the text extraction process used by OpenAI and other NLP practitioners is to identify text to *keep*, not text to *remove*. The scripts are designed to extract readable natural language text from HTML files for use as training data. Any loss of CMI during this process is an incidental byproduct of the text extraction methodology—it is not a purposeful or intentional removal of CMI." ECF 1748 (C. Callison-Burch Decl.) ¶ 88. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1355. A "byline" is not "readable human language" "of the sort that is typically desired for LLM pretraining." Ex. 635 [Callison-Burch July 15, 2026 Expert Depo. Tr.] at 246:2-6.	1355. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 1355 are disputed because the cited evidence is incorrect. As stated, Paragraph 1355 implies that Dr. Callison-Burch testified regarding "readable human language." However, that is the phrasing of the questioner, not Dr. Callison-Burch. A. Blanco Opp. Decl. Ex. 181 (C. Callison-Burch Tr. (July 15, 2026)) at 246:2-6. Dr. Callison-Burch additionally testified that the byline is a short phrase. *Id.* at 246:5–8. Further disputed to the extent that, as phrased, Paragraph 1355 suggests that OpenAI intentionally bypassed the byline because it isn't

	useful for training, when in fact, any alleged CMI bypassing is incidental to data cleaning. *See* ECF 1748 (C. Callison-Burch Decl.) ¶¶ 82–89; News Ex. 676 (C. Callison-Burch Rebuttal Rep.) Section VIII ("Plaintiffs' Experts Misunderstand CMI Removal in Training Data"). Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1356. A byline is not desired for LLM pretraining because it is "a short, you know, three-word phrase" rather than "large natural language text." Ex. 635 [Callison-Burch July 15, 2026 Expert Depo. Tr.] at 243:6-16, 246:2-16.	1356. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 1356 are disputed because the cited evidence is incorrect. Plaintiffs omit the portion of Dr. Callison-Burch's testimony reiterating that "his understanding is limited" as to what constitutes CMI and that this byline discussion occurred in the context of whether a byline is considered CMI. A. Blanco Opp. Decl. Ex. 181 (C. Callison-Burch Tr. (Jul. 15, 2026)) at 242:3–9, 243:2–3; *see also* Fed. R. Evid. 106. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1357. The purpose of OpenAI's text extraction was to remove text that "[y]ou wouldn't want the model to be outputting." Ex. 529 [Chess 30(b)(1) Depo. Tr.] at 154:19-155:2.	1357. **Disputed.** OpenAI responds as follows: The facts set forth in Paragraph 1357 are not supported by the cited evidence and contradicted by the record. Specifically, pursuant to Fed. R. Evid. 106, Dr. Callison-Burch states in his Reply Report, "OpenAI's text extraction pipeline is designed to retain language data that reflects human language and ignore non-language data that would impede

	the models' ability to generate human language." News Ex. 675 (C. Callison-Burch Reply Rep.) ¶ 21.a. Further disputed as to the suggestion that the extractor "removes text." While the extractor selectively extracts material that is suitable for model training. Not extracting data is not equivalent to removing data Pursuant to Fed. R. Evid. 106, Dr. Callison-Burch states in his declaration, "the purpose of the text extraction process used by OpenAI and other NLP practitioners is to identify text to *keep*, not text to *remove*. The scripts are designed to extract readable natural language text from HTML files for use as training data. Any loss of CMI during this process is an incidental byproduct of the text extraction methodology—it is not a purposeful or intentional removal of CMI." ECF 1748 (C. Callison-Burch Decl). ¶ 88. Further disputed as omitting context necessary to understanding the fact as stated. Chess testified [redacted] A. Blanco Opp. Decl. Ex. 184 (B. Chess 30(b)(1) Tr.) at 154:19–155:10. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1358. OpenAI did not want its models to output "copyright notices." Ex. 529 [Chess. 30(b)(1) Depo. Tr.] at 156:11-16.	1358. **Disputed.** OpenAI responds as follows: This fact is disputed as stated as it omits necessary context and is not supported by the cited evidence. Specifically, OpenAI wants the models to train on actual human language, which copyright notices are not. Pursuant to

	Fed. R. Evid. 106, Plaintiffs have cited to no support for the proposition that the inclusion of copyright notices in training data would help with misattribution, which remains entirely hypothetical. *See, e.g.*, A. Blanco Opp. Decl. Ex. 197 (O. Etzioni Tr.) 170:2–17. Pursuant to Fed. R. Evid. 106, Dr. Callison-Burch states in his declaration, "the purpose of the text extraction process used by OpenAI and other NLP practitioners is to identify text to *keep*, not text to *remove*. The scripts are designed to extract readable natural language text from HTML files for use as training data. Any loss of CMI during this process is an incidental byproduct of the text extraction methodology—it is not a purposeful or intentional removal of CMI." ECF 1748 (C. Callison-Burch Decl.) ¶ 88. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1359. A link to a website's Terms of Service "would not be something that [OpenAI] would want inputted into the model." Ex. 529 [Chess. 30(b)(1) Depo. Tr.] at 155:21-156:5.	1359. **Disputed in Part.** OpenAI responds as follows: This fact is disputed as stated as it omits necessary context and is not supported by the cited evidence. Specifically, Chess testified [redacted] News Ex. 529 (B. Chess 30(b)(1) Tr.) at 156:2–10. Further disputed to the extent Plaintiffs suggest that

	OpenAI intentionally bypassed the Terms of Service because it isn't useful for training, when in fact, any alleged CMI bypassing is incidental to data cleaning. Pursuant to Fed. R. Evid. 106, Dr. Callison-Burch states in his declaration, "the purpose of the text extraction process used by OpenAI and other NLP practitioners is to identify text to *keep*, not text to *remove*. The scripts are designed to extract readable natural language text from HTML files for use as training data. Any loss of CMI during this process is an incidental byproduct of the text extraction methodology—it is not a purposeful or intentional removal of CMI." ECF 1748 (C. Callison-Burch Decl.) ¶ 88. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1360. When OpenAI processed the raw HTML content from the datasets OpenAI acquired from Common Crawl that contained News Plaintiffs' Asserted Works, OpenAI removed CMI, including the title and other information identifying the work, the name of and other identifying information about the author of the work, the name of and other identifying information about the copyright owner of the work, and terms and conditions for the use of a work (or a link to the same). Goldstein Opening Report and Errata thereto, Tables 12-20; Delorey Opening Report at ¶ 248–263; Wenger Opening Rep. at 7.	1360. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 1360 are disputed because the cited evidence is incorrect. These facts are disputed because Plaintiffs' experts did not reliably establish that OpenAI bypassed CMI when extracting text from copies of News Plaintiffs' Asserted Works. Their analyses assume, without reliably demonstrating, that the identified training records were copies of the corresponding Asserted Works, that the particular source records OpenAI acquired contained the asserted CMI, and that the comparison WARC files were the same files OpenAI processed. Their methodologies also

	frequently misidentified sitewide or unrelated webpage material as CMI and searched only limited portions of OpenAI's training-data records, thereby overlooking title, author, publisher, and other source-identifying information retained in URLs and metadata. Nor did Dr. Wenger establish that any particular extractor was used to process the training records on which she based her opinions. Fed. R. Evid. 106; S. Krishnan Opp. Decl. (News) Ex. H (S. Krishnan Rebuttal Rep. (Wenger)) ¶¶ 136, 142–205; S. Krishnan Opp. Decl. (News) Ex. A (S. Krishnan Rebuttal Rep. (Goldstein)) ¶¶ 153–232; News Ex. 1167 (S. Krishnan Rebuttal Rep. (Delorey)) ¶¶ 140–217. OpenAI's text extractors do not "remove" anything. Rather, they extract material that is suitable for model training. Pursuant to Fed. R. Evid. 106, not extracting data is not equivalent to removing data. *See, e.g.*, ECF 1748 (C. Callison-Burch Decl.) ¶ 88 ("Importantly, the purpose of the text extraction process used by OpenAI and other NLP practitioners is to identify text to *keep*, not text to *remove*. The scripts are designed to extract readable natural language text from HTML files for use as training data. Any loss of CMI during this process is an incidental byproduct of the text extraction methodology—it is not a purposeful or intentional removal of CMI."). Microsoft responds that these facts do not pertain to claims asserted against Microsoft.

2. OpenAI's Removal of The Times's and DNP's CMI

1361. The Times and Daily News Plaintiffs' Rule 1006 summaries for CMI removal set forth, *inter alia*, CMI removal information for each OpenAI training data record for which The Times and the Daily News Plaintiffs are moving for summary judgment. Maisel Declaration, Exhibits C and D (CMI Removal Rule 1006 summaries).	1361. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 1361 are disputed because the cited evidence is incorrect. Specifically, OpenAI disputes that those spreadsheets set forth actual "CMI removal information." The spreadsheets are outputs of Dr. Goldstein's disputed expert analysis and embed his conclusions that a training-data record matches an Asserted Work, that a particular Common Crawl WARC file corresponds to that record, that the WARC file contained work-specific CMI, and that OpenAI removed that CMI. Pursuant to Fed. R. Evid. 106, Dr. Krishnan identified fundamental errors at each of those steps, including false-positive work matches, an "oldest match" heuristic estimated to select the wrong WARC file 82% of the time, erroneous identification of purported CMI, and failure to identify CMI retained outside the training record's text field. S. Krishnan Opp. Decl. (News) Ex. A (S. Krishnan Rebuttal Rep. (Goldstein)) ¶¶ 68–131, 153–232. Even without accounting for all of these errors, Dr. Krishnan demonstrated that more than 83% of Dr. Goldstein's alleged CMI removals were erroneous. S. Krishnan Opp. Decl. (News) Ex. A (S. Krishnan Rebuttal Rep. (Goldstein)) ¶¶ 68–131, 153–232; Fed. R. Evid. 106. OpenAI further objects to characterizing these expert-generated analytical outputs as Rule 1006 summaries because they do not merely summarize voluminous underlying records; they incorporate disputed methodological judgments, expert opinions, and the legal

	conclusion that CMI was "removed," while also depending on the admissibility and accuracy of the underlying materials. Undisputed that Plaintiffs created spreadsheets labeled "Rule 1006 summaries" and populated them with rows for the training-data records on which Plaintiffs purport to seek summary judgment. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1362. More specifically, the Rule 1006 summaries for CMI removal contain the following information for each CMI removal on which The Times and Daily News Plaintiffs are moving for summary judgment:	1362. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Plaintiffs' spreadsheets contain columns bearing the listed titles or that certain columns reproduce identifiers, links, or numerical outputs. However, the facts set forth in Paragraph 1362 are disputed because the cited evidence is incorrect. Specifically, OpenAI disputes the descriptions of those columns to the extent they present Dr. Goldstein's conclusions and extrapolations as established facts. *First*, Columns F through H and N do not reliably identify a "matching Asserted Work." The shared-16-gram and coverage-score criteria are unvalidated similarity heuristics that produce false positives, including matches based on quotations, syndicated material, public-domain material, boilerplate, and other third-party content. Restricting Plaintiffs' motion to records with an 80% coverage score does not establish that the paired texts are copies of the same work, much less copies of expression

Column	Column Title	Explanation
A	CMI Removal Rule 1006 Summary Index	Index of entry in Rule 1006 Summary
B	OpenAI Record - Bates Number	Identifies the Bates number of the produced training data record
C	OpenAI Record - Record Number	Identifies the Record Number of the produced training data record; this is an offset, which identifies the particular record in the file corresponding to the Bates Number.
D	OAI Training Dataset Name	Identifies the OpenAI training dataset corresponding to the training data record.
E	OAI Model Name	Identifies the OpenAI model name corresponding to the training data record.
F	Registration Number	Identifies the Registration Number of the matching Asserted Work, where the matching Asserted Work has at least a shared 16-gram and 80% or higher coverage score.
G	Asserted Work URL	Identifies the URL of the matching Asserted Work.
H	Asserted Work Record Text	Links to the text of the Asserted Work.
I	Training Data Record Text	Links to the text of the Training Data Record.
J	WARC File - Crawl Identifier	Identifies the Common Crawl WARC File (by Crawl Identifier) corresponding to the training data record.
K	WARC File - Record Id	Identifies the Common Crawl WARC File (by Record Id) corresponding to the training data record.
L	WARC File Content	Links to the Common Crawl WARC File corresponding to the training data record.
M	Training Data Source URL	Identifies the source URL from which the training data record is derived.
N	Coverage Score Between	Identifies the coverage score between the Asserted Work and the Training Data Record.

owned by Plaintiffs. S. Krishnan Opp. Decl. (News) Ex. A (S. Krishnan Rebuttal Rep. (Goldstein)) ¶¶ 68–131; Fed. R. Evid. 106.

Second, Columns J through L do not reliably identify the Common Crawl WARC file "corresponding" to a training-data record. Dr. Goldstein did not possess the WARC files OpenAI actually processed. He instead selected the oldest available WARC file with a matching URL, even though the same URL was ordinarily captured multiple times and those captures frequently contained different content and different purported CMI. *Id.* ¶¶ 166–171. Dr. Krishnan found that 98% of sampled records could be associated with multiple WARC files, 95.2% had an alternative WARC yielding different purported CMI, and Dr. Goldstein's estimated likelihood of selecting the wrong WARC file was 82%. Id. ¶ 170. The coverage score reported in Column O does not cure that defect because it compares article body text, not the CMI that varied among captures. *Id.* ¶¶ 163–172; Fed. R. Evid. 106.

Third, Columns P through AB do not establish whether CMI was "removed" or identify CMI that was actually removed. Dr. Goldstein's extraction process frequently identified incorrect titles, authors, copyright notices, publishers, and sitewide footer language rather than CMI conveyed with a particular Asserted Work. He then searched only the training record's text field, overlooking source-identifying information retained in URLs and other metadata. Dr. Krishnan's analysis demonstrated, among other things, that OpenAI retained

	Asserted Work and Training Data Record	
O	Coverage Score Between WARC Article Text and Training Data Record	Identifies the coverage score between the Common Crawl WARC File text and the Training Data Record.
P	Any CMI Removed?	Identifies whether any type of CMI has been removed.
Q, S, U, W, Y, AA	[CMI Type] Removed? For [CMI Type] = Publisher Name, Title, Author Name, Publish Date, Terms of Service, Copyright Notice	For each [CMI Type], identifies whether the [CMI Type] has been removed.
R,T,V,X,Z,AB	[CMI Type] in WARC File	For each [CMI Type], if the [CMI Type] has been removed, identifies the text corresponding to that [CMI Type] that has been removed.

Goldstein Declaration, ¶8.

every purported publisher and more than 80% of the purported titles that Dr. Goldstein classified as removed, and that more than 83% of Dr. Goldstein's total alleged CMI removals were erroneous. *Id.* ¶¶ 174–232; Fed. R. Evid. 106.

Microsoft responds that these facts do not pertain to claims asserted against Microsoft.

1363. There are 3,002,796 OpenAI training data records which match to a Times Asserted Work with a coverage score of at least 80%, which were created by processing a copy of a Times webpage containing CMI conveyed in connection with the Asserted Work, and for which at least one of (1) publisher, (2) title, (3) author, (4) publication date, (5) Terms of Service, and (6) copyright notice information was absent from the training data record copy of the Asserted Work. Goldstein Declaration, Exhibits K and L; Maisel Declaration, Exhibits C and D (CMI Removal Rule 1006 summaries).

1363. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Dr. Goldstein's table reports 3,002,796 records satisfying Dr. Goldstein's proposed filters. However, the facts set forth in Paragraph 1363 are disputed because the cited evidence is incorrect. Specifically, OpenAI disputes that those records have been shown to be copies of Times Asserted Works created from webpages containing work-specific CMI that was subsequently absent from training data records. An 80% coverage score does not necessarily establish that a training-data record is a copy of the matching Asserted Work, particularly where the overlap may reflect syndicated, public-domain, quoted, or other third-party material. Nor did Dr. Goldstein

	establish that any record was created from the WARC file he selected: he used an "oldest match" heuristic that, according to Dr. Krishnan's analysis, was approximately 82% likely to select the wrong WARC file. S. Krishnan Opp. Decl. (News) Ex. A (S. Krishnan Rebuttal Rep. (Goldstein)) ¶¶ 166–171. Finally, Dr. Goldstein searched only the record's text field for exact strings, thereby classifying information as "absent" even when it was retained in the URL or other fields. *Id.* ¶ 147. Pursuant to Fed. R. Evid. 106, Dr. Krishnan's broader analysis of the same methodology verified that at least 17,042,418 of Dr. Goldstein's 20,634,870 individual alleged CMI removals for The Times—82.6%—were erroneous, without accounting for all identified matching and WARC-selection errors. *Id.* ¶¶ 68–131, 153–232 & Table 15. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1364. Table 12 provided with the Goldstein Declaration Exhibit K accurately states the number of OpenAI training data records which match to a Times Asserted Work with a coverage score of at least 80%, which were created by processing a copy of a Times webpage containing, respectively, (1) publisher, (2) title, (3) author, (4) publication date, (5) Terms of Service, and (6) copyright notice information conveyed in connection with the Asserted Work, and for which the corresponding (1) publisher, (2) title, (3) author, (4) publication date, (5)	1364. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Table 12 accurately reports the outputs of Dr. Goldstein's methodology. However, the facts set forth in Paragraph 1364 are disputed because the cited evidence is incorrect. Specifically, OpenAI disputes that Table 12 accurately establishes the number of records satisfying the substantive descriptions in the statement. Dr. Goldstein did not reliably establish that each listed record matched a

Terms of Service, and (6) copyright notice information was absent from the training data record copy of the Asserted Work. Goldstein Declaration, Exhibits K and L; Maisel Declaration, Exhibits C and D (CMI Removal Rule 1006 summaries).	Times Asserted Work, was created by processing the WARC file he selected, originally contained work-specific CMI, or omitted that information from the complete training-data record. S. Krishnan Opp. Decl. (News) Ex. A (S. Krishnan Rebuttal Rep. (Goldstein)) ¶¶ 153–232 & Table 15; Fed. R. Evid. 106. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1365. There are 467,758 OpenAI training data records which were sourced from The Mercury News website and which match to a Mercury News Asserted Work with a coverage score of at least 80%, which were created by processing a copy of a Mercury News webpage containing CMI conveyed in connection with the Asserted Work, and for which at least one of (1) publisher, (2) title, (3) author, (4) publication date, (5) terms of service, and (6) copyright notice information was absent from the training data record copy of the Asserted Work. Goldstein Declaration (providing updated Tables 12-20 showing CMI removal instances).	1365. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Dr. Goldstein's table reports 467,758 records satisfying Dr. Goldstein's proposed filters. However, the facts set forth in Paragraph 1365 are disputed because the cited evidence is incorrect. Specifically, OpenAI disputes that those records have been shown to be copies of Mercury News Asserted Works created from webpages containing work-specific CMI that was subsequently absent from training data records for the reasons described in OpenAI's response to Fact 1363. Pursuant to Fed. R. Evid. 106, Dr. Krishnan's broader analysis verified that at least 3,048,618 of Dr. Goldstein's 3,199,101 individual alleged Mercury News CMI removals—95.3%—were erroneous, without accounting for every matching and WARC-selection error. S. Krishnan Opp. Decl. (News) Ex. A (S. Krishnan Rebuttal Rep. (Goldstein)) ¶¶ 68–131, 153–232, Table 6 & Table 15.

	Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1366. Table 13 provided with the Goldstein Declaration Exhibit K accurately states the number of OpenAI training data records which match to a Mercury News Asserted Work with a coverage score of at least 80%, which were created by processing a copy of a Mercury News webpage containing, respectively, (1) publisher, (2) title, (3) author, (4) publication date, (5) terms of service, and (6) copyright notice information conveyed in connection with the Asserted Work, and for which the corresponding (1) publisher, (2) title, (3) author, (4) publication date, (5) terms of service, and (6) copyright notice information was absent from the training data record copy of the Asserted Work. Goldstein Declaration, Exhibits K and L; Maisel Declaration, Exhibits C and D (CMI Removal Rule 1006 summaries).	1366. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: OpenAI does not dispute Table 13 accurately reports the outputs of Dr. Goldstein's methodology. However, the facts set forth in Paragraph 1366 are disputed because the cited evidence is incorrect. Specifically, OpenAI disputes that Table 13 accurately establishes the number of records satisfying the substantive descriptions in the statement. Dr. Goldstein did not reliably establish that each listed record matched a Mercury News Asserted Work, was created by processing the WARC file he selected, originally contained work-specific CMI, or omitted that information from the complete training-data record. S. Krishnan Opp. Decl. (News) Ex. A (S. Krishnan Rebuttal Rep. (Goldstein)) ¶¶ 153–232, Table 6 & Table 15; Fed. R. Evid. 106. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1367. There are 575,250 OpenAI training data records which match to a Chicago Tribune Asserted Work with a coverage score of at least 80%, which were created by processing a copy of a Chicago Tribune webpage containing CMI conveyed in connection with the Asserted Work, and for which at least one of (1) publisher, (2) title, (3) author,	1367. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Dr. Goldstein's table reports 575,250 records satisfying Dr. Goldstein's proposed filters. However, the facts set forth in Paragraph 1367 are disputed because the cited evidence is

(4) publication date, (5) terms of service, and (6) copyright notice information was absent from the training data record copy of the Asserted Work. Goldstein Declaration, Exhibits K and L; Maisel Declaration, Exhibits C and D (CMI Removal Rule 1006 summaries).	incorrect. Specifically, OpenAI disputes that those records have been shown to be copies of Chicago Tribune Asserted Works created from webpages containing work-specific CMI that was subsequently absent from training data records for the reasons described in OpenAI's response to Fact 1363. Pursuant to Fed. R. Evid. 106, Dr. Krishnan's broader analysis verified that at least 2,799,267 of Dr. Goldstein's 3,407,951 individual alleged Chicago Tribune CMI removals—82.1%—were erroneous, without accounting for all matching and WARC-selection errors. S. Krishnan Opp. Decl. (News) Ex. A (S. Krishnan Rebuttal Rep. (Goldstein)) ¶¶ 68–131, 153–232, Table 6 & Table 15. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1368. Table 14 provided with the Goldstein Declaration Exhibit K accurately states the number of OpenAI training data records which match to a Chicago Tribune Asserted Work with a coverage score of at least 80%, which were created by processing a copy of a Chicago Tribune webpage containing, respectively, (1) publisher, (2) title, (3) author, (4) publication date, (5) terms of service, and (6) copyright notice information conveyed in connection with the Asserted Work, and for which the corresponding (1) publisher, (2) title, (3) author, (4) publication date, (5) terms of service, and (6) copyright notice information was absent from the training data record copy of the Asserted Work. Goldstein Declaration, Exhibits	1368. **Disputed in Part- Not Supported by Cited Evidence.** OpenAI responds as follows: OpenAI does not dispute Table 14 accurately reports the outputs of Dr. Goldstein's methodology. However, the facts set forth in Paragraph 1368 are disputed because the cited evidence is incorrect. Specifically, OpenAI disputes that Table 14 accurately establishes the number of records satisfying the substantive descriptions in the statement. Dr. Goldstein did not reliably establish that each listed record matched a Chicago Tribune Asserted Work, was created by processing the WARC file he selected, originally contained work-specific CMI, or omitted that information from the

K and L; Maisel Declaration, Exhibits C and D (CMI Removal Rule 1006 summaries).	complete training-data record. S. Krishnan Opp. Decl. (News) Ex. A (S. Krishnan Rebuttal Rep. (Goldstein)) ¶¶ 153–232, Table 6 & Table 15; Fed. R. Evid. 106. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1369. There are 315,076 OpenAI training data records which were sourced from the Denver Post website and which match to a Denver Post Asserted Work with a coverage score of at least 80%, which were created by processing a copy of a Denver Post webpage containing CMI conveyed in connection with the Asserted Work, and for which at least one of (1) publisher, (2) title, (3) author, (4) publication date, (5) terms of service, and (6) copyright notice information was absent from the training data record copy of the Asserted Work. Goldstein Declaration, Exhibits K and L; Maisel Declaration, Exhibits C and D (CMI Removal Rule 1006 summaries).	1369. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Dr. Goldstein's table reports 315,076 records satisfying Dr. Goldstein's proposed filters. However, the facts set forth in Paragraph 1369 are disputed because the cited evidence is incorrect. Further, these facts are disputed because OpenAI disputes that those records have been shown to be copies of Denver Post Asserted Works created from webpages containing work-specific CMI that was subsequently absent from training data records for the reasons described in OpenAI's response to Fact 1363. Pursuant to Fed. R. Evid. 106, Dr. Krishnan's broader analysis verified that at least 2,527,690 of Dr. Goldstein's 2,767,491 individual alleged Denver Post CMI removals—91.3%—were erroneous, without accounting for all matching and WARC-selection errors. S. Krishnan Opp. Decl. (News) Ex. A (S. Krishnan Rebuttal Rep. (Goldstein)) ¶¶ 68–131, 153–232, Table 6 & Table 15. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.

1370. Table 15 provided with the Goldstein Declaration Exhibit K accurately states the number of OpenAI training data records which match to a Denver Post Asserted Work with a coverage score of at least 80%, which were created by processing a copy of a Denver Post webpage containing, respectively, (1) publisher, (2) title, (3) author, (4) publication date, (5) terms of service, and (6) copyright notice information conveyed in connection with the Asserted Work, and for which the corresponding (1) publisher, (2) title, (3) author, (4) publication date, (5) terms of service, and (6) copyright notice information was absent from the training data record copy of the Asserted Work. Goldstein Declaration, Exhibits K and L; Maisel Declaration, Exhibits C and D (CMI Removal Rule 1006 summaries).	1370. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: OpenAI does not dispute Table 15 accurately reports the outputs of Dr. Goldstein's methodology. However, the facts set forth in Paragraph 1370 are disputed because the cited evidence is incorrect. Specifically, OpenAI disputes that Table 15 accurately establishes the number of records satisfying the substantive descriptions in the statement. Pursuant to Fed. R. Evid. 106, Dr. Goldstein did not reliably establish that each listed record matched a Denver Post Asserted Work, was created by processing the WARC file he selected, originally contained work-specific CMI, or omitted that information from the complete training-data record. S. Krishnan Opp. Decl. (News) Ex. A (S. Krishnan Rebuttal Rep. (Goldstein)) ¶¶ 153–232, Table 6 & Table 15. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1371. There are 363,246 OpenAI training data records which were sourced from the Orlando Sentinel website and which match to an Orlando Sentinel Asserted Work with a coverage score of at least 80%, which were created by processing a copy of an Orlando Sentinel webpage containing CMI conveyed in connection with the Asserted Work, and for which at least one of (1) publisher, (2) title, (3) author, (4) publication date, (5) terms of service, and (6) copyright notice information was absent	1371. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Dr. Goldstein's table reports 363,246 records satisfying Dr. Goldstein's proposed filters. However, the facts set forth in Paragraph 1371 are disputed because the cited evidence is incorrect. Specifically, OpenAI disputes that those records have been shown to be copies of Orlando Sentinel Asserted Works created from webpages containing work-specific

from the training data record copy of the Asserted Work. Goldstein Declaration, Exhibits K and L; Maisel Declaration, Exhibits C and D (CMI Removal Rule 1006 summaries).	CMI that was subsequently absent from training data records for the reasons described in OpenAI's response to Fact 1363. Pursuant to Fed. R. Evid. 106, Dr. Krishnan's broader analysis verified that at least 1,608,896 of Dr. Goldstein's 2,269,107 individual alleged Orlando Sentinel CMI removals—70.9%—were erroneous, without accounting for all matching and WARC-selection errors. S. Krishnan Opp. Decl. (News) Ex. A (S. Krishnan Rebuttal Rep. (Goldstein)) ¶¶ 68–131, 153–232, Table 6 & Table 15. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1372. Table 16 provided with the Goldstein Declaration Exhibit K accurately states the number of OpenAI training data records which match to an Orlando Sentinel Asserted Work with a coverage score of at least 80%, which were created by processing a copy of an Orlando Sentinel webpage containing, respectively, (1) publisher, (2) title, (3) author, (4) publication date, (5) terms of service, and (6) copyright notice information conveyed in connection with the Asserted Work, and for which the corresponding (1) publisher, (2) title, (3) author, (4) publication date, (5) terms of service, and (6) copyright notice information was absent from the training data record copy of the Asserted Work. Goldstein Declaration, Exhibits K and L; Maisel Declaration, Exhibits C and D (CMI Removal Rule 1006 summaries).	1372. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: OpenAI does not dispute Table 16 accurately reports the outputs of Dr. Goldstein's methodology. However, the facts set forth in Paragraph 1372 are disputed because the cited evidence is incorrect. Specifically, OpenAI disputes that Table 16 accurately establishes the number of records satisfying the substantive descriptions in the statement. Pursuant to Fed. R. Evid. 106, Dr. Goldstein did not reliably establish that each listed record matched an Orlando Sentinel Asserted Work, was created by processing the WARC file he selected, originally contained work-specific CMI, or omitted that information from the complete training-data record. S. Krishnan Opp. Decl. (News) Ex. A (S. Krishnan Rebuttal Rep. (Goldstein)) ¶¶ 153–232, Table 6 & Table 15.

	Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1373. There are 344,198 OpenAI training data records which were sourced from the New York Daily News website and which match to a New York Daily News Asserted Work with a coverage score of at least 80%, which were created by processing a copy of a New York Daily News webpage containing CMI conveyed in connection with the Asserted Work, and for which at least one of (1) publisher, (2) title, (3) author, (4) publication date, (5) terms of service, and (6) copyright notice information was absent from the training data record copy of the Asserted Work. Goldstein Declaration, Exhibits K and L; Maisel Declaration, Exhibits C and D (CMI Removal Rule 1006 summaries).	1373. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Dr. Goldstein's table reports 344,198 records satisfying Dr. Goldstein's proposed filters. However, the facts set forth in Paragraph 1373 are disputed because the cited evidence is incorrect. Specifically, OpenAI disputes that those records have been shown to be copies of New York Daily News Asserted Works created from webpages containing work-specific CMI that was subsequently absent from training data records for the reasons described in OpenAI's response to Fact 1363. Pursuant to Fed. R. Evid. 106, Dr. Krishnan's broader analysis verified that at least 1,308,717 of Dr. Goldstein's 1,682,127 individual alleged New York Daily News CMI removals—77.8%—were erroneous, without accounting for all matching and WARC-selection errors. S. Krishnan Opp. Decl. (News) Ex. A (S. Krishnan Rebuttal Rep. (Goldstein)) ¶¶ 68–131, 153–232, Table 6 & Table 15. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1374. Table 17 provided with the Goldstein Declaration Exhibit K accurately states the number of OpenAI training data records which match to a New York	1374. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: OpenAI

Daily News Asserted Work with a coverage score of at least 80%, which were created by processing a copy of a New York Daily News webpage containing, respectively, (1) publisher, (2) title, (3) author, (4) publication date, (5) terms of service, and (6) copyright notice information conveyed in connection with the Asserted Work, and for which the corresponding (1) publisher, (2) title, (3) author, (4) publication date, (5) terms of service, and (6) copyright notice information was absent from the training data record copy of the Asserted Work. Goldstein Declaration, Exhibits K and L; Maisel Declaration, Exhibits C and D (CMI Removal Rule 1006 summaries).	does not dispute Table 17 accurately reports the outputs of Dr. Goldstein's methodology. However, the facts set forth in Paragraph 1374 are disputed because the cited evidence is incorrect. Specifically, OpenAI disputes that Table 17 accurately establishes the number of records satisfying the substantive descriptions in the statement. Pursuant to Fed. R. Evid. 106, Dr. Goldstein did not reliably establish that each listed record matched a New York Daily News Asserted Work, was created by processing the WARC file he selected, originally contained work-specific CMI, or omitted that information from the complete training-data record. S. Krishnan Opp. Decl. (News) Ex. A (S. Krishnan Rebuttal Rep. (Goldstein)) ¶¶ 153–232, Table 6 & Table 15. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1375. There are 190,683 OpenAI training data records which were sourced from the Sun Sentinel website and which match to a Sun Sentinel Asserted Work with a coverage score of at least 80%, which were created by processing a copy of a Sun Sentinel webpage containing CMI conveyed in connection with the Asserted Work, and for which at least one of (1) publisher, (2) title, (3) author, (4) publication date, (5) terms of service, and (6) copyright notice information was absent from the training data record copy of the Asserted Work. Goldstein Declaration, Exhibits	1375. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Dr. Goldstein's table reports 190,683 records satisfying Dr. Goldstein's proposed filters. However, the facts set forth in Paragraph 1375 are disputed because the cited evidence is incorrect. Specifically, OpenAI disputes that those records have been shown to be copies of Sun Sentinel Asserted Works created from webpages containing work-specific CMI that was subsequently absent from training data records for the reasons described in OpenAI's response to

K and L; Maisel Declaration, Exhibits C and D (CMI Removal Rule 1006 summaries).	Fact 1363. Pursuant to Fed. R. Evid. 106, Dr. Krishnan's broader analysis verified that at least 941,641 of Dr. Goldstein's 1,203,662 individual alleged Sun Sentinel CMI removals—78.2%—were erroneous, without accounting for all matching and WARC-selection errors. S. Krishnan Opp. Decl. (News) Ex. A (S. Krishnan Rebuttal Rep. (Goldstein)) ¶¶ 68–131, 153–232, Table 6 & Table 15. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1376. Table 18 provided with the Goldstein Declaration Exhibit K accurately states the number of OpenAI training data records which match to a Sun Sentinel Asserted Work with a coverage score of at least 80%, which were created by processing a copy of a Sun Sentinel webpage containing, respectively, (1) publisher, (2) title, (3) author, (4) publication date, (5) terms of service, and (6) copyright notice information conveyed in connection with the Asserted Work, and for which the corresponding (1) publisher, (2) title, (3) author, (4) publication date, (5) terms of service, and (6) copyright notice information was absent from the training data record copy of the Asserted Work. Goldstein Declaration, Exhibits K and L; Maisel Declaration, Exhibits C and D (CMI Removal Rule 1006 summaries).	1376. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: OpenAI does not dispute Table 18 accurately reports the outputs of Dr. Goldstein's methodology. However, the facts set forth in Paragraph 1376 are disputed because the cited evidence is incorrect. Specifically, OpenAI disputes that Table 18 accurately establishes the number of records satisfying the substantive descriptions in the statement. Pursuant to Fed. R. Evid. 106, Dr. Goldstein did not reliably establish that each listed record matched a Sun Sentinel Asserted Work, was created by processing the WARC file he selected, originally contained work-specific CMI, or omitted that information from the complete training-data record. S. Krishnan Opp. Decl. (News) Ex. A (S. Krishnan Rebuttal Rep. (Goldstein)) ¶¶ 153–232, Table 6 & Table 15.

	Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1377. There are 188,148 OpenAI training data records which were sourced from the Pioneer Press website and which match to a Pioneer Press Asserted Work with a coverage score of at least 80%, which were created by processing a copy of a Pioneer Press webpage containing CMI conveyed in connection with the Asserted Work, and for which at least one of (1) publisher, (2) title, (3) author, (4) publication date, (5) terms of service, and (6) copyright notice information was absent from the training data record copy of the Asserted Work. Goldstein Declaration, Exhibits K and L; Maisel Declaration, Exhibits C and D (CMI Removal Rule 1006 summaries).	1377. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Dr. Goldstein's table reports 188,148 records satisfying Dr. Goldstein's proposed filters. However, the facts set forth in Paragraph 1377 are disputed because the cited evidence is incorrect. Specifically, OpenAI disputes that those records have been shown to be copies of Pioneer Press Asserted Works created from webpages containing work-specific CMI that was subsequently absent from training data records for the reasons described in OpenAI's response to Fact 1363. Pursuant to Fed. R. Evid. 106, Dr. Krishnan's broader analysis verified that at least 1,010,918 of Dr. Goldstein's 1,052,677 individual alleged Pioneer Press CMI removals—96.0%—were erroneous, without accounting for all matching and WARC-selection errors. S. Krishnan Opp. Decl. (News) Ex. A (S. Krishnan Rebuttal Rep. (Goldstein)) ¶¶ 68–131, 153–232, Table 6 & Table 15. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1378. Table 19 provided with the Goldstein Declaration Exhibit K accurately states the number of OpenAI training data records which match to a Pioneer Press Asserted Work with a coverage score of at least 80%, which were created by processing a copy of a Pioneer Press	1378. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: OpenAI does not dispute Table 19 accurately reports the outputs of Dr. Goldstein's methodology.

webpage containing, respectively, (1) publisher, (2) title, (3) author, (4) publication date, (5) terms of service, and (6) copyright notice information conveyed in connection with the Asserted Work, and for which the corresponding (1) publisher, (2) title, (3) author, (4) publication date, (5) terms of service, and (6) copyright notice information was absent from the training data record copy of the Asserted Work. Goldstein Declaration, Exhibits K and L; Maisel Declaration, Exhibits C and D (CMI Removal Rule 1006 summaries).	However, the facts set forth in Paragraph 1378 are disputed because the cited evidence is incorrect. Specifically, OpenAI disputes that Table 19 accurately establishes the number of records satisfying the substantive descriptions in the statement. Pursuant to Fed. R. Evid. 106, Dr. Goldstein did not reliably establish that each listed record matched a Pioneer Press Asserted Work, was created by processing the WARC file he selected, originally contained work-specific CMI, or omitted that information from the complete training-data record. S. Krishnan Opp. Decl. (News) Ex. A (S. Krishnan Rebuttal Rep. (Goldstein)) ¶¶ 153–232, Table 6 & Table 15. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1379. There are 194,138 OpenAI training data records which were sourced from the Orange County Register website and which match to an Orange County Register Asserted Work with a coverage score of at least 80%, which were created by processing a copy of an Orange County Register webpage containing CMI conveyed in connection with the Asserted Work, and for which at least one of (1) publisher, (2) title, (3) author, (4) publication date, (5) terms of service, and (6) copyright notice information was absent from the training data record copy of the Asserted Work. Goldstein Declaration, Exhibits K and L; Maisel Declaration, Exhibits C and D (CMI Removal Rule 1006 summaries).	1379. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: Undisputed that Dr. Goldstein's table reports 194,138 records satisfying Dr. Goldstein's proposed filters. However, the facts set forth in Paragraph 1379 are disputed because the cited evidence is incorrect. Specifically, OpenAI disputes that those records have been shown to be copies of Orange County Register Asserted Works created from webpages containing work-specific CMI that was subsequently absent from training data records for the reasons described in OpenAI's response to Fact 1363. Pursuant to Fed. R. Evid. 106, Dr. Krishnan's broader analysis verified that at least 826,906

	of Dr. Goldstein's 880,928 individual alleged Orange County Register CMI removals—93.9%—were erroneous, without accounting for all matching and WARC-selection errors. S. Krishnan Opp. Decl. (News) Ex. A (S. Krishnan Rebuttal Rep. (Goldstein)) ¶¶ 68–131, 153–232, Table 6 & Table 15. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1380. Table 20 provided with the Goldstein Declaration Exhibit K accurately states the number of OpenAI training data records which match to an Orange County Register Asserted Work with a coverage score of at least 80%, which were created by processing a copy of an Orange County Register webpage containing, respectively, (1) publisher, (2) title, (3) author, (4) publication date, (5) terms of service, and (6) copyright notice information conveyed in connection with the Asserted Work, and for which the corresponding (1) publisher, (2) title, (3) author, (4) publication date, (5) terms of service, and (6) copyright notice information was absent from the training data record copy of the Asserted Work. Goldstein Declaration, Exhibits K and L; Maisel Declaration, Exhibits C and D (CMI Removal Rule 1006 summaries).	1380. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: OpenAI does not dispute Table 20 accurately reports the outputs of Dr. Goldstein's methodology. However, the facts set forth in Paragraph 1380 are disputed because the cited evidence is incorrect. Specifically, OpenAI disputes that Table 20 accurately establishes the number of records satisfying the substantive descriptions in the statement. Pursuant to Fed. R. Evid. 106, Dr. Goldstein did not reliably establish that each listed record matched an Orange County Asserted Work, was created by processing the WARC file he selected, originally contained work-specific CMI, or omitted that information from the complete training-data record. S. Krishnan Opp. Decl. (News) Ex. A (S. Krishnan Rebuttal Rep. (Goldstein)) ¶¶ 153–232, Table 6 & Table 15. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.

1381. OpenAI failed to retain a copyright notice in 99.9% of OpenAI training data records which match to a Times Asserted Work with a coverage score of at least 80%, which were created by processing a copy of a Times webpage containing a copyright notice conveyed in connection with the Asserted Work. Goldstein Declaration, Exhibits K and L; Maisel Declaration, Exhibits C and D (CMI Removal Rule 1006 summaries).	1381. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: OpenAI does not dispute that Dr. Goldstein's table reports a 99.9% rate under his proposed methodology. OpenAI otherwise disputes the statement as inaccurate. Specifically, OpenAI disputes that 99.9% of the identified records reflect OpenAI's failure to retain a valid, work-specific copyright notice for the reasons described in OpenAI's response to Fact 1363. Pursuant to Fed. R. Evid. 106, Dr. Krishnan's broader analysis verified that at least 3,721,881 of Dr. Goldstein's 3,754,577 individual alleged Times copyright-notice removals—99.1%—were erroneous, without accounting for all matching and WARC-selection errors. S. Krishnan Opp. Decl. (News) Ex. A (S. Krishnan Rebuttal Rep. (Goldstein)) ¶¶ 68–131, 153–232, Table 6 & Table 15. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1382. OpenAI failed to retain a copyright notice in 99.4% of OpenAI training data records which match to a Mercury News Asserted Work with a coverage score of at least 80%, which were created by processing a copy of a Mercury News webpage containing a copyright notice conveyed in connection with the Asserted Work. Goldstein Declaration, Exhibits K and L; Maisel Declaration, Exhibits C and D (CMI Removal Rule 1006 summaries).	1382. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: OpenAI does not dispute that Dr. Goldstein's table reports a 99.4% rate under Dr. Goldstein's proposed methodology. OpenAI otherwise disputes the statement as inaccurate. Specifically, OpenAI disputes that 99.4% of the identified records have been shown to reflect OpenAI's failure to retain a valid, work-specific copyright notice for the reasons described in OpenAI's response to Fact 1363. Pursuant to Fed. R. Evid. 106, Dr. Krishnan's

	broader analysis verified that at least 682,411 of Dr. Goldstein's 686,933 individual alleged Mercury News copyright-notice removals—99.3%—were erroneous, without accounting for all matching and WARC-selection errors. S. Krishnan Opp. Decl. (News) Ex. A (S. Krishnan Rebuttal Rep. (Goldstein)) ¶¶ 68–131, 153–232, Table 6 & Table 15. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1383. OpenAI failed to retain a copyright notice in 99.9% of OpenAI training data records which match to a Chicago Tribune Asserted Work with a coverage score of at least 80%, which were created by processing a copy of a Chicago Tribune webpage containing a copyright notice conveyed in connection with the Asserted Work. Goldstein Declaration, Exhibits K and L; Maisel Declaration, Exhibits C and D (CMI Removal Rule 1006 summaries).	1383. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: OpenAI does not dispute that Dr. Goldstein's table reports a 99.9% rate under Dr. Goldstein's proposed methodology. OpenAI otherwise disputes the statement as inaccurate. Specifically, OpenAI disputes that 99.9% of the identified records have been shown to reflect OpenAI's failure to retain a valid, work-specific copyright notice for the reasons described in OpenAI's response to Fact 1363. Pursuant to Fed. R. Evid. 106, Dr. Krishnan's broader analysis verified that at least 623,867 of Dr. Goldstein's 691,957 individual alleged Chicago Tribune copyright-notice removals—90.2%—were erroneous, without accounting for all matching and WARC-selection errors. S. Krishnan Opp. Decl. (News) Ex. A (S. Krishnan Rebuttal Rep. (Goldstein)) ¶¶ 68–131, 153–232, Table 6 & Table 15.

	Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1384. OpenAI failed to retain a copyright notice in 99.5% of OpenAI training data records which match to a Denver Post Asserted Work with a coverage score of at least 80%, which were created by processing a copy of a Denver Post webpage containing a copyright notice conveyed in connection with the Asserted Work. Goldstein Declaration, Exhibits K and L; Maisel Declaration, Exhibits C and D (CMI Removal Rule 1006 summaries).	1384. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: OpenAI does not dispute that Dr. Goldstein's table reports a 99.5% rate under Dr. Goldstein's proposed methodology. OpenAI otherwise disputes the statement as inaccurate. Specifically, OpenAI disputes that 99.5% of the identified records have been shown to reflect OpenAI's failure to retain a valid, work-specific copyright notice for the reasons described in OpenAI's response to Fact 1363. Dr. Krishnan's broader analysis verified that at least 600,796 of Dr. Goldstein's 608,902 individual alleged Denver Post copyright-notice removals—98.7%—were erroneous, without accounting for all matching and WARC-selection errors. S. Krishnan Opp. Decl. (News) Ex. A (S. Krishnan Rebuttal Rep. (Goldstein)) ¶¶ 68–131, 153–232, Table 6 & Table 15; Fed. R. Evid. 106. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1385. OpenAI failed to retain a copyright notice in 100.0% of OpenAI training data records which match to an Orlando Sentinel Asserted Work with a coverage score of at least 80%, which were created by processing a copy of an Orlando Sentinel webpage containing a copyright notice conveyed in connection with the Asserted Work. Goldstein	1385. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: OpenAI does not dispute that Dr. Goldstein's table reports a 100.0% rate under Dr. Goldstein's proposed methodology. OpenAI otherwise disputes the statement as inaccurate. Specifically, OpenAI disputes that 100.0% of the identified records have been shown to reflect OpenAI's

Declaration, Exhibits K and L; Maisel Declaration, Exhibits C and D (CMI Removal Rule 1006 summaries).	failure to retain a valid, work-specific copyright notice for the reasons described in OpenAI's response to Fact 1363. Pursuant to Fed. R. Evid. 106, Dr. Krishnan's broader analysis verified that at least 364,419 of Dr. Goldstein's 457,641 individual alleged Orlando Sentinel copyright-notice removals—79.6%—were erroneous, without accounting for all matching and WARC-selection errors. S. Krishnan Opp. Decl. (News) Ex. A (S. Krishnan Rebuttal Rep. (Goldstein)) ¶¶ 68–131, 153–232, Table 6 & Table 15. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1386. OpenAI failed to retain a copyright notice in 100.0% of OpenAI training data records which match to a New York Daily News Asserted Work with a coverage score of at least 80%, which were created by processing a copy of a New York Daily News webpage containing a copyright notice conveyed in connection with the Asserted Work. Goldstein Declaration, Exhibits K and L; Maisel Declaration, Exhibits C and D (CMI Removal Rule 1006 summaries).	1386. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: OpenAI does not dispute that Dr. Goldstein's table reports a 100.0% rate under Dr. Goldstein's proposed methodology. OpenAI otherwise disputes the statement as inaccurate. OpenAI disputes that 100.0% of the identified records have been shown to reflect OpenAI's failure to retain a valid, work-specific copyright notice for the reasons described in OpenAI's response to Fact 1363. Pursuant to Fed. R. Evid. 106, Dr. Krishnan's broader analysis verified that at least 358,144 of Dr. Goldstein's 406,239 individual alleged New York Daily News copyright-notice removals—88.2%—were erroneous, without accounting for all matching and WARC-selection errors. S. Krishnan Opp. Decl. (News) Ex.

	A (S. Krishnan Rebuttal Rep. (Goldstein)) ¶¶ 68–131, 153–232, Table 6 & Table 15. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1387. OpenAI failed to retain a copyright notice in 100.0% of OpenAI training data records which match to a Sun Sentinel Asserted Work with a coverage score of at least 80%, which were created by processing a copy of a Sun Sentinel webpage containing a copyright notice conveyed in connection with the Asserted Work. Goldstein Declaration, Exhibits K and L; Maisel Declaration, Exhibits C and D (CMI Removal Rule 1006 summaries).	1387. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: OpenAI does not dispute that Dr. Goldstein's table reports a 100.0% rate under Dr. Goldstein's proposed methodology. OpenAI otherwise disputes the statement as inaccurate. Specifically, OpenAI disputes that 100.0% of the identified records have been shown to reflect OpenAI's failure to retain a valid, work-specific copyright notice for the reasons described in OpenAI's response to Fact 1363. Pursuant to Fed. R. Evid. 106, Dr. Krishnan's broader analysis verified that at least 207,331 of Dr. Goldstein's 243,748 individual alleged Sun Sentinel copyright-notice removals—85.1%—were erroneous, without accounting for all matching and WARC-selection errors. S. Krishnan Opp. Decl. (News) Ex. A (S. Krishnan Rebuttal Rep. (Goldstein)) ¶¶ 68–131, 153–232, Table 6 & Table 15. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1388. OpenAI failed to retain a copyright notice in 99.9% of OpenAI training data records which match to a Pioneer Press Asserted Work with a coverage score of at	1388. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: OpenAI does not dispute that Dr. Goldstein's table reports a 99.9%

least 80%, which were created by processing a copy of a Pioneer Press webpage containing a copyright notice conveyed in connection with the Asserted Work. Goldstein Declaration, Exhibits K and L; Maisel Declaration, Exhibits C and D (CMI Removal Rule 1006 summaries).	rate under Dr. Goldstein's proposed methodology. OpenAI otherwise disputes the statement as inaccurate. OpenAI disputes that 99.9% of the identified records have been shown to reflect OpenAI's failure to retain a valid, work-specific copyright notice for the reasons described in OpenAI's response to Fact 1363. Pursuant to Fed. R. Evid. 106, Dr. Krishnan's broader analysis verified that at least 222,535 of Dr. Goldstein's 222,810 individual alleged Pioneer Press copyright-notice removals—99.9%—were erroneous, without accounting for all matching and WARC-selection errors. S. Krishnan Opp. Decl. (News) Ex. A (S. Krishnan Rebuttal Rep. (Goldstein)) ¶¶ 68–131, 153–232, Table 6 & Table 15. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1389. OpenAI failed to retain a copyright notice in 100.0% of OpenAI training data records which match to an Orange County Register Asserted Work with a coverage score of at least 80%, which were created by processing a copy of an Orange County Register webpage containing a copyright notice conveyed in connection with the Asserted Work. Goldstein Declaration, Exhibits K and L; Maisel Declaration, Exhibits C and D (CMI Removal Rule 1006 summaries).	1389. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: OpenAI does not dispute that Dr. Goldstein's table reports a 100.0% rate under Dr. Goldstein's proposed methodology. OpenAI otherwise disputes the statement as inaccurate. Specifically, OpenAI disputes that 100.0% of the identified records have been shown to reflect OpenAI's failure to retain a valid, work-specific copyright notice for the reasons described in OpenAI's response to Fact 1363. Pursuant to Fed. R. Evid. 106, Dr. Krishnan's broader analysis verified that at least 52,616 of Dr. Goldstein's 52,668 individual alleged Orange County

	Register copyright-notice removals—99.9%—were erroneous, without accounting for all matching and WARC-selection errors. S. Krishnan Opp. Decl. (News) Ex. A (S. Krishnan Rebuttal Rep. (Goldstein)) ¶¶ 68–131, 153–232, Table 6 & Table 15. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.

3. OpenAI's Removal of Ziff Davis CMI

1390. At least 6,271,629 OpenAI training data records match the text of at least 1,636,586 Ziff Davis Asserted Works, including Exact Matches and Revised Versions of a Ziff Davis Asserted Work. Delorey Supplemental Report ¶¶ 34–35.	1390. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: OpenAI does not dispute that Mr. Delorey's tables report 6,271,629 training-data records associated with 1,636,586 purported Ziff Davis matches under his proposed methodology. Otherwise OpenAI disputes the statement as incorrect. Specifically, OpenAI disputes that those records have been reliably shown to "match" Ziff Davis Asserted Works or to constitute "Exact Matches" or "Revised Versions." Mr. Delorey's categories depend on unvalidated word-level similarity and containment heuristics that are sensitive to common vocabulary, boilerplate, document length, and third-party content. Pursuant to Fed. R. Evid. 106, his Asserted Works corpus also retained duplicates, wire-service reports, press releases, and other third-party material. Indeed, more than 90% of the works Mr. Delorey characterized as "exact" copies were identified through

	non-exact similarity heuristics. News Ex. 1167 (S. Krishnan Rebuttal Rep. (Delorey)) ¶¶ 24–25, 60–118. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1391. Of the 6,271,629 OpenAI training data records that match a Ziff Davis Asserted Work, at least 6,017,818 matches are found in OpenAI's Common Crawl datasets, including 67,457 matches in the midtraining dataset, and another 5,950,361 in the other Common Crawl datasets. Delorey Supplemental Report ¶ 38, n. 21.	1391. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: OpenAI does not dispute that Mr. Delorey's tables assign 6,017,818 of his reported results to Common Crawl-derived datasets, including the listed dataset-level figures. The statement is otherwise disputed as inaccurate. Specifically, OpenAI disputes that those results have been reliably shown to be "matches" to Ziff Davis Asserted Works. Pursuant to Fed. R. Evid. 106, Mr. Delorey's figures incorporate false positives caused by an inadequately audited Asserted Works corpus and unvalidated word-level similarity and containment heuristics, including purported matches attributable to duplicate works, wire-service reports, press releases, quotations, and other third-party content. News Ex. 1167 (S. Krishnan Rebuttal Rep. (Delorey)) ¶¶ 24–25, 60–118. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1392. As explained above, OpenAI created these training data records from its library of WARC files, which OpenAI downloaded from Common Crawl and stored on an azure storage account. Delorey Opening Report ¶¶ 85-91,	1392. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: OpenAI does not dispute that it downloaded Common Crawl data in WARC format. Otherwise OpenAI disputes the

156-57, 212, 215, 224, 229-31, 254-263, 270; Delorey Supplemental Report n.23 Ex. 521 Yuan 11/18/2025 Dep. Tr. at 13:14-16, 13:25-14:2; 21:22-27:21; Ex. 631 [Barbara Frederiksen-Cross Dep. Tr.] 129:4–23.	statement as inaccurate. Specifically, OpenAI disputes that Mr. Delorey identified the particular WARC files OpenAI downloaded or established that any particular training-data record was derived from a particular WARC file containing a particular Ziff Davis Asserted Work. Pursuant to Fed. R. Evid. 106, Mr. Delorey instead downloaded substitute WARC files from Common Crawl and selected the capture nearest the inferred training-data cutoff; 15% of his selected files postdated that cutoff, and Dr. Krishnan estimated that his methodology selected the wrong WARC file approximately 75% of the time. OpenAI also disputes the characterization of the stored data as a "library." Evidence that raw Common Crawl data was stored in a cloud-storage account for processing does not establish that OpenAI created or maintained a curated "library" of works. News Ex. 1167 (S. Krishnan Rebuttal Rep. (Delorey)) ¶¶ 140–151. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1393. As noted, OpenAI did not produce its copy of these WARC files and, instead, directed News Plaintiffs to download WARC files directly from Common Crawl. Dkt. 569; Delorey Reply Report ¶ 51.	1393. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 1393 are disputed because they are not supported by the cited evidence. OpenAI disputes the implication that it possessed but failed to produce copies of the WARC files. Specifically, OpenAI provided the applicable training-data cutoff dates and directed Plaintiffs

	to Common Crawl, from which the WARC files could be downloaded directly. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1394. From the set of confirmed matches of Ziff Davis Asserted Works found in OpenAI's training data in Common Crawl datasets, 3,468,909 confirmed matches had corresponding Common Crawl captures with at least one CMI element (*e.g.*, publisher, title, author, publication date, terms of service, or copyright notice) conveyed in connection with the Asserted Work. Delorey Supplemental Report ¶ 40.	1394. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: OpenAI does not dispute that Mr. Delorey reports 3,468,909 records satisfying his proposed filters. OpenAI otherwise disputes the statement as unsupported by the cited evidence. Specifically, OpenAI disputes that those records have been shown to be confirmed copies of Ziff Davis Asserted Works derived from the selected Common Crawl captures, or that the selected captures contained work-specific CMI conveyed in connection with those works. Pursuant to Fed. R. Evid. 106, Mr. Delorey relied on unvalidated matching criteria, did not identify the WARC captures actually used to create particular training-data records, and frequently treated sitewide or unrelated webpage material as CMI associated with a particular work. News Ex. 1167 (S. Krishnan Rebuttal Rep. (Delorey)) ¶¶ 24–28, 60–118, 140–200. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1395. For at least 3,462,149 (99.8%) of the 3,468,909 confirmed matches of Ziff Davis Asserted Works found in OpenAI's training data, for which the	1395. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: OpenAI does not dispute that Mr. Delorey reports 3,462,149

corresponding Common Crawl page contained at least one CMI element, OpenAI removed at least one CMI element from the corresponding training data records. Delorey Supplemental Report ¶¶40-41.	records satisfying his proposed filters. OpenAI disputes that those records reflect actual removal of CMI from copies of Ziff Davis Asserted Works. Specifically, Mr. Delorey did not reliably establish the asserted work, the corresponding training-data record, the WARC capture from which that record was created, whether the identified material constituted work-specific CMI, or whether that material was absent from the training-data record as a whole. Pursuant to Fed. R. Evid. 106, Dr. Krishnan explained that these errors are cumulative and that it is impossible to determine which alleged removals, if any, would remain after all errors were corrected. News Ex. 1167 (S. Krishnan Rebuttal Rep. (Delorey)) ¶¶ 60–118, 140–217. Furthermore, OpenAI's text extractors do not "remove" anything. Rather, they extract material that is suitable for model training. Not extracting data is not equivalent to removing data. *See, e.g.*, ECF No. 1748 (C. Callison-Burch Decl.) ¶ 88 ("Importantly, the purpose of the text extraction process used by OpenAI and other NLP practitioners is to identify text to keep, not text to remove. The scripts are designed to extract readable natural language text from HTML files for use as training data. Any loss of CMI during this process is an incidental byproduct of the text extraction methodology—it is not a purposeful or intentional removal of CMI."). Microsoft responds that these facts do not pertain to claims asserted against Microsoft.

1396. Further, for at least 2,043,072 (58.9%) of the 3,468,909 confirmed matches of Ziff Davis Asserted Works found in OpenAI's training data, for which the corresponding Common Crawl page contained at least one CMI element, OpenAI removed all the CMI elements from the corresponding training data records. Delorey Supplemental Report ¶¶40-41.	1396. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: OpenAI does not dispute that Mr. Delorey reports 2,043,072 records satisfying his proposed filters. OpenAI disputes that his analysis establishes that OpenAI removed all CMI from those records. Among other defects, Mr. Delorey searched for selected strings using exact-substring comparisons against only a selected text field, while disregarding URLs, metadata, and other fields in which the purported information remained. Pursuant to Fed. R. Evid. 106, his analysis also relied on unvalidated matches, unverified WARC selections, and material that did not necessarily constitute work-specific CMI. News Ex. 1167 (S. Krishnan Rebuttal Rep. (Delorey)) ¶¶ 60–118, 140–217. Furthermore, OpenAI's text extractors do not "remove" anything. Rather, they extract material that is suitable for model training. Not extracting data is not equivalent to removing data. *See, e.g.*, ECF No. 1748, C. Callison-Burch Decl. ¶ 88 ("Importantly, the purpose of the text extraction process used by OpenAI and other NLP practitioners is to identify text to keep, not text to remove. The scripts are designed to extract readable natural language text from HTML files for use as training data. Any loss of CMI during this process is an incidental byproduct of the text extraction methodology—it is not a purposeful or intentional removal of CMI."). Microsoft responds that these facts do not pertain to claims asserted against Microsoft.

1397. The table below illustrates certain CMI elements that were conveyed in connection with a Ziff Davis Asserted Work in the Common Crawl WARC files, the number of instances where that particular CMI element may have been preserved in a matching OpenAI training data copies, and the minimum number (expressed as a percentage) of that particular CMI element that was removed from the Ziff Davis Asserted Works. Delorey Supplemental Report ¶ 41.

Element	OpenAI Training Data Matches for Ziff Davis Asserted Works Corresponding to Common Crawl Captures with this CMI element	OpenAI Training Data Matches preserving this CMI element	Percentage of CMI stripped
Any copyright notice	3,211,980	6,266	99.80%
Terms of service reference	2,318,115	40,355	98.26%
Title	3,129,854	611,389	80.47%
Author attribution	1,929,836	456,943	76.32%

Element	OpenAI Training Data Matches for Ziff Davis Asserted Works Corresponding to Common Crawl Captures with this CMI element	OpenAI Training Data Matches preserving this CMI element	Percentage of CMI stripped
Publication identity	2,533,159	1,069,815	57.77%

OpenAI stripped CMI from Ziff Davis's works on a massive scale. Delorey Supplemental Report ¶ 16

1397. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: OpenAI does not dispute that Mr. Delorey's table reports the outputs of his methodology. However, OpenAI disputes that the table establishes that the identified material was CMI conveyed in connection with a Ziff Davis Asserted Work, that the training-data records were copies of the identified works created from the selected WARC captures, or that the reported material was preserved or removed. Pursuant to Fed. R. Evid. 106, the table's labels embed those disputed conclusions. News Ex. 1167 (S. Krishnan Rebuttal Rep. (Delorey)) ¶¶ 60–118, 140–217.

Furthermore, OpenAI's text extractors do not "remove" anything. Rather, they extract material that is suitable for model training. Not extracting data is not equivalent to removing data. *See, e.g.*, ECF No. 1748 (C. Callison-Burch Decl.) ¶ 88 ("Importantly, the purpose of the text extraction process used by OpenAI and other NLP practitioners is to identify text to keep, not text to remove. The scripts are designed to extract readable natural language text from HTML files for use as training data. Any loss of CMI during this process is an incidental byproduct of the text extraction methodology—it is not a purposeful or intentional removal of CMI.").

Microsoft responds that these facts do not pertain to claims asserted against Microsoft.

1398. OpenAI failed to retain "any copyright notice" (including copyright notices without a holder name and those with an identifiable rights holder) in 99.8% of the 3,211,980 comparisons between OpenAI training data records and the corresponding Common Crawl captures of a Ziff Davis webpage containing a "copyright notice" conveyed in connection with the Asserted Work. Delorey Supplemental Report ¶42.	1398. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: OpenAI does not dispute that Mr. Delorey reports a 99.8% rate under his proposed methodology. However, OpenAI disputes that the reported comparisons establish removal of copyright notices conveyed in connection with particular Ziff Davis Asserted Works. Specifically, Mr. Delorey treated copyright notices appearing anywhere in flattened webpage content—including notices belonging to third parties and notices appearing indiscriminately in sitewide footers or navigation elements—as CMI associated with an Asserted Work. Pursuant to Fed. R. Evid. 106, in Dr. Krishnan's sample, 94% of the alleged copyright notices appeared in sitewide elements, and the notice year frequently did not correspond to the work's publication year. News Ex. 1167 (S. Krishnan Rebuttal Rep. (Delorey)) ¶¶ 152–188, Tables 11–12. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1399. OpenAI failed to retain a "terms of service" reference in 98.26% of the 2,318,115 comparisons between OpenAI training data records and the corresponding Common Crawl captures of a Ziff Davis webpage containing a "terms of service" reference conveyed in connection with the Asserted Work. Delorey Supplemental Report ¶ 41.	1399. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: OpenAI does not dispute that Mr. Delorey reports a 98.26% rate under his proposed methodology. However, OpenAI disputes that the reported comparisons establish removal of terms-of-service information conveyed in connection with particular Ziff Davis Asserted Works. Specifically, Mr. Delorey treated generic references to terms or privacy

	policies appearing anywhere in flattened webpage content as work-specific CMI. Pursuant to Fed. R. Evid. 106, Dr. Krishnan found that more than 99.9% of the alleged terms-of-service material in his sample appeared in sitewide headers, footers, or navigation elements applied indiscriminately across webpages. News Ex. 1167 (S. Krishnan Rebuttal Rep. (Delorey)) ¶¶ 152–188, Table 12. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1400. OpenAI failed to retain an article "title" in 80.47% of the 3,129,854 comparisons between OpenAI training data records and the corresponding Common Crawl captures of a Ziff Davis webpage containing a "title" conveyed in connection with the Asserted Work. Delorey Supplemental Report ¶ 41.	1400. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: OpenAI does not dispute that Mr. Delorey reports an 80.47% rate under his proposed methodology. However, OpenAI disputes that his analysis establishes removal of article titles from copies of Ziff Davis Asserted Works. Specifically, Mr. Delorey's title-extraction methodology could select text other than an article's title, and his preservation analysis relied on exact-string searches of a selected text field while ignoring other fields. Pursuant to Fed. R. Evid. 106, Dr. Krishnan identified 142,977 instances in which Mr. Delorey's extracted title differed substantially from the asserted title and 3,543,236 instances in which the purported title appeared in another training-data field. News Ex. 1167 (S. Krishnan Rebuttal (Delorey)) ¶¶ 189–217, Tables 13–14.

	Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1401. OpenAI failed to retain "author attribution" in 76.32% of the 1,929,836 comparisons between OpenAI training data records and the corresponding Common Crawl captures of a Ziff Davis webpage containing an "author attribution" conveyed in connection with the Asserted Work. Delorey Supplemental Report ¶ 41.	1401. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: OpenAI does not dispute that Mr. Delorey reports a 76.32% rate under his proposed methodology. However, OpenAI disputes that his analysis establishes removal of author attribution from copies of Ziff Davis Asserted Works. Specifically, Mr. Delorey's author-extraction methodology generated erroneous attributions, and his preservation analysis relied on exact-string searches of a selected text field. Pursuant to Fed. R. Evid. 106, Dr. Krishnan identified 702,092 instances in which Mr. Delorey's extracted author differed substantially from the asserted author, as well as additional instances in which the author information appeared in another training-data field or was missed because of Mr. Delorey's strict matching rule. News Ex. 1167 (S. Krishnan Rebuttal Rep. (Delorey)) ¶¶ 189–217, Tables 13–14. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1402. OpenAI also failed to retain "publication identity" in 57.77% of the 2,533,159 comparisons between OpenAI training data records and the corresponding Common Crawl captures of a Ziff Davis webpage containing a "publication identity" conveyed in connection with the Asserted Work. Delorey Supplemental Report ¶ 41.	1402. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: OpenAI does not dispute that Mr. Delorey reports a 57.77% rate under his proposed methodology. However, OpenAI disputes that the og:site_name field Mr. Delorey labeled "publication identity" necessarily identifies the publisher

	or copyright owner of a particular Asserted Work, or otherwise constitutes work-specific CMI. OpenAI further disputes Mr. Delorey's preservation analysis because it ignored URLs, metadata, and other training-data fields. Pursuant to Fed. R. Evid. 106, Dr. Krishnan identified 3,487,652 instances in which the purported publisher information appeared in another field. News Ex. 1167 (S. Krishnan Rebuttal Rep. (Delorey)) ¶¶ 198–217, Table 14. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1403. The CMI removal findings are conservative because, among other things, (i) Mr. Delorey's detection methods credit CMI as preserved whenever it appears even incidentally in the Asserted Work body, even where the structured CMI tags or bylines themselves were stripped; (ii) the specific numbers identified here exclude hundreds of thousands of Asserted Works for which OpenAI did not produce the associated WARC files and those WARC files could not be readily identified; and (iii) Mr. Delorey applied other restrictions to identify the WARC files for purposes of this motion. Delorey Opening Report ¶¶ 262, 263, Delorey Rebuttal Report ¶¶ 15-21, 34.	1403. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 1403 are disputed because the cited evidence is incorrect. Specifically, these facts are disputed because OpenAI's text extractors do not "remove" anything. Rather, they extract material that is suitable for model training. Pursuant to Fed. R. Evid. 106, not extracting data is not equivalent to removing data. *See, e.g.*, ECF No. 1748 (C. Callison-Burch Decl.) ¶ 88 ("Importantly, the purpose of the text extraction process used by OpenAI and other NLP practitioners is to identify text to keep, not text to remove. The scripts are designed to extract readable natural language text from HTML files for use as training data. Any loss of CMI during this process is an incidental byproduct of the text extraction methodology—it is not a purposeful or intentional removal of CMI.") OpenAI disputes that Mr. Delorey's findings are

	conservative. His methodology systematically inflated the alleged removals by relying on unreliable matches, selecting WARC captures without establishing that OpenAI used them, treating sitewide and unrelated webpage material as work-specific CMI, and using brittle exact-string searches limited to a selected text field. Nor does the inability to identify a WARC file establish that additional CMI removal occurred. Pursuant to Fed. R. Evid. 106, Dr. Krishnan explained that these interlocking errors prevent determining which alleged copies or removals, if any, would survive a complete correction. News Ex. 1167 (S. Krishnan Rebuttal Rep. (Delorey)) ¶¶ 60–118, 140–217. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1404. OpenAI's experts did not conduct training data analysis to determine if copyright notices were in fact preserved in OpenAI's training data copies of Ziff Davis Asserted Works. Ex. 673 [Krishnan Dep. Tr.] at 435:7–20.	1404. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 1404 are disputed because the cited evidence is incorrect. Specifically, these facts are disputed because OpenAI disputes the implication that its experts did not analyze Mr. Delorey's allegations concerning copyright notices. Pursuant to Fed. R. Evid. 106, Dr. Krishnan examined the methodology underlying those allegations and found, among other things, that Mr. Delorey treated third-party and sitewide notices as work-specific CMI; that 94% of the sampled alleged notices appeared in sitewide elements; and that the notice year frequently did not correspond to the work's publication year. Those defects

	undermine the premise that the identified material was a copyright notice conveyed in connection with the particular Asserted Work before any preservation analysis is undertaken. News Ex. 1167 (S. Krishnan Rebuttal Rep. (Delorey)) ¶¶ 152–188, Tables 11–12. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1405. Further, OpenAI expert Dr. Krishnan confirmed that after accounting for all his critiques, OpenAI's training data did not preserve the "author" conveyed in connection with Ziff Davis Asserted Works for 1,793,446 training data copies of Ziff Davis Asserted Works. Ex. 673 [Krishnan Dep. Tr.] at 422:14–423:3.	1405. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 1405 are disputed because the cited evidence is incorrect. Specifically, these facts are disputed because Dr. Krishnan did not "confirm" that OpenAI failed to preserve author information in 1,793,446 copies of Ziff Davis Asserted Works. That figure is labeled "Total Remaining" in Table 14 after limited corrective analyses; it is not a count of validated copies or confirmed removals after accounting for all of Dr. Krishnan's critiques. Pursuant to Fed. R. Evid. 106, Dr. Krishnan expressly explained that his examples were illustrative rather than exhaustive and that, because the errors are cumulative and interlocking, it is impossible to determine which alleged copies or CMI removals, if any, would remain after all errors were corrected. News Ex. 1167 (S. Krishnan Rebuttal Rep. (Delorey)) ¶¶ 207–217, Table 14. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.

1406. Dr. Krishnan also confirmed that after accounting for all his critiques, OpenAI's training data did not preserve the "title" conveyed in connection with Ziff Davis Asserted Works for 715,285 training data copies of Ziff Davis Asserted Works. Ex. 673 [Krishnan Dep. Tr.] at 423:8–17.	1406. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 1406 are disputed because the cited evidence is incorrect. Specifically, these facts are disputed because Dr. Krishnan did not "confirm" that OpenAI failed to preserve title information in 715,285 copies of Ziff Davis Asserted Works. That figure is labeled "Total Remaining" in Table 14 after limited corrective analyses; it is not a count of validated copies or confirmed removals after accounting for all of Dr. Krishnan's critiques. Pursuant to Fed. R. Evid. 106, Dr. Krishnan expressly explained that his examples were illustrative rather than exhaustive and that, because the errors are cumulative and interlocking, it is impossible to determine which alleged copies or CMI removals, if any, would remain after all errors were corrected. News Ex. 1167 (S. Krishnan Rebuttal Rep. (Delorey)) ¶¶ 207–217, Table 14. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1407. Dr. Krishnan also confirmed that after accounting for all his critiques, OpenAI's training data did not preserve the "publisher" conveyed in connection with Ziff Davis Asserted Works for 184,153 training data copies of Ziff Davis Asserted Works. Ex. 673 [Krishnan Dep. Tr.] at 423:20–424:2.	1407. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 1407 are disputed because the cited evidence is incorrect. Specifically, these facts are disputed because Krishnan did not "confirm" that OpenAI failed to preserve publisher information in 184,153 copies of Ziff Davis Asserted Works. That figure is labeled "Total Remaining" in Table 14 after limited corrective analyses; it is not a count

	of validated copies or confirmed removals after accounting for all of Dr. Krishnan's critiques. Moreover, the information Mr. Delorey characterized as "publisher" was generally an og:site_name value that identified a website and did not necessarily identify the publisher or copyright owner of the particular work. Pursuant to Fed. R. Evid. 106, Dr. Krishnan expressly stated that it is impossible to determine which alleged copies or CMI removals, if any, would remain after all errors were corrected. News Ex. 1167 (S. Krishnan Rebuttal Rep. (Delorey)) ¶¶ 198–217, Table 14. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.

4. OpenAI's Removal of CIR's and The Intercept's CMI

1408. Appendix 198 to Dr. Wenger's Supplemental Report accurately states, for each of The Intercept's SJ Asserted Works, and for each of OpenAI's Nonprofit News Plaintiffs' Common Crawl Training Sets excerpts of which were produced to The Intercept, the Jaccard similarity score over 3-grams between each SJ Accused Copy and the original, and whether author, title, copyright notice, and terms of use information was absent from that copy. Wenger Supplemental Rep. at 22-23; Ex. 1172 [Wenger Supplemental Rep. Appendix 198]; Ex. 985 [OPCO_INTERCEPT_0000001]; Ex. 987 [OPCO_INTERCEPT_0000002]; Ex. 989	1408. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: OpenAI does not dispute that Appendix 198 reports the 3-gram Jaccard scores generated by Dr. Wenger's methodology and her classifications concerning whether the exact title string she searched for appeared in the selected text. OpenAI disputes that the appendix accurately identifies "SJ Accused Copies," the relevant "original," or the actual absence of title information from a copy. Pursuant to Fed. R. Evid. 106, Dr. Wenger's unvalidated Jaccard methodology cannot reliably distinguish copies from false positives, and her exact-string search ignored title

[OPCO_INTERCEPT_0000003]; Ex. 991 [OPCO_INTERCEPT_0000004]; Ex. 993 [OPCO_INTERCEPT_0000005]; Ex. 995 [OPCO_INTERCEPT_0000006]; Ex. 997 [OPCO_INTERCEPT_0000007]; Ex. 999 [OPCO_INTERCEPT_0000008]; Ex. 1001 [OPCO_INTERCEPT_0000009]; Ex. 1003 [OPCO_INTERCEPT_0000010]; Ex. 1005 [OPCO_INTERCEPT_0000011]; Macala Decl. Ex. B; Macala Decl. Ex. C.	information retained in URLs and other fields or expressed with minor variations. *See* S. Krishnan Opp. Decl. (News) Ex. H (S. Krishnan Rebuttal Rep. (Wenger)) ¶¶ 82–118, 142–188, 201–205. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1409. Appendix 199 to Dr. Wenger's Supplemental Report accurately states, for each of the Mother Jones SJ Asserted Works, and for each of OpenAI's Nonprofit News Plaintiffs' Common Crawl Training Sets excerpts of which were produced to CIR, the Jaccard similarity score over 3-grams between each SJ Accused Copy and the original, and whether author, title, copyright notice, and terms of service information was absent from that copy. Wenger Supplemental Rep. at 22-23; Ex. 1173 [Wenger Supplemental Rep. Appendix 199]; Ex. 963 [OPCO_NEWS_1144089]; Ex. 965 [OPCO_NEWS_1144090]; Ex. 967 [OPCO_NEWS_1144091]; Ex. 969 [OPCO_NEWS_1144092]; Ex. 971 [OPCO_NEWS_1144093]; Ex. 973 [OPCO_NEWS_1144094]; Ex. 975 [OPCO_NEWS_1144095]; Ex. 977 [OPCO_NEWS_1144096]; Ex. 979	1409. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: OpenAI does not dispute that Appendix 199 reports the 3-gram Jaccard scores generated by Dr. Wenger's methodology and her classifications concerning whether the exact title string she searched for appeared in the selected text. OpenAI disputes that the appendix accurately identifies "SJ Accused Copies," the relevant "original," or the actual absence of title information from a copy. Pursuant to Fed. R. Evid. 106, Dr. Wenger's unvalidated Jaccard methodology cannot reliably distinguish copies from false positives, and her exact-string search ignored title information retained in URLs and other fields or expressed with minor variations. *See* S. Krishnan Opp. Decl. (News) Ex. H (S. Krishnan Rebuttal Rep. (Wenger)) ¶¶ 82–118, 142–188, 201–205.

[OPCO_NEWS_1144097]; Ex. 981 [OPCO_NEWS_1144098]; Ex. 983 [OPCO_NEWS_1144099]; Bauerlein Decl. Ex. D; Bauerlein Decl. Ex. E; Bauerlein Decl. Ex. F; Bauerlein Decl. Ex. G.	Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1410. Appendix 201 to Dr. Wenger's Supplemental Report accurately states, for each of the Reveal SJ Asserted Works, and for each of OpenAI's Nonprofit News Plaintiffs' Common Crawl Training Sets excerpts of which were produced to CIR, the Jaccard similarity score over 3-grams between each SJ Accused Copy and the original, and whether title information was absent from that copy. Wenger Supplemental Rep. at 22-23; Ex. 1174 [Wenger Supplemental Rep. Appendix 201]; Ex. 963 [OPCO_NEWS_1144089]; Ex. 965 [OPCO_NEWS_1144090]; Ex. 967 [OPCO_NEWS_1144091]; Ex. 969 [OPCO_NEWS_1144092]; Ex. 971 [OPCO_NEWS_1144093]; Ex. 973 [OPCO_NEWS_1144094]; Ex. 975 [OPCO_NEWS_1144095]; Ex. 977 [OPCO_NEWS_1144096]; Ex. 979 [OPCO_NEWS_1144097]; Ex. 981 [OPCO_NEWS_1144098]; Ex. 983 [OPCO_NEWS_1144099]; Bauerlein Decl. Ex. D; Bauerlein Decl. Ex. E; Bauerlein Decl. Ex. F; Bauerlein Decl. Ex. G.	1410. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: OpenAI does not dispute that Appendix 201 reports the 3-gram Jaccard scores generated by Dr. Wenger's methodology and her classifications concerning whether the exact title string she searched for appeared in the selected text. OpenAI disputes that the appendix accurately identifies "SJ Accused Copies," the relevant "original," or the actual absence of title information from a copy. Dr. Wenger's unvalidated Jaccard methodology cannot reliably distinguish copies from false positives, and her exact-string search ignored title information retained in URLs and other fields or expressed with minor variations. Pursuant to Fed. R. Evid. 106, *see* S. Krishnan Opp. Decl. (News) Ex. H (S. Krishnan Rebuttal Rep. (Wenger)) ¶¶ 82–118, 142–188, 201–205. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1411. OpenAI removed at least one piece of CMI from at least 7,064 different Intercept SJ Asserted Works.	1411. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 1411 are disputed because the cited evidence

Wenger Supplemental Rep. at 33; Ex. 1172 [Wenger Supplemental Rep. Appendix 198].	is incorrect. Specifically, OpenAI disputes that this establishes that OpenAI removed CMI from 7,064 different Intercept Asserted Works. Specifically, Dr. Wenger did not reliably establish that the records were copies of the identified works, that the particular source records acquired by OpenAI contained the asserted CMI, or that the searched-for information was absent from the complete training-data records. She also treated sitewide footer material as work-specific CMI. Pursuant to Fed. R. Evid. 106, *see* S. Krishnan Opp. Decl. (News) Ex. H (S. Krishnan Rebuttal Rep. (Wenger)). Further, these facts are disputed because OpenAI's text extractors do not "remove" anything. Rather, they extract material that is suitable for model training. Not extracting data is not equivalent to removing data. Pursuant to Fed. R. Evid. 106, *see, e.g.*, ECF No. 1748 (C. Callison-Burch Decl.) ¶ 88 ("Importantly, the purpose of the text extraction process used by OpenAI and other NLP practitioners is to identify text to keep, not text to remove. The scripts are designed to extract readable natural language text from HTML files for use as training data. Any loss of CMI during this process is an incidental byproduct of the text extraction methodology—it is not a purposeful or intentional removal of CMI."). Microsoft responds that these facts do not pertain to claims asserted against Microsoft.

1412. OpenAI removed at least one piece of CMI from at least 13,353 copies of Intercept SJ Asserted Works. Wenger Supplemental Rep. at 33; Ex. 1172 [Wenger Supplemental Rep. Appendix 198].	1412. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 1412 are disputed because the cited evidence is incorrect. Specifically, OpenAI disputes that those records have been shown to be copies of Intercept Asserted Works from which OpenAI removed CMI. Dr. Wenger's count rests on unreliable matching, assumes without establishing that the relevant source records contained the asserted CMI, and uses narrow searches of only part of each training-data record that overlook retained information. Pursuant to Fed. R. Evid. 106, *see* S. Krishnan Opp. Decl. (News) Ex. H (S. Krishnan Rebuttal Rep. (Wenger)) ¶¶ 119–205. Further, these facts are disputed because OpenAI's text extractors do not "remove" anything. Rather, they extract material that is suitable for model training. Not extracting data is not equivalent to removing data. Pursuant to Fed. R. Evid. 106, *see, e.g.*, ECF 1748 (C. Callison-Burch Decl.) ¶ 88 ("Importantly, the purpose of the text extraction process used by OpenAI and other NLP practitioners is to identify text to keep, not text to remove. The scripts are designed to extract readable natural language text from HTML files for use as training data. Any loss of CMI during this process is an incidental byproduct of the text extraction methodology—it is not a purposeful or intentional removal of CMI."). Microsoft responds that these facts do not pertain to claims asserted against Microsoft.

1413. OpenAI removed at least 46,142 pieces of CMI from copies of Intercept SJ Asserted Works. Wenger Supplemental Rep. at 33; Ex. 1172 [Wenger Supplemental Rep. Appendix 198].	1413. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 1413 are disputed because the cited evidence is incorrect. Specifically, OpenAI disputes that those classifications represent 46,142 "pieces of CMI" removed by OpenAI. The figure aggregates outputs of an unreliable pipeline that did not reliably establish the underlying copy, the presence of work-specific CMI in the actual source record, or the absence of the information from the complete training-data record. Pursuant to Fed. R. Evid. 106, *see* S. Krishnan Opp. Decl. (News) Ex. H (S. Krishnan Rebuttal Rep. (Wenger)) ¶¶ 119–205. Further, these facts are disputed because OpenAI's text extractors do not "remove" anything. Rather, they extract material that is suitable for model training. Not extracting data is not equivalent to removing data. Pursuant to Fed. R. Evid. 106, *see, e.g.*, ECF 1748 (C. Callison-Burch Decl.) ¶ 88 ("Importantly, the purpose of the text extraction process used by OpenAI and other NLP practitioners is to identify text to keep, not text to remove. The scripts are designed to extract readable natural language text from HTML files for use as training data. Any loss of CMI during this process is an incidental byproduct of the text extraction methodology—it is not a purposeful or intentional removal of CMI."). Microsoft responds that these facts do not pertain to claims asserted against Microsoft.

1414. OpenAI removed at least one piece of CMI from at least 37,360 different CIR SJ Asserted Works. Wenger Supplemental Rep. at 33; Ex. 1172 [Wenger Supplemental Rep. Appendix 198]; Ex. 1174 [Wenger Supplemental Rep. Appendix 201].	1414. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 1414 are disputed because the cited evidence is incorrect. Specifically, OpenAI disputes that this establishes that OpenAI removed CMI from 37,360 different CIR Asserted Works. Dr. Wenger did not reliably establish that the records were copies of the identified works, that the particular source records acquired by OpenAI contained the asserted CMI, or that the searched-for information was absent from the complete training-data records. She also treated sitewide footer material as work-specific CMI. Pursuant to Fed. R. Evid. 106, *see* S. Krishnan Opp. Decl. (News) Ex. H (S. Krishnan Rebuttal Rep. (Wenger)) ¶¶ 119–205. Further, these facts are disputed because OpenAI's text extractors do not "remove" anything. Rather, they extract material that is suitable for model training. Not extracting data is not equivalent to removing data. Pursuant to Fed. R. Evid. 106, *see, e.g.*, ECF 1748 (C. Callison-Burch Decl.) ¶ 88 ("Importantly, the purpose of the text extraction process used by OpenAI and other NLP practitioners is to identify text to keep, not text to remove. The scripts are designed to extract readable natural language text from HTML files for use as training data. Any loss of CMI during this process is an incidental byproduct of the text extraction methodology—it is not a purposeful or intentional removal of CMI.").

	Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1415. OpenAI removed at least one piece of CMI from at least 101,092 copies of CIR SJ Asserted Works. Wenger Supplemental Rep. at 33; Ex. 1172 [Wenger Supplemental Rep. Appendix 198]; Ex. 1174 [Wenger Supplemental Rep. Appendix 201].	1415. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 1415 are disputed because the cited evidence is incorrect. Specifically, these facts are disputed because OpenAI disputes that the appendix accurately identifies "SJ Accused Copies," the relevant "original," or the actual absence of title information from a copy. Dr. Wenger's unvalidated Jaccard methodology cannot reliably distinguish copies from false positives, and her exact-string search ignored title information retained in URLs and other fields or expressed with minor variations. Pursuant to Fed. R. Evid. 106, *see* S. Krishnan Opp. Decl. (News) Ex. H (S. Krishnan Rebuttal Rep. (Wenger)) ¶¶ 82–118, 142–188, 201–205. Further, these facts are disputed because OpenAI's text extractors do not "remove" anything. Rather, they extract material that is suitable for model training. Not extracting data is not equivalent to removing data. Pursuant to Fed. R. Evid. 106, *see, e.g.*, ECF 1748 (C. Callison-Burch Decl.) ¶ 88 ("Importantly, the purpose of the text extraction process used by OpenAI and other NLP practitioners is to identify text to keep, not text to remove. The scripts are designed to extract readable natural language text from HTML files for use as training data. Any loss of CMI during this process is an incidental

	byproduct of the text extraction methodology—it is not a purposeful or intentional removal of CMI."). Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1416. OpenAI removed at least 370,993 pieces of CMI from copies of Mother Jones SJ Asserted Works. Wenger Supplemental Rep. at 33; Ex. 1172 [Wenger Supplemental Rep. Appendix 198]; Ex. 1174 [Wenger Supplemental Rep. Appendix 201].	1416. **Disputed - Not Supported by Cited Evidence.** OpenAI responds as follows: The facts set forth in Paragraph 1416 are disputed because the cited evidence is incorrect. Specifically, OpenAI disputes that those classifications represent 370,993 "pieces of CMI" removed from copies of Mother Jones Asserted Works. The figure aggregates outputs of an unreliable pipeline that did not reliably establish the underlying copy, the presence of work-specific CMI in the actual source record, or the absence of the information from the complete training-data record. Pursuant to Fed. R. Evid. 106, *see* S. Krishnan Opp. Decl. (News) Ex. H (S. Krishnan Rebuttal Rep. (Wenger)) ¶¶ 119–205. Further, these facts are disputed because OpenAI's text extractors do not "remove" anything. Rather, they extract material that is suitable for model training. Not extracting data is not equivalent to removing data. Pursuant to Fed. R. Evid. 106, *see, e.g.*, ECF 1748 (C. Callison-Burch Decl.) ¶ 88 ("Importantly, the purpose of the text extraction process used by OpenAI and other NLP practitioners is to identify text to keep, not text to remove. The scripts are designed to extract readable natural language text from HTML files for use as training

	data. Any loss of CMI during this process is an incidental byproduct of the text extraction methodology—it is not a purposeful or intentional removal of CMI."). Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1417. The following table accurately states the number of SJ Accused Copies for which each type of The Intercept's CMI was either absent or present: CMI Type / Absent / Present Author / 12,762 / 591 Title / 6,682 / 6,671 Copyright Notice / 13,353 / 0 Terms of Use / 13,345 / 8 Wenger Supplemental Rep. at 29; Ex. 1172 [Wenger Supplemental Rep. Appendix 198].	1417. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: OpenAI does not dispute that the table reports Dr. Wenger's binary classifications for Intercept records under her proposed search methodology. OpenAI disputes that the "Absent" and "Present" columns accurately state whether CMI was absent from or present in copies of Intercept Asserted Works. Dr. Wenger used overly strict, unvalidated string searches of only selected text; ignored URLs and other fields; searched for only "all rights reserved" and "terms of service" as proxies for entire categories of CMI; and treated sitewide footer material as CMI conveyed with particular works. Pursuant to Fed R. Evid. 106, *see* S. Krishnan Opp. Decl. (News) Ex. H (S. Krishnan Rebuttal Rep. (Wenger)) ¶¶ 137–141, 175–197, 201–205. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1418. The following table accurately states the number of SJ Accused Copies for which each type of Mother Jones's CMI was either absent or present:	1418. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: OpenAI does not dispute that the table reports Dr. Wenger's binary classifications for Mother Jones records under her

CMI Type	Absent	Present
Author	87,003	4,409
Title	73,905	17,507
Copyright Notice	91,406	6
Terms of Use	91,325	87

Wenger Supplemental Rep. at 29; Ex. 1173 [Wenger Supplemental Rep. Appendix 199].

proposed search methodology. OpenAI disputes that the "Absent" and "Present" columns accurately state whether CMI was absent from or present in copies of Mother Jones Asserted Works. Dr. Wenger used overly strict, unvalidated string searches of only selected text; ignored URLs and other fields; searched for only "all rights reserved" and "terms of service" as proxies for entire categories of CMI; and treated sitewide footer material as CMI conveyed with particular works. Pursuant to Fed. R. Evid. 106, *see* S. Krishnan Opp. Decl. (News) Ex. H (S. Krishnan Rebuttal Rep. (Wenger)) ¶¶ 137–141, 175–197, 201–205.

Microsoft responds that these facts do not pertain to claims asserted against Microsoft.

1419. The following table accurately states the number of SJ Accused Copies for which each type of Reveal's CMI was either absent or present:

CMI Type	Absent	Present
Author	2,407	7,273
Title	5,593	4,087
Copyright Notice	9,680	0
Terms of Use	9,674	6

Wenger Supplemental Rep. at 29; Ex. 1174 [Wenger Supplemental Rep. Appendix 201].

1419. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: OpenAI does not dispute that the table reports Dr. Wenger's binary classifications for Reveal records under her proposed search methodology. OpenAI disputes that the "Absent" and "Present" columns accurately state whether CMI was absent from or present in copies of Reveal Asserted Works. Pursuant to Fed. R. Evid. 106, Dr. Wenger used overly strict, unvalidated string searches of only selected text; ignored URLs and other fields; searched for only "all rights reserved" and "terms of use" as proxies for entire categories of CMI; and treated sitewide footer material as CMI conveyed with particular works. S. Krishnan Opp.

	Decl. (News) Ex. H (S. Krishnan Rebuttal Rep. (Wenger)) ¶¶ 137–141, 175–197, 201–205. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1420. The following table accurately states the number of SJ Accused Copies for which each type of CIR's CMI was either absent or present: CMI Type / Absent / Present Author / 89,410 / 11,682 Title / 79,498 / 21,594 Copyright Notice / 101,086 / 6 Terms of Use / 100,999 / 93 Wenger Supplemental Rep. at 29; Ex. 1172 [Wenger Supplemental Rep. Appendix 198]; Ex. 1174 [Wenger Supplemental Report Appendix 201].	1420. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: OpenAI does not dispute that the table reports Dr. Wenger's binary classifications for CIR records under her proposed search methodology. OpenAI disputes that the "Absent" and "Present" columns accurately state whether CMI was absent from or present in copies of CIR Asserted Works. Pursuant to Fed. R. Evid. 106, Dr. Wenger used overly strict, unvalidated string searches of only selected text; ignored URLs and other fields; searched for only "all rights reserved" and "terms of use" as proxies for entire categories of CMI; and treated sitewide footer material as CMI conveyed with particular works. *See* S. Krishnan Opp. Decl. Ex. H (S. Krishnan Rebuttal (Wenger)) ¶¶ 137–141, 175–197, 201–205. Microsoft responds that these facts do not pertain to claims asserted against Microsoft.
1421. The tables at pages 31-32 of Dr. Wenger's Supplemental Report accurately state the number of SJ Accused Copies from which OpenAI removed author, title, copyright notice, and terms of use information from CIR's and The Intercept's SJ Asserted Works, broken down by	1421. **Disputed in Part - Not Supported by Cited Evidence.** OpenAI responds as follows: OpenAI does not dispute that the tables at pages 31–32 report Dr. Wenger's counts, broken down by training dataset, of records in which her methodology did not find selected

Nonprofit News Plaintiffs' Common Crawl Training Set. Wenger Supplemental Rep. at 31-32; Ex. 985 [OPCO_INTERCEPT_0000001]; Ex. 987 [OPCO_INTERCEPT_0000002]; Ex. 989 [OPCO_INTERCEPT_0000003]; Ex. 991 [OPCO_INTERCEPT_0000004]; Ex. 993 [OPCO_INTERCEPT_0000005]; Ex. 995 [OPCO_INTERCEPT_0000006]; Ex. 997 [OPCO_INTERCEPT_0000007]; Ex. 999 [OPCO_INTERCEPT_0000008]; Ex. 1001 [OPCO_INTERCEPT_0000009]; Ex. 1003 [OPCO_INTERCEPT_0000010]; Ex. 1005 [OPCO_INTERCEPT_0000011]; Macala Decl. Ex. B; Macala Decl. Ex. C; Ex. 963 [OPCO_NEWS_1144089]; Ex. 965 [OPCO_NEWS_1144090]; Ex. 967 [OPCO_NEWS_1144091]; Ex. 969 [OPCO_NEWS_1144092]; Ex. 971 [OPCO_NEWS_1144093]; Ex. 973 [OPCO_NEWS_1144094]; Ex. 975 [OPCO_NEWS_1144095]; Ex. 977 [OPCO_NEWS_1144096]; Ex. 979 [OPCO_NEWS_1144097]; Ex. 981 [OPCO_NEWS_1144098]; Ex. 983 [OPCO_NEWS_1144099]; Bauerlein Decl. Ex. D; Bauerlein Decl. Ex. E; Bauerlein Decl. Ex. F; Bauerlein Decl. Ex. G.	strings. OpenAI disputes that the tables accurately state the number of copies from which OpenAI "removed" CMI. Pursuant to Fed. R. Evid. 106, Dr. Wenger's methodology did not reliably establish that the records were copies of the identified works, that the actual source records acquired by OpenAI contained work-specific CMI, or that the information was absent from the complete training-data records. *See* S. Krishnan Opp. Decl. (News) Ex. H (S. Krishnan Rebuttal Rep. (Wenger)) ¶¶ 82–118, 119–205. Further, OpenAI's text extractors do not "remove" anything. Pursuant to Fed. R. Evid. 106, rather, they extract material that is suitable for model training. Not extracting data is not equivalent to removing data. *See, e.g.*, ECF No. 1748, C. Callison-Burch Decl. ¶ 88 ("Importantly, the purpose of the text extraction process used by OpenAI and other NLP practitioners is to identify text to keep, not text to remove. The scripts are designed to extract readable natural language text from HTML files for use as training data. Any loss of CMI during this process is an incidental byproduct of the text extraction methodology—it is not a purposeful or intentional removal of CMI."). Microsoft responds that these facts do not pertain to claims asserted against Microsoft.

Dated: October 23, 2026

Respectfully submitted,

*/s/ Allison S. Blanco**

LATHAM & WATKINS LLP
Andrew M. Gass (*pro hac vice*)
andrew.gass@lw.com
505 Montgomery Street, Suite 2000
San Francisco, CA 94111
Telephone: 415.391.0600

Sarang V. Damle
sy.damle@lw.com
Luke A. Budiardjo
luke.budiardjo@lw.com
1271 Avenue of the Americas
New York, NY 10020
Telephone: 212.906.1200

Elana Nightingale Dawson (*pro hac vice*)
elana.nightingaledawson@lw.com
555 Eleventh Street, NW, Suite 1000
Washington, D.C. 20004
Telephone: 202.637.2200

/s/ Joseph C. Gratz

MORRISON & FOERSTER LLP
Joseph C. Gratz (*pro hac vice*)
jgratz@mofo.com
Tiffany Cheung (*pro hac vice*)
tcheung@mofo.com
Caitlin Sinclaire Blythe (*pro hac vice*)
cblythe@mofo.com
425 Market Street
San Francisco, CA 94105
Telephone: 415.268.7000

Lily Li (*pro hac vice*)
YLi@mofo.com
755 Page Mill Road
Palo Alto, CA 94304
Telephone: 650.813.5600

/s/ Robert A. Van Nest

KEKER, VAN NEST & PETERS LLP
Robert A. Van Nest (*pro hac vice*)
rvannest@keker.com
R. James Slaughter (*pro hac vice*)
rslaughter@keker.com
Paven Malhotra
pmalhotra@keker.com
Michelle S. Ybarra (*pro hac vice*)
mybarra@keker.com
Reid P. Mullen (*pro hac vice*)
rmullen@keker.com
Nicholas S. Goldberg (*pro hac vice*)
ngoldberg@keker.com
633 Battery St.
San Francisco, CA 94111
Telephone: 415.391.5400

Attorneys for Defendant OpenAI

* All parties whose electronic signatures are included in this filing have provided consent in accordance with Rule 8.5(b) of the Court's ECF Rules and Instructions.

/s/ Lisa Simpson

ORRICK, HERRINGTON & SUTCLIFFE LLP
Annette L. Hurst (admitted pro hac vice)
Daniel Justice (admitted pro hac vice)
Robert Scott Shwarts
Catherine Lui (admitted pro hac vice)
Amanda H. Schwartz (admitted pro hac vice)
The Orrick Building
405 Howard Street
San Francisco, CA 94105-2669
Telephone: (415) 773-5700
Facsimile: (415) 773-5759
ahurst@orrick.com
djustice@orrick.com
aschwartz@orrick.com

Christopher J. Cariello
Lisa T. Simpson
Marc Shapiro
Vanessa Sorrentino
Hilda Obeng
51 West 52nd Street
New York, NY 10019
Telephone: (212) 506-3778
Facsimile: (212) 506-5151
ccariello@orrick.com
lsimpson@orrick.com
mrshapiro@orrick.com
vsorrentino@orrick.com
hobeng@orrick.com

Sheryl Koval Garko (admitted pro hac vice)
Laura Brooks Najemy (admitted pro hac vice)
222 Berkley Street
Boston, MA 02116
Telephone: (617) 880-1800
Facsimile: (617) 880-1801
sgarko@orrick.com
lnajemy@orrick.com

Isaac Sawlih Behnawa (admitted pro hac vice)
Johnathan Vaknin (admitted pro hac vice)
355 S. Grand Ave. Suite 2700
Los Angeles, CA 90071
Telephone: (213) 612-2179
Facsimile: (213) 612-2499
ibehnawa@orrick.com
jvaknin@orrick.com

Monica Svetoslavov (admitted pro hac vice)
Louis Andrew Schreiber
2100 Pennsylvania Avenue NW
Washington, DC 20037
Telephone: (202) 339-8400
Facsimile: (202) 339-8500
msvetoslavov@orrick.com
aschreiber@orrick.com

FAEGRE DRINKER BIDDLE & REATH LLP
Jeffrey S. Jacobson
1177 Avenue of the Americas
New York, New York 10036
Telephone: (212) 248-3191
jeffrey.jacobson@faegredrinker.com

Jared B. Briant (admitted pro hac vice)
Kirstin L. Stoll-DeBell (admitted pro hac vice)
Timothy Scull (admitted pro hac vice)
Shelby Pickar-Dennis (admitted pro hac vice)
1144 15th Street, Suite 3400
Denver, Colorado 80202
Telephone: (303) 607-3588
jared.briant@faegredrinker.com
kirstin.stolldebell@faegredrinker.com
timothy.scull@faegredrinker.com
shelby.pickar-dennis@faegredrinker.com

Lora A. Brzezynski (admitted pro hac vice)
Brianna Lynn Silverstein (admitted pro hac vice)
Faith Maggard (admitted pro hac vice)
1500 K Street NW, Suite 1100
Washington, DC 20005
Telephone: (202) 230-5000
Facsimile: (202) 842-8465
lora.brzezynski@faegredrinker.com
brianna.silverstein@faegredrinker.com
faith.maggard@faegredrinker.com

Carrie A. Beyer (admitted pro hac vice)
320 South Canal Street, Suite 3300
Chicago, IL 60606-5707
Telephone: (312) 569-1000
Facsimile: (312) 569-3000
carrie.beyer@faegredrinker.com

Elizabeth M.C. Scheibel (admitted pro hac vice)
2200 Wells Fargo Center, 90 S. 7th Street
Minneapolis, MN 55402
Telephone: (612) 766-7000
Facsimile: (612) 766-1600
elizabeth.scheibel@faegredrinker.com

Andrew Lawrence Van Houter
600 Campus Drive
Florham Park, NJ 07932
Telephone: (973) 549-7018
Facsimile: (973) 360-9831
Andrew.vanhouter@faegredrinker.com

Attorneys for Defendant Microsoft Corporation