**UNITED STATES DISTRICT COURT**
**SOUTHERN DISTRICT OF NEW YORK**

| | |
|---|---|
| IN RE OPENAI, INC., COPYRIGHT INFRINGEMENT LITIGATION<br><br>This document relates to:<br>No. 1:23-cv-08292<br>No. 1:23-cv-10211<br>No. 1:23-cv-11195<br>No. 1:24-cv-03285<br>No. 1:24-cv-04872<br>No. 1:24-cv-01515<br>No. 1:25-cv-04315 | Case No. 25-md-3143-SHS-OTW<br><br>Hon. Sidney H. Stein |

**PLAINTIFFS' OPPOSITION TO DEFENDANTS'**
**RULE 702 MOTION TO EXCLUDE EXPERT TESTIMONY**

# TABLE OF CONTENTS

**ARGUMENT** ..... 1

**I. BOOK PLAINTIFFS** ..... 1

**A. Zhao** ..... 1

**B. Lasinski** ..... 3

**C. Seeley** ..... 7

**II. NEWS PLAINTIFFS' TECHNICAL EXPERTS** ..... 10

**A. Goldstein** ..... 10

1. Text matching. ..... 10

2. CMI removal. ..... 11

*a. Goldstein validated his CMI removal analysis.* ..... 11

*b. Goldstein's search for CMI was not "deficient"* ..... 14

3. Microsoft's Copilot logs. ..... 15

*a. 4-gram counts.* ..... 15

*b. RAG leak analysis.* ..... 16

*c. Web-domain and "impressions" counts.* ..... 16

*d. Supplemental Report analysis.* ..... 17

4. Defendants' technical collaboration. ..... 18

**B. Wenger** ..... 19

1. Text matching. ..... 19

2. Keyword matching and summarization. ..... 20

3. CMI-related opinions. ..... 20

*a. CMI acquisition.* ..... 20

*b. CMI removal.* ..... 22

*c. "False positives."* ..... 22

*d. "CMI" retained in URLs.* ..... 22

4. Copilot extrapolations. ..... 23

**C. Delorey** ..... 23

1. Delorey's "Copies" Opinion is Reliable ..... 23

2. Delorey's CMI Analysis is Reliable ..... 25

**D. Etzioni** ..... 26

**III. NEWS PLAINTIFFS' INDUSTRY EXPERTS** ..... 27

**A. Stricchiola** ..... 27

1. Stricchiola's Click-Through-Rate Opinions Satisfy FRE 702. ..... 27

2. Stricchiola's Meta Tag Opinions Satisfy FRE 702. ..... 29

**B. Krein** ..... 31

**C. Bell, Owen, and Cross** ..... 33

1. Market Harm and Substitution Opinions ..... 33

2. Bell and Owen's Reliance on Audience Surveys ..... 35

3. Bell's News and "Pink Slime" Lists ..... 36

**D. Tease** ..... 36

**IV. NEWS PLAINTIFFS' HARM EXPERTS** ..... 37

**A. Malackowski** ..... 37

**B. McDonald** ..... 40

**CONCLUSION** ..... 41

# TABLE OF AUTHORITIES

**Cases**

*A.V.E.L.A., Inc. v. Est. of Marilyn Monroe, LLC*,
364 F.Supp.3d 291 (S.D.N.Y. 2019) ..... 21

*Associated Press v. Meltwater U.S. Holdings, Inc.*,
931 F.Supp.2d 537 (S.D.N.Y. 2013) ..... 28, 39

*Authors Guild v. Google, Inc.*,
804 F.3d 202 (2d Cir. 2015) ..... 7

*Baker v. Saint-Gobain Performance Plastics Corp.*,
No. 1:16-cv-917-LEK/DJS, 2024 WL 4189385 (N.D.N.Y. Sept. 12, 2024)..... 18

*Bazemore v. Friday*,
478 U.S. 385 (1986)..... 40

*Berlyn, Inc. v. Gazette Newspapers, Inc.*,
214 F.Supp.2d 530 (D. Md. 2002)..... 39

*Campbell ex rel. Campbell v. Metro. Prop. & Cas. Ins. Co*.,
239 F.3d 179 (2d Cir. 2001) ..... 32

*Cassava Scis., Inc. v. Bredt*,
No. 1:22-cv-9409-GHW, 2024 WL 1347362 (S.D.N.Y. Mar. 28, 2024)..... 32

*Chen-Oster v. Goldman, Sachs & Co.*,
No. 10-cv-6950(AT)(RWL), 2022 WL 814074 (S.D.N.Y. Mar. 17, 2022) ..... 21

*Comput. Assocs. Int'l, Inc. v. Altai, Inc.*,
982 F.2d 693 (2d Cir. 1992) ..... 11

*Daubert v. Merrell Dow Pharms., Inc*.,
509 U.S. 579 (1993)..... 32, 39

*Deere & Co. v. Kinze Mfg., Inc.*,
No. 4:20-cv-00389-RGE-HCA, 2024 WL 1235569 (S.D. Iowa Feb. 8, 2024)..... 37

*Emmerich Newspapers, Inc. v. Particle Media, Inc.*,
190 F.4th 356 (5th Cir. 2026) ..... 14

*Fed. Hous. Fin. Agency v. Nomura Holding Am., Inc.*,
No. 11-cv-6201(DLC), 2015 WL 539489 (S.D.N.Y. Feb. 10, 2015) ..... 41

*Floyd v. City of New York*,
861 F.Supp.2d 274 (S.D.N.Y. 2012) ..... 40

*Forte v. Liquidnet Holdings, Inc.*,
675 Fed.App'x 21 (2d Cir. 2017) ..... 36

*Goldstein v. Metro. Reg'l Info. Sys., Inc.*,
No. TDC-15-2400, 2016 WL 4257457 (D. Md. 2016) ..... 13, 25

*Gussack Realty Co. v. Xerox Corp.*,
224 F.3d 85 (2d Cir. 2000) ..... 8

*Hachette Book Grp. v. Internet Archive,*
115 F.4th 163 (2d. Cir. 2024) ..... passim

*Hayden v. Int'l Bus. Machs.. Corp.*,
No. 21-cv-2485(VB), 2025 WL 1697021 (S.D.N.Y. June 17, 2025) ..... 36

*Highland Cap. Mgmt., L.P. v. Schneider*,
551 F.Supp.2d 173 (S.D.N.Y. 2008) ..... 37

*Icon-IP Pty Ltd. v. Specialized Bicycle Components, Inc.*,
87 F.Supp.3d 928 (N.D. Cal. 2015) ..... 23

*In re AXA Equitable Life Ins. Co. COI Litig.*,
595 F.Supp.3d 196 (S.D.N.Y. 2022) ..... 2

*In re Mirena IUD Prods. Liab. Litig.*,
169 F.Supp.3d 396 (S.D.N.Y. 2016) ..... 34

*In re Mirena IUS Levonorgestrel-Related Prods. Liab. Litig. (No. II)*,
341 F.Supp.3d 213 (S.D.N.Y. 2018) ..... 6, 7, 40

*In re NYSE Specialists Sec. Litig.*,
260 F.R.D. 55 (S.D.N.Y. 2009) ..... 3, 8

*In re: Gen. Motors LLC*,
No. 14-md-2543, 2015 WL 9480448 (S.D.N.Y. Dec. 29, 2015) ..... 3

*Johnson Elec. N. Am. Inc. v. Mabuchi Motor Am. Corp.*,
103 F.Supp.2d 268 (S.D.N.Y. 2000) ..... 6

*Kumho Tire Co., Ltd. V. Carmichael*,
526 U.S. 137 (1999) ..... 33

*Larball Publ'g Co., Inc. v. Lipa*,
No. 22-cv-1872(KPF), 2025 WL 936536 (S.D.N.Y. Mar. 27, 2025) ..... 11

*Laumann v. National Hockey League*,
117 F.Supp.3d 299 (S.D.N.Y. 2015) ..... 6, 41

*LifeTree Trading Pte., Ltd. v. Washakie Renewable Energy, LLC*,
No. 14-cv-9075(JPO), 2018 WL 2192186 (S.D.N.Y. May 14, 2018) ..... 27

*Louis Vuitton Malletier S.A. v. Sunny Merch. Corp.*,
97 F.Supp.3d 485 (S.D.N.Y. 2015) ..... 1

*Malletier v. Dooney & Bourke, Inc.*,
525 F.Supp.2d 558 (S.D.N.Y. 2007) ..... 32

*Micro Chem., Inc. v. Lextron, Inc.*,
317 F.3d 1387 (Fed. Cir. 2003) ..... 32

*Middle Mkt. Fin. Corp. v. D'Orazio*,
No. 96-cv-8138, 2002 WL 31108260 (S.D.N.Y. Sept. 23, 2002) ..... 37

*Munoz v. Orr*,
200 F.3d 291 (5th Cir. 2000) ..... 36

*Nimely v. City of New York*,
414 F.3d 381 (2d Cir. 2005) ..... 30

*Pierson v. Infinity Music & Ent., Inc.*,
300 F.Supp.3d 390 (D. Conn. 2018) ..... 12

*R.F.M.A.S., Inc. v. So*,
748 F.Supp.2d 244 (S.D.N.Y. 2010) ..... 30

*Rutledge v. Walgreen Co.*,
182 F.4th 168, 193 (2d Cir. 2026) ..... 7, 31, 33

*Ryoo Dental v. Han*,
No. sacv-15-308-JLS, 2016 WL 6138413 (C.D. Cal. April 25, 2016) ..... 30

*Schering Corp. v. Pfizer Inc.*,
189 F.3d 218 (2d Cir. 1999) ..... 39

*Scott v. Chipotle Mexican Grill, Inc.*,
315 F.R.D. 33 (S.D.N.Y. 2016) ..... 18

*U.S. Bank Nat'l Ass'n v. PHL Variable Life Ins. Co.*,
112 F.Supp.3d 122 (S.D.N.Y. 2015) ..... 5, 8

*U.S. v. Locascio*,
6 F.3d 924 (2d Cir. 1993) ..... 39

*United States v. M/Y Amadea*,
770 F.Supp.3d 558 (S.D.N.Y. 2025) ..... 7, 8

*United States v. Malka*,
602 F.Supp.3d 510 (S.D.N.Y. 2022) ..... 10

**Statutes**

17 U.S.C. § 107 ..... 33

17 U.S.C. § 1202 ..... passim

17 U.S.C. § 1203 ..... 14

17 U.S.C. § 401 ..... 25

**Rules**

Fed. R. Civ. P. 26(e) ..... 32

Fed. R. Evid. 702 ..... 1, 5, 27, 29

Fed. R. Evid. 703 ..... 34

## TABLE OF CITATION CONVENTIONS

| Citation | Document | ECF No. |
|---|---|---|
| Mot. | Defendants' Notice of Motion and Rule 702 Motion to Exclude Expert Testimony | 2050 |
| Mem. | Memorandum of Law in Support of Defendants' Rule 702 Motion to Exclude Expert Testimony | 2052 |
| Malhotra | Declaration of Paven Malhotra in Support of Defendants' Joint Rule 702 Motion to Exclude Expert Testimony | 2060 |
| Microsoft-News-MSJ | Defendant Microsoft Corporation's Memorandum of Law In Support Of Motion For Summary Judgment In News Plaintiffs' Consolidated Cases | 1698 (public)<br>1699 (sealed) |
| Microsoft-News-MSJ-SOF | Defendant Microsoft Corporation's Rule 56.1 Statement Of Undisputed Material Facts In Support Of Motion For Summary Judgment In News Plaintiffs' Consolidated Cases | 1827 (public)<br>1702 (sealed) |
| News-MSJ-Opp | News Plaintiffs' Omnibus Memorandum of Law in Opposition to Defendants' Motions for Summary Judgment | 2118 (sealed) |

# ARGUMENT

## I. BOOK PLAINTIFFS

### A. Zhao

Zhao is the Neubauer Professor of Computer Science at the University of Chicago. His work focuses on "the use of content by generative AI models," the subject of his testimony here. ECF1787-1 ¶¶1–2. Defendants' motion to exclude certain of his opinions should be denied.

**1. Zhao does not recite facts without "connecting" them to any conclusion.** Mem.2. [redacted] ECF1787-1 ¶¶12, 14–72; ECF1787-3 ¶¶8–11. He likewise cites evidence that there is no [redacted] ECF1787-2 ¶¶83–100. Zhao thus does not "merely recite[]" evidence or "a narrative of the case." Mem.2 & citations. He uses his expertise to connect the evidence to his conclusions—precisely what Rule 702(b) requires. *See Louis Vuitton Malletier S.A. v. Sunny Merch. Corp.*, 97 F. Supp. 3d 485, 507 (S.D.N.Y. 2015) (rejecting similar argument where expert "synthesize[d] [evidence] and pull[ed] together common themes in reaching his conclusions").

**2. Defendants' assertion that Zhao offers "scienter and legal opinions" is unsupported.** Mem.2. Defendants identify two paragraphs that they claim show that Zhao offers state-of-mind or legal opinions.[1] Mem.2–3. [redacted]

[1] Defendants also cite a Paragraph 110 of Zhao's reply report that does not exist.

[redacted] ECF1787-1 ¶20; ECF1787-2 ¶111. In any event, concerns about "isolated opinions or sentences in Plaintiffs' experts' reports [] can be adequately addressed through particularized objections at trial." *In re AXA Equitable Life Ins. Co. COI Litig.*, 595 F. Supp. 3d 196, 255 (S.D.N.Y. 2022).

**3. Defendants' request to exclude Zhao's "memorization" opinions is unclear and unavailing.** To start, Defendants do not clearly identify *which* opinions they challenge; that is alone reason to deny the motion.

Defendants appear to challenge Zhao's opinions that LLM training "necessarily results in a degree of memorization" and "regurgitation." ECF1787-2 ¶¶14–82; ECF1787-3 ¶¶ 17–34. Those opinions are a corollary of Zhao's broader opinion—set forth in nearly 100 paragraphs of analysis—that LLMs "are approximations of their training data that encode expressive content" via a process equivalent to "lossy compression," an opinion that rests on [redacted], machine learning research, and Zhao's experiment. *See, e.g.,* ECF1787-2 § IV; ECF1787-3 § V. His opinion is even endorsed by Microsoft's technical expert Dr. John Lafferty, who wrote that generative language models "predict[] or, equivalently, compress[] natural text." ECF2143-10 at ix.

Defendants say that Zhao fails to "bridge the gap" by conducting tests of the accused models. Mem.3. But this misunderstands Zhao's opinions and is, at best, material for cross-examination. Zhao opines "that the nature of LLM training" results in memorization and even [redacted]. ECF1787-2 ¶¶14–82; *id.* ¶¶51 (citing Dr. Shan's reports), 57 (same), 78 (same); *see also* ECF1814-1 ¶¶19–21, 140–156;

ECF1814-4 ¶¶6-12 & § IV. Defendants' "disagreement" with Zhao's supported conclusions is no basis to exclude them. *In re NYSE Specialists Sec. Litig.*, 260 F.R.D. 55, 68 (S.D.N.Y. 2009).

**4. Zhao's AI risk opinions are relevant rebuttal.** Defendants reprise their argument that Zhao merely recites evidence. Defendants are wrong again. Zhao cites ample evidence and *connects it to his conclusion*, based on his expertise in AI systems, as he re-affirmed at his deposition. ECF1787-2 § VII; *e.g.*, Zhao Dep. Tr. at 142:23-144:3. Defendants also argue his opinions are irrelevant. [redacted] *See, e.g., In re: Gen. Motors LLC*, 2015 WL 9480448, at *7 (S.D.N.Y. Dec. 29, 2015) (where a party "opens the door by introducing expert testimony . . . he cannot complain when the [other side] offers similar testimony in response" (cleaned up)).

**B. Lasinski**

The Court should deny Defendants' motion to exclude the opinions of Lasinski, a Partner in HKA's Intellectual Property Group with more than three decades of experience valuing intellectual property. ECF1825-1 ¶¶1–5.

First, the premise of Defendants' motion—that an expert assessing factor four must trace declines in a particular plaintiff's sales or revenue to Defendants' products—is contrary to Second Circuit law. Mem.4–6. The Second Circuit has "never held that the rightsholder bears the burden of showing actual market harm"; rightsholders need not even "present empirical data." *Hachette Book Grp. v. Internet Archive,* 115 F.4th 163, 194 (2d. Cir. 2024). Factor four asks about "the effect of the use upon the potential market for or value of the copyrighted work," and reaches "not only the market harm caused by the particular actions of the alleged infringer, but also the market

harm that would result from unrestricted and widespread conduct of the same sort." *Id.* at 189. "[L]ogical inferences" are permitted and it is *defendants'* burden to "disprove" harm. *Id.* at 190, 193.

Lasinski's analysis fits squarely within this framework. He explains that "Defendants' alleged wrongful conduct and infringement of the Copyrighted Works has negatively impacted the Plaintiffs in the form of lost revenues from book sales and/or foregone licensing opportunities." ECF1825-1 ¶14. "The harm to authors operates through at least two concurrent market mechanisms." ECF1825-2 ¶411. "The first is volume-driven: the entry of a large quantity of new, competing content directly into the marketplace where Plaintiffs' works are sold." *Id.* "The second is quality-driven: to the extent that the introduction of AI-generated content meets or exceeds consumer acceptance levels, those works can more directly compete with Plaintiffs' works, thereby introducing a level of competition that did not exist before." *Id.*

Those opinions rest on data showing a flood of AI-generated works into the marketplace, including works generated using Defendants' models. [redacted]

[redacted] This record amply supports Lasinski's opinions.

Defendants' challenges to this data raise merits disputes, not Rule 702 issues. [redacted] Defendants' disagreements with Lasinski's data go to weight, not admissibility. *See U.S. Bank Nat'l Ass'n v. PHL Variable Life Ins. Co.*, 112 F. Supp. 3d 122, 134 (S.D.N.Y. 2015) ("weakness in the factual basis of an opinion bears on the weight of the evidence, not its admissibility").

Defendants' contention that marketplace dilution "is not a cognizable form of harm" is addressed in the summary judgment briefing. Mem.4; *see* ECF1848 26–28, 33–38. An influx of competing substitutes is the paradigm of "market harm that would result from unrestricted and widespread conduct of the same sort." *Hachette*, 115 F.4th at 189.

Second, Defendants' assertion that Lasinski "failed to assess the sales data Defendants' experts obtained from third-party provider Circana" is wrong and, even if true, would not justify exclusion. Mem.5. Lasinski analyzed Leonard's Circana-based regressions and showed they undercut Leonard's conclusions. ECF1825-2 ¶434. He documented substantial discrepancies between Plaintiff-produced sales data and the Circana data for a sample of the Asserted Works. *Id.* ¶¶422–424. And he showed that Circana captures only print trade books, excluding e-books and audiobooks. *Id.* ¶¶415, 425. [redacted]

[redacted]

ECF1825-2 ¶415, 425; ECF2265-2 ¶141.

Defendants' reliance on Lasinski's deposition fares no better. Asked whether he had conducted "any quantitative analysis" showing that book prices "have decreased because of ChatGPT and Copilot," Lasinski answered that he had not performed that price analysis, then explained what his analysis does show: "since the flood of … AI-generated books, the revenue per book, even without any AI generation in it, goes down." Malhotra Ex. 9, 97:16–98:02. Defendants recast that as a concession that he is "not making a 'causal claim'" about book revenues. Mem.5. But declining to run a particular price regression is not a disclaimer of causation, and the answer Defendants omit describes the very relationship they say he disclaimed.

Third, Defendants' authorities are irrelevant. *Laumann v. National Hockey League*, 117 F. Supp. 3d 299 (S.D.N.Y. 2015) is an antitrust case addressing expert testimony about class-wide antitrust injury. The fourth fair use factor is an entirely different inquiry, as it addresses not only *past* market impact but consideration of the impact of widespread conduct of the same type and does not require any empirical evidence whatsoever. *Hachette,* 115 F.4th at 190, 193. *Johnson Elec. N. Am. Inc. v. Mabuchi Motor Am. Corp.*, 103 F. Supp. 2d 268 (S.D.N.Y. 2000), is equally far afield. It too is not a copyright case and involved an expert who failed to consider invoices in assessing patent damages. *Id.* at 283–84. In all events, Lasinski analyzed sales and revenue data—[redacted], Circana's, and Plaintiffs' own—and showed how it demonstrates an effect on the potential market for or value of the copyrighted works at issue.

Likewise, *In re Mirena IUS Levonorgestrel-Related Prods. Liab. Litig. (No. II)*, 341 F. Supp. 3d 213 (S.D.N.Y. 2018) is about causation in a personal injury case. Defendants cite *Mirena* for the unremarkable proposition that experts generally should not rely on outside studies in ways

that "exceed the limitations the authors themselves place[d] on" them. *Id.* at 241. Unlike in *Mirena,* Lasinski does not cite a single study to establish "causation"—the fourth factor is not limited to evaluation of market effects caused by the defendant—but instead relies on a 2026 Reimers study along with a host of other evidence to arrive at his conclusions related to the effect on the market and value of Plaintiffs' works. ECF1825-2 ¶377; ECF2034-30 29; ECF1825-2 ¶¶385-412, 417. Regardless, challenges to an expert's "reliance on secondary sources concern the weight, not the admissibility, of [the] testimony." *United States v. M/Y Amadea*, 770 F. Supp. 3d 558, 580 (S.D.N.Y. 2025). What is more, the Second Circuit recently cautioned against over-policing experts' reliance on relevant scientific studies, particularly where an expert "synthesizes many scientific results" to arrive at his conclusions. *Rutledge v. Walgreen Co*., 182 F.4th 168, 193 (2d Cir. 2026).

Finally, *Authors Guild v. Google, Inc.*, 804 F.3d 202 (2d Cir. 2015), does not help Defendants. Mem.6. That case held that search-index snippets do not give users "access to effectively competing substitutes." *Id.* at 223–24. [redacted]

### C. Seeley

Mark Seeley has spent more than 36 years in publishing, including 22 years as General Counsel of Elsevier coordinating anti-piracy enforcement against Library Genesis. ECF1804-1 ¶¶5, 8, 23, 32; Appx. A. Defendants devote three paragraphs, citing no authority, to excluding virtually all his testimony. Mem.8.

Defendants first argue that Seeley's testimony on how digital piracy harms the market for copyrighted books is irrelevant. Mem.8. But, as explained above, fair use requires consideration of the effect on the potential market.

Defendants next attack Seeley's sources. But as noted, such challenges go to weight, not admissibility. *Amadea*, 770 F. Supp. 3d at 580; *U.S. Bank*, 112 F. Supp. 3d at 134.

The attacks are meritless regardless. Defendants complain that certain studies concern video piracy. Mem.8. But Seeley explained that the reliance was reasonable: piracy depresses traffic to the sites where consumers lawfully buy content regardless of medium. ECF1804-3 ¶¶23–24; *see also* Malhotra Ex. 14 at 99:18-24. Defendants do not explain why that parallel is unreasonable. And his opinion rests independently on multiple other sources. *See* ECF1804-1 ¶¶13-15, 18, 55, 57-60, 66.

Defendants also fault Seeley for not reconstructing the methodologies underlying the studies he cites. Mem.8. But an expert may rely on published studies in his field without having designed, conducted, or replicated them himself. *See Gussack Realty Co. v. Xerox Corp.*, 224 F.3d 85, 95 (2d Cir. 2000) ("The expert need not have conducted her own tests."). Any such gap is fodder for cross-examination, not exclusion.

[redacted]

Seeley reads them differently and answered each critique. ECF1804-3 ¶¶18–28, 48–53. Such disagreement "is not a legitimate basis on which to strike that expert's opinion." *NYSE*, 260 F.R.D. at 68.

Defendants call Seeley's "existential threat" opinion irrelevant and his normalization opinion "conjecture." Mem.8. Neither label holds. That threat is precisely the factor-four effect

from "widespread conduct of the same sort." *Hachette*, 115 F.4th at 189. And the normalization opinion rests on the record. Seeley explains that Defendants are among the industry's most visible developers and among the first to take books at scale from shadow libraries, signaling to others that the practice was acceptable and the risk tolerable. ECF1804-3 ¶¶36–44. He cites OpenAI researchers who pirated LibGen at OpenAI joining Anthropic and engaging in the same piracy there, Meta's leadership approving LibGen because it was "known that OpenAI and Mistral are using the library for their models," statements from Anna's Archive that shadow libraries "were dying" until the arrival of AI and that "[v]irtually all major companies building LLMs contacted us," and more. *Id.*; ECF1804-1 ¶¶ 64-66. At worst, Defendants' objections are the type of issue the Court can police at trial—not a basis for wholesale exclusion of Seeley's testimony.

Finally, Defendants complain that Seeley relies on piracy web-traffic reports without showing Defendants are the source of such traffic, but cite no authority requiring him to apportion that growth between Defendants and their imitators. Mem.8. Courts in this Circuit draw logical inferences on factor four and do not condition admissibility on quantification. *Hachette*, 115 F.4th at 189, 193–95.

* * *

Defendants have not justified excluding any opinions of Books Plaintiffs' experts, and the motion should be denied in full.

Regardless, the Court should not exclude the following opinions listed in Defendants' notice but not discussed in their memorandum:

- Zhao: Opening Sections II (¶¶7-8), III (¶12(b)), and V (¶¶73-81); Rebuttal Section VI (¶¶103-110); and Reply Section V.F. (¶36).

- Lasinski: Opening Sections 2.0 (¶14), 6.3 (¶¶208, 211, 214), and 6.4 (¶215); Rebuttal Sections 2.1 (¶10), 2.2 (¶12), 2.5 (¶16), 9.7.1 (¶¶414-419), and 9.7.5 (¶442); and Reply Section 2.4 (¶12).

- Seeley: Opening Sections I (¶4), IV (¶¶13–15), VI (¶¶36–52), VII (¶¶53–62), and VIII (¶¶63–75); Rebuttal Sections I (¶¶3–4), IV(a) (¶¶14–21), and IV(c) (¶¶41–46); and Reply Sections I (¶¶3–7), IV (¶¶13–28, 30–47), and V (¶¶48–70).

Defendants never explain, much less justify, excluding those opinions. *See United States v. Malka*, 602 F. Supp. 3d 510, 529 (S.D.N.Y. 2022).

## II. NEWS PLAINTIFFS' TECHNICAL EXPERTS

### A. Goldstein

News Plaintiffs present three highly qualified experts (Goldstein, Wenger, and Delorey) who independently conducted reliable (i) text matching analyses (comparing Asserted Works to training datasets and outputs) and (ii) copyright management information ("CMI") removal analyses on a massive scale.

1. Text matching.

Defendants do not challenge the technical aspects of Goldstein's text matching analysis. Instead, they fault Goldstein for not conducting legal analysis regarding copyrightability—*i.e.*, "he did not audit authorship, verify copyright registrations, or otherwise attempt to determine whether the material supplied to him matched text that came from unprotected sources;" and did not "filter out non-protectable elements[.]" Mem.9-10.

Defendants identify just four works out of millions that they contend Goldstein should have filtered out due to purported lack of ownership. Mem.9-10. News Plaintiffs are no longer asserting those works, and Defendants' arguments with respect to them are moot.

Moreover, Goldstein's role is technical, not legal, and he appropriately relied on counsel to identify the works at issue. Ex. 66 [Goldstein-Reply-Rpt.], ¶183. Goldstein's opinion is fully consistent with "[t]he well-established general rule in this circuit," that "expert testimony may be used to assist the fact finder in ascertaining whether the defendant had copied any part of the

plaintiff's work[.]" *Comput. Assocs. Int'l, Inc. v. Altai, Inc.*, 982 F.2d 693, 713 (2d Cir. 1992); *see also Larball Publ'g Co., Inc. v. Lipa*, No. 22-cv-1872(KPF), 2025 WL 936536, at *6-7 (S.D.N.Y. Mar. 27, 2025) (rejecting argument that expert must "conduct a filtering analysis for his or her report to be admissible," and explaining that "Defendants conflate the legal standard for definitively establishing that substantial similarity exists with the requirements for a reliable expert report that could assist the trier of fact in concluding that such a similarity exists").

2. CMI removal.

Defendants challenge the accuracy of the CMI Goldstein extracted and the sufficiency of his searches for CMI in OpenAI's training data. Neither challenge has merit.

*a. Goldstein validated his CMI removal analysis.*

Goldstein extracted CMI from the Asserted Works contained in the Common Crawl datasets using, *inter alia*, the tool Newspaper3K that OpenAI also used when it "processed" those same datasets. Newspaper3K is a program designed to extract newspaper article text and bylines (e.g., title and author) from the surrounding content on a webpage. Ex. 67 [Goldstein-Opening-Rpt.], ¶¶524-527. Defendants do not object to use of this tool, argue it is unreliable, or otherwise attack Goldstein's methodology. Rather, they assert Goldstein "failed to validate" whether his extracted CMI was "accurate" because of alleged mistakes in Goldstein's identification of CMI. Mem.11. Defendants' attacks are misguided.

*First*, Goldstein validated his results. *See* Ex. 67, ¶550 (p.220)[2]; Ex. 66, ¶115 (validation included "looking at thousands of examples of WARC files [the raw HTML making up News Plaintiffs' webpages] that had CMI extracted, and confirming that the script used had correctly

[2] Note that certain paragraph numbers in the Goldstein Opening Report are duplicated; page numbers are provided along with certain paragraph numbers for clarity.

extracted the CMI"). Any dispute regarding the sufficiency of Goldstein's validation goes to weight, not admissibility.

*Second*, none of the alleged mistakes support exclusion.

Defendants complain that the article title Goldstein identified sometimes "included publication names, column names, and other descriptions that were not the actual 'title' of the work[.]" Mem.11. But under the statute, CMI includes "[t]he title and other information identifying the work[.]" 17 U.S.C. § 1202(c)(1).

Defendants complain that the publisher Goldstein identified was sometimes "populated by a metadata tag for the website's name[.]" Mem.11. But the website's name—one of News Plaintiffs' domains—accurately reflects "[t]he name of, and other identifying information about, the copyright owner of the work[.]" 17 U.S.C. § 1202(c)(3).

Defendants complain that the copyright notice and terms of service that Goldstein identified "appeared in sitewide footers" "applied indiscriminately to every page on the website." Mem.11. This is a legal question, not a basis for exclusion. Moreover, it is common to place copyright notices and terms of service links in webpage footers; it is still CMI. *E.g.*, *Pierson v. Infinity Music & Ent., Inc.*, 300 F.Supp.3d 390, 396 (D. Conn. 2018) (terms of use "at the bottom of each relevant webpage" qualified as CMI).

Defendants complain about News Plaintiffs' webpages "that do not contain Asserted Works or even potentially copyrightable works" but have a copyright notice (Mem.12), but Goldstein did not count those webpages and Defendants do not explain how copyright notices on works *not in this case* have any bearing on whether a copyright notice on Asserted Works constitutes CMI. For example, Defendants refer to "a NYT webpage that presents ***only*** the full

text of the U.S. Constitution[.]" Mem.12 (emphasis added). But the underlying webpage[3] includes ***more than*** the full text of the U.S. Constitution; it also includes current headlines and curated links with thumbnail images (under "Editor's Picks"). The other examples similarly fail.

Defendants' two additional complaints about the copyright notices Goldstein identified fall flat. *First*, they claim Goldstein counted instances of "New York Times" "without a copyright claim or year, in ***over 2 million*** instances[.]" Mem.12. But OpenAI's own expert admitted this is not true because he looked at the wrong field for his contrary opinion. Ex. 93 [Krishnan-7/29/2026-Depo.-Tr.], 175:12-176:8; *see also* Ex. 66, ¶¶124-26. *Second*, Defendants argue that in some cases "the extracted year [of the copyright notice] did not match the copyright registration year of the Asserted Work[.]"[4] Mem.12. This is a legal issue, not a basis for exclusion. Moreover, the statutory language regarding copyright notices is not that the notice itself is CMI, but that "the information set forth in a notice of copyright" is CMI. *See* 17 U.S.C. §1202(c)(1), (3). This information is present, and qualifies as CMI, whether or not the year in the notice is the year in the registration. *Cf. Goldstein v. Metro. Reg'l Info. Sys., Inc.*, No. TDC-15-2400, 2016 WL 4257457, at *7 (D. Md. 2016) (rejecting argument that CMI must "include all of the elements of a 'notice of copyright,'" and finding that "Metro Photograph, © www.goldsteinphoto.com"—without a year—constituted CMI); *see also* 17 U.S.C. §1202(c)(1), (3) (defining CMI).

If Defendants wish to challenge particular instances of CMI removal, they can do so at trial, but such challenges go to the weight of Goldstein's testimony (or legal questions), not its admissibility.

---

[3] *See* https://www.nytimes.com/1987/07/04/us/we-the-people-the-constitution-of-the-united-states-of-america.html (from Malhotra Ex. 35 ¶213).

[4] For example, some webpages hosting Asserted Works display the current year next to the copyright notice, regardless of when an Asserted Work was first published.

*b. Goldstein's search for CMI was not "deficient"*

OpenAI does not dispute that it removed CMI from News Plaintiffs' articles. Instead, OpenAI argues that because an article's URL may include some CMI[5] that OpenAI admittedly removed from the body of the article (like title, date, or publisher information), OpenAI did not remove this CMI. Accordingly, Defendants fault Goldstein for purportedly "ignor[ing] relevant fields in" OpenAI's training data records, because "he only searched the training record's 'text' field, while excluding the preserved 'url' and 'domain' fields." Mem.12. OpenAI is wrong for three independent reasons.

*First*, the statute prohibits intentional removal of "***any*** copyright management information" (17 U.S.C. §1202(b)(1)) and provides damages "for ***each*** violation" (17 U.S.C. § 1203(c)(3)(B)). OpenAI's retention of the URL does not negate that OpenAI removed other occurrences of the CMI.[6] Goldstein reliably identified the CMI that OpenAI removed, which is a "violation."

*Second*, OpenAI created additional copies of the Asserted Works during training that removed the "url" and "domain" fields it alleges to be CMI. *See* Ex. 66, ¶¶127-129. Even if it initially preserved the URLs, OpenAI eventually removed them during training.

*Third*, OpenAI has not shown that the URLs here constitute CMI conveyed with the work. *See Emmerich Newspapers, Inc. v. Particle Media, Inc.*, 190 F.4th 356, 381 (5th Cir. 2026). It is especially dubious here because Krishnan's opinion included cases of "inexact" matches—a term

[5] OpenAI's expert's example is "http://www.nytimes.com/**1992/05/07**/sports/**cycling-lemond-bids-for-elusive-title**.html?src=pm", with the date and title portions bolded. Malhotra Ex. 35 [Krishnan-Rebuttal-Rpt.], Figure 23. OpenAI contends the publisher information is found in the domain (here, nytimes.com).

[6] While OpenAI argues that publication dates (but not other CMI) were extracted from the URL, Mem.13, it ignores Goldstein's analysis showing that publication dates were present in the Asserted Works outside of the URL. *See* Ex. 66 [Goldstein-Reply-Rpt.], ¶¶130-131.

he could not define—between "CMI" in the URL and CMI on Plaintiffs' webpages. Ex. 93, 235:25-237:25.

3. Microsoft's Copilot logs.

The premise of every Microsoft critique is that Goldstein's analyses bear exclusively on quantifying the ***number*** of infringements, but Goldstein's analysis of Microsoft's Copilot logs is relevant both to showing *prima facie* copyright infringement and rebutting Defendants' fair use defense. His analysis quantifies how frequently Copilot uses News Plaintiffs' content, explains how Copilot works, and quantifies how frequently users turn to Copilot for news. *See* Ex. 67, ¶¶394-398, 401-411, 415-420, Tables 9-11; Ex. 64 [Goldstein-Rebuttal-Rpt.], ¶¶77-79, 82-85, 106-107, 109-120, 129-158; Ex. 68 [Goldstein-Supplemental-Rpt.], ¶¶21-41, 46-49. Microsoft's critique goes to weight, not admissibility, and should be denied.

*a. 4-gram counts.*

As part of Goldstein's analysis, he presents 4-gram counts (spans of 4 consecutive words) from News Plaintiffs' Asserted Works that appear in Microsoft's Copilot logs. Goldstein found significant numbers of Copilot conversations having at least 50 consecutive four-word sequences from Asserted Works. *See* Ex. 67, ¶398. Microsoft alleges these counts are "inflate[d.]" Mem.13. But Microsoft has not argued that Goldstein computed the 4-gram counts incorrectly. Rather, Microsoft claims the 4-gram counts do not show infringement because they include user's prompts and aggregate counts from multiple Asserted Works. Mem.14-15. But occurrences of News Plaintiffs' content in the prompt (or in the grounding and output fields) of Copilot conversations are relevant for several purposes, including to rebut Microsoft's fair use argument that consumers rarely use Copilot to obtain information about current events. *See* Microsoft-News-MSJ.28. Similarly, the amount of content—even spread across multiple articles—is relevant to how Copilot

copies News Plaintiffs' content. Microsoft does not explain how copying from two asserted articles is less culpable than copying from one.

Microsoft also ignores the context in which these counts are offered. They are not provided by themselves to show infringement but are part of a larger set of analyses which collectively quantify Copilot's use of News Plaintiffs' content.

*b. RAG leak analysis.*

In rebuttal to Defendants' fair use defense, Goldstein presents evidence that text spans of 16-words or longer provided to the model in the grounding process are systematically output ("leaked") in responses to end users. *See* Ex. 64, ¶75 (disputing opinion that Defendants used no more original content than necessary). Microsoft does not dispute the results of Goldstein's leak analysis but argues instead the analysis is "doom[ed]" because the Copilot conversations Goldstein identifies "may have no displayed match whatsoever to Plaintiffs' Asserted Works." Mem.15. That critique misses the point.

Copilot's propensity to display its grounding data in outputs is relevant to how Copilot works and is proper for rebutting Defendants' fair use defense, including whether the RAG process copies more content than necessary or simply repackages the expression it copies for the same purpose as the original. Microsoft has argued that web content used for grounding "is not revealed to the user." Microsoft-News-MSJ.17. Goldstein's analysis shows this is untrue—Copilot systematically displays grounding content in outputs to the user in approximately 20% of conversations. Ex. 64, ¶82.

*c. Web-domain and "impressions" counts.*

Goldstein provides counts for how often News Plaintiffs' domains appear in Copilot output logs, as well as Microsoft's own "impressions" data "track[ing] the number of times that web domains appeared as citations." *See* Ex. 67, ¶¶394-97, 415-18. He also considered Microsoft

testimony on how long snippets can be and provided per-plaintiff information on snippet length. Ex. 64, ¶¶102-07 (citing Microsoft's 30(b)(6) witness testimony that "[redacted]"); Ex. 67, ¶¶201-04 (analyzing snippet length based on documents and testimony), 400-414 (analyzing snippet lengths from exemplary conversations), ¶416 and n.382. Goldstein concludes "that Copilot used News Plaintiffs' content for grounding a significant number of times" and "Microsoft likely copied substantial portions of News Plaintiffs' articles and website content as part of the RAG process and in generating output to users." Ex. 67, ¶¶395, 418.

Microsoft argues that "the inference from citation count to volume copied" is not measured, and these counts do not help a trier of fact determine what portions were copied. Mem.16-17. But Microsoft ignores the technical analysis described above that Goldstein provided to support his conclusion. In addition, the domain and impression counts for News Plaintiffs' websites that Copilot accessed rebut Microsoft's argument that consumers rarely use Copilot to obtain information about current events. Microsoft-News-MSJ.28.

*d. Supplemental Report analysis.*

In his supplemental report (which addresses Copilot output logs that Microsoft produced only at the end of fact discovery and then supplemented four weeks after the close of fact discovery), Goldstein quantifies how much of News Plaintiffs' Asserted Works appeared in Copilot output logs. Goldstein only included matching text if it met a particular threshold (having at least 16 consecutive, or nearly consecutive, words from an Asserted Work) and further required a News Plaintiff domain. *See* Ex. 68, ¶¶21-35. This evidence is sufficient to show how much of News Plaintiffs' Asserted Works appear in the Copilot logs. *See* Ex. 69 [Goldstein-7/31/2026-Depo-Tr.], 517:19-518:1. Goldstein then applied an additional filter: Google's Gemini LLM "to

check whether the span [of text from an Asserted Work] represents prose by the article author, or whether it is quoted/repeated from an outside source." *See* Ex. 68, ¶28. Microsoft contends that the entire analysis should be excluded because the Gemini filtering process "has an unknown, untested error rate." Mem.18. But Microsoft claims that LLMs are "capable of a wide variety of natural language processing tasks, such as summarization, translation, reasoning, content generation, and question answering." Microsoft-News-MSJ-SOF-SF3. Here, Goldstein used Google's Gemini for a natural language processing task that he assessed it was capable to perform. Goldstein's results are limited to conversations containing both a certain threshold of News Plaintiffs' Asserted Works and a News Plaintiff domain. No more was needed to make the analysis reliable. *See* Ex. 69, 517:19-518:1. Microsoft has not shown, or even alleged, any error in the Gemini filtering process. Goldstein's additional filtering with Gemini is a sound methodology and renders the results more conservative by removing conversations that would otherwise be counted.

4. Defendants' technical collaboration.

Microsoft contends that Goldstein's opinions on Defendants' collaboration in training the models at issue in this case "should be excluded as an improper narrative of the record with legal conclusions attached." Mem.19. Not true. Goldstein's testimony explains each Defendant's technical role and contributions towards creating the millions of copies of News Plaintiffs' works documented in his reports, based on Defendants' highly technical documents and deposition testimony. "[A]n expert's testimony is admissible when it is required to aid the jury in its understanding of highly technical documents." *See Baker v. Saint-Gobain Performance Plastics Corp.*, No. 1:16-cv-917-LEK/DJS, 2024 WL 4189385, at *2 (N.D.N.Y. Sept. 12, 2024) (citing *Scott v. Chipotle Mexican Grill, Inc.*, 315 F.R.D. 33, 45 (S.D.N.Y. 2016)).

Goldstein opined on Microsoft's ***technical ability*** to "supervise and control training" based on Microsoft's "(1) developing, operating, and providing technical support for the supercomputing

platform used to train the models; (2) providing Bing Index data for OpenAI to use to develop and evaluate its LLM models; and (3) providing code, services, and URLs for a joint project to scrape Plaintiffs' Asserted Works under the 'GPTBot' alias." *See* Ex. 67, ¶547 (p.209). He analyzed extensive technical documents and testimony about the infrastructure, data, code, services, and URLs that Microsoft provided to OpenAI to support this opinion. *See generally* Ex. 67, §XIV.

For avoidance of doubt, Goldstein will not testify about Microsoft's legal duties.

**B. Wenger**

1. Text matching.

Like Goldstein, Wenger performed a text-matching analysis to identify instances of Plaintiffs' works in OpenAI's training sets. Defendants seek to exclude it because she did not first "validate" the Asserted Works for ownership or copyrightability, and identified training-data matches containing alleged third-party text. Mem.19. But Defendants point to only *18 examples* from hundreds of thousands of copies. *Id*. Like Goldstein, Wenger is not a lawyer and relied on Plaintiffs to identify their works. Defendants may question Wenger on these legal challenges on cross-examination.

Defendants next object that some of the 18 examples are "false positive" matches without any text. Mem.19. Defendants ignore the two-step nature of her analysis: first searching for URLs matching Asserted Works, then reporting the "Jaccard" overlap between the Asserted Work and the training-data record. Ex. 106 [Wenger-Opening-Rpt.], pp.37-48. When a URL had no adjacent text, Wenger reported zero overlap. Malhotra Ex. 36 [Krishnan-Rebuttal-Rpt.], ¶¶112-13; Ex. 106, p.42. Wenger did not opine that these matches are substantially similar "copies." She reported the number of matches within each similarity decile plus the overlap for each training-data record. This reliable methodology will assist the jury to determine the number of copies at its chosen threshold and exclude any records it deems dissimilar.

Defendants also object to Wenger's "text-matching 'regurgitation' analyses." Mem.19. They conflate two analyses. The first, to which Defendants raise another "third-party-false-positive" objection, concerns Wenger's base model inspection. *Id.* (citing Malhotra Ex. 58 [Wenger-Depo.-Tr.], 181:24-184:24). Here, however, Wenger did not purport to identify infringing "regurgitations." Instead, Wenger prompted certain base models to ascertain the models' ***capacity*** to memorize and regurgitate. Ex. 106, pp.71-79; Ex. 71 [Wenger-Reply-Rpt.], pp.99-100. Including some purported third-party content would not undermine that conclusion.

Second, Defendants object that, when inspecting output logs, Wenger failed to distinguish user input from model-generated text. Mem.19. This is incorrect. In two outputs, the prompt included the work's text. Malhotra Ex. 58, 189:24-190:6. Defendants offer no argument that the resulting output is not an infringement, or any reason why Wenger may not opine on the others.

2. Keyword matching and summarization.

Defendants object to Wenger's findings that Defendants' products summarize Plaintiffs' content and include Plaintiff-relevant terms in chats because these outputs do not measure regurgitation. Mem.20. Wenger's findings, however, pertain to other aspects of the case. ChatGPT summarizing Plaintiffs' articles and discussing its publications are probative of substitution.

3. CMI-related opinions.

*a. CMI acquisition.*

Wenger opines that OpenAI obtained Plaintiffs' CMI when it downloaded copies of Asserted Works from (1) Common Crawl, and (2) Plaintiffs' websites (the latter leading to copies in the WebText training sets). Wenger reached this opinion on two independent grounds.

First, she examined source code: (1) the crawler on which Common Crawl's was based, and (2) the WebText2 scraper. From this, Wenger concluded that the scrapers took Plaintiffs'

entire webpages, including all CMI. Ex. 106, pp.30-32, 36. This analysis supports Wenger's opinion that OpenAI obtained Plaintiffs' CMI when it downloaded the Asserted Works. Because Defendants ignore this analysis entirely, there is no basis to exclude the opinion.

Second, to confirm this analysis, Wenger inspected a sample of 17,631 Common Crawl files associated with Plaintiffs' Asserted Works. *Id*. at 32-35; Ex. 71, pp.22-26. Using an automated search methodology uncontested by Defendants, followed by spot-checking, Wenger concluded that whenever an asserted work appeared in the Common Crawl files, so did its CMI. Ex. 106, pp.32-35; Ex. 71, pp.22-26. Krishnan checked Wenger's analysis but found no errors. Ex. 93, 195:3-17.

Defendants claim Wenger did not spot-check enough samples. Mem.21. But she chose her sample size based on what an academic peer reviewer would deem reasonable. Malhotra Ex. 58, 139:2-140:3. Defendants can address sample size on cross. *Chen-Oster v. Goldman, Sachs & Co.*, No. 10-cv-6950(AT)(RWL), 2022 WL 814074, at *17 (S.D.N.Y. Mar. 17, 2022); *A.V.E.L.A., Inc. v. Est. of Marilyn Monroe, LLC*, 364 F.Supp.3d 291, 326-27 (S.D.N.Y. 2019).

Defendants also contend that CMI may not be present in Common Crawl for various reasons, Mem.21, but Wenger accounted for them. As she acknowledged, her automated search sometimes found no CMI, and she manually inspected each instance. Ex. 106, p.34. Without fail, the asserted work itself was absent due to website errors. *Id*. at 34-35. These instances do not undermine her opinion that whenever OpenAI *did* obtain the asserted work, it also obtained the CMI. *Id*.[7] Defendants offer no basis to exclude her opinions.

[7] Defendants also cite Wenger's testimony that CMI could change if the underlying HTML changed, Mem.21-22, but cite no evidence of any changes that would undermine her opinion.

Defendants also criticize Wenger for not examining raw WebText downloads, Mem.21, but ignore that OpenAI did not produce them notwithstanding Plaintiffs' discovery requests. Ex. 71, p.23. Regardless, Wenger's analysis of the WebText2 scraper supports her opinion.

*b. CMI removal.*

Defendants accuse Wenger of using "mismatched tests for CMI presence and removal." Mem.21. Specifically, they say Wenger compared the training-data records with certain CMI supplied by Plaintiffs' counsel, not the CMI in Common Crawl, and claim Wenger conceded the two sources "could differ." *Id*. But Wenger testified she "had no reason to believe they were different." Malhotra Ex. 58, 146:25-147: 8. Defendants identify no differences.

Defendants also contend Wenger inconsistently searched training-data records for "all rights reserved" and the Common Crawl data for "copyright" or "©." Mem.21. But Plaintiffs' copyright notices have both. Ex. 106, pp.33, 53. Defendants show no effects on her results.

*c. "False positives."*

Defendants also claim Wenger's CMI methodology led to false positives. Mem.21-22. They cite one purported instance out of hundreds of thousands of CMI removals. *Id*.; Malhotra Ex. 36, ¶171 (Fig. 28). But as Krishnan Figure 28's title makes clear, that supposed false positive is a downstream effect of Wenger's text-matching methodology, not her CMI-removal analysis.

*d. "CMI" retained in URLs.*

Defendants object that Wenger ignored URLs that supposedly are CMI. Mem.22. Plaintiffs incorporate their response to the materially identical objection to Goldstein. *See supra* at § II.A.2.b.

4. Copilot extrapolations.

In calculating Copilot regurgitations, Wenger first quantified matches between the Copilot conversation logs and the asserted works, then extrapolated them on an annualized basis. Ex. 106, pp.105-110. Defendants accuse Wenger of employing "flawed" match criteria, but their motion identifies no flaws, so any challenge is forfeited. Mem.22.

Defendants object to Wenger's extrapolation as based on "a single day's Copilot user rate," Mem.22, but offer no evidence that rate is wrong. Defendants also say the extrapolation comes from a "biased census" because the sample included CIR-related keywords. *Id.*[8] Wenger opined her extrapolations were sound anyway, including because her CIR rates mirrored her uncontested Intercept rates, and she would not expect many regurgitations to contain CIR's keywords, which were removed from OpenAI's CMI-less training data underlying Copilot. Ex. 71, pp.116-17.

**C. Delorey**

1. Delorey's "Copies" Opinion is Reliable.

Delorey applies decades of data-science expertise to apply well-established textual comparisons of ZD's Asserted Works with OpenAI's training data records. OpenAI does not criticize Delorey's credentials or the soundness of his science. Instead, as with Goldstein and Wenger, OpenAI's criticisms of Delorey attack extraneous issues that at best, go to weight, not admissibility. *Icon-IP Pty Ltd. v. Specialized Bicycle Components, Inc.*, 87 F.Supp.3d 928, 941 (N.D. Cal. 2015) (imperfect models must "be attacked by cross examination, contrary evidence, and attention to the burden of proof, not exclusion").

[8] Defendants do not criticize Wenger's Intercept extrapolation; nor could they, since Intercept-related keywords were not used to select the sample. Ex. 71, p.115.

OpenAI argues that Delorey's text matching analysis lacked validation. Mem.23. But, in reality it was OpenAI itself that identified "[redacted]" to ZD Asserted Works. Ex. 103 [Delorey-Depo.-Tr.], 102:3-104:25; Ex. 93, 337:19-338:18. Most of Delorey's matches were based on those results, already identified by OpenAI as matches.

OpenAI then argues that Delorey's filtration process failed to exclude what OpenAI characterizes as "third-party" content from the Asserted Works. But Delorey does not have expertise in legal issues, like ownership. Rather, he successfully employed filters to exclude certain content identified by counsel, and OpenAI's identification of a few "edge" examples—compared to ~1.7M Asserted Works—does not undermine Delorey's filtration methodology. Ex. 95 [Delorey-Supplemental-Rpt.] §5-8 (removing works OpenAI identified as purportedly third party and providing data to evaluate each work individually). Even OpenAI's expert admitted that filtering a corpus this large was a "[redacted]." Ex. 94 [Krishnan-7/30/2026-Depo.-Tr.], 440:20-441:12.

OpenAI's next criticism is even weaker. Out of Delorey's nearly six million "confirmed matches", OpenAI's expert identified "[redacted]" and was "[redacted]" Ex. 93, 346:13-355:18; Ex. 95, ¶¶35-37. He also conceded that he "[redacted]"[9]. Ex. 94, 743:21-745:19. For most of the ZD brands, Krishnan failed to identify any false positives, further limiting his critique. Ex. 93, 368:4-15, 370:6-371:1.

[9] Exs. 26, 27, 38, 39 of the Malhotra Decl. are erroneous examples of "false positives," as Delorey has already excluded these. Mem.23; Ex. 95, ¶¶7-8 (excluding .gov URLs).

2. Delorey's CMI Analysis is Reliable.

Delorey also applies well-accepted methodologies to identify instances where OpenAI removed CMI.

Again, OpenAI does not dispute Delorey's credentials or that he properly identified OpenAI's removal of ZD's CMI from its training records. Ex. 96 [Delorey-Opening-Rpt.] ¶¶259, 26. Defendants' criticism is narrower: some removed CMI supposedly did not meet the legal definition of a "copyright notice" under 17 U.S.C. § 401(b). Mem.24-25. But Delorey does not purport to offer any opinion on the legal sufficiency of the copyright notice; he simply presents evidence of CMI removal. Further, OpenAI's legal arguments are misplaced: the "plain language of the statute [§1202(c)(3)] does not equate CMI with a notice of copyright." *Supra* § II.A.2.a; *Goldstein*, 2016 WL 4257457, at *7. Indeed, as OpenAI's expert admitted, "[redacted]" can make it "[redacted]" Ex. 93, 26:17-27. OpenAI also cites no authority that § 401(b) limits the type of CMI giving rise to a removal claim, which is governed by § 1202. The legal sufficiency of copyright notices is not a basis to exclude Delorey's technical opinions.

OpenAI's suggestion that Delorey's supplement "conced[ed]" he "failed to reliably identify CMI" is wrong. Mem.24. Delorey properly supplemented his report in response to untimely reports served by Krishnan. Ex. 95, ¶18 (emphasizing his disagreement with Krishnan while mooting Krishnan's ever-changing criticisms). Importantly, Delorey's ultimate conclusion—that OpenAI stripped ZD's copyright notices from its training records in 99.81% of the instances—did not change and is not challenged. *Id*. ¶20; Ex. 96, ¶261.

Defendants make identical arguments against Delorey and Goldstein for failing to inspect "other fields" in OpenAI's training data. Mem.12-13, 25. Those fail for the same reasons described above. *Supra* § II.A.2.b.

**D. Etzioni**

Defendants object to four figures (graphs) with incorrect numbers in Etzioni's reports, which Etzioni later withdrew. Mem.25-27. On that basis, Defendants seek to exclude fifty-seven paragraphs of testimony regarding the importance of news content for LLM training. *Id.*; Mot.8. This extraordinary request should be denied.

*First*, the figures ***have already been withdrawn*** by Etzioni.[10] Thus, they will not be presented at trial and are not relied upon for summary judgment purposes. There is no need for the Court to take any action.

*Second*, Defendants' argument turns on the mistaken premise that Etzioni's testimony was "based on" the three figures. Mem.25-27. But the withdrawn figures were neither evidence nor the basis for Etzioni's opinions; they were demonstratives meant to summarize a subset of evidence cited and analyzed throughout the relevant sections of Etzioni's report that Defendants do not address, let alone challenge. That evidence includes: 16 academic papers, 36 documents produced by Defendants, and 14 transcripts of Defendants' employees' depositions (Ex. 70(A) [Etzioni-Corrected-Opening-Rpt.], ¶¶71-122; Ex. 70(B) [Etzioni-Corrected-Rebuttal-Rpt.], ¶¶311-15), including testimony from an OpenAI employee confirming that "[redacted]" Ex. 70(A), ¶89. Defendants cite no authority to support their argument, much less a case in which a court excluded unchallenged and well-supported expert testimony *en masse* because the expert made mistakes in demonstratives. And even if Etzioni had not withdrawn the

[10] Etzioni used an AI tool to illustrate the findings of certain academic papers, which resulted in numerical errors. Thus, Dr. Etzioni withdrew the figures, and the corrections are described in his errata. Ex. 70(D). For the Court's convenience, Plaintiffs provide both the original and corrected versions of Dr. Etzioni's Opening and Rebuttal reports. ECF1806/1545 (original); Ex. 70(A)-(B) (corrected).

demonstratives, they would have been "the subject of cross-examination"—not a basis for exclusion. *LifeTree Trading Pte., Ltd. v. Washakie Renewable Energy, LLC*, No. 14-cv-9075(JPO), 2018 WL 2192186, at *7 & n.4. (S.D.N.Y. May 14, 2018).

Separately, Defendants ask the Court to exclude ¶227 of Etzioni's Opening Report. Mem.27-28. Though unexplained, Defendants' assertion appears to be that Etzioni's conclusion in ¶227—that Copilot users may have otherwise visited News Plaintiffs' websites—is unsupported. But the surrounding paragraphs support this conclusion using Microsoft's own internal documents and studies, including a study showing that [redacted], and a document explaining that [redacted]" Ex. 70(A), ¶¶225-228. Defendants offer no basis to preclude Etzioni from discussing internal Microsoft documents and what they show.

## III. NEWS PLAINTIFFS' INDUSTRY EXPERTS

### A. Stricchiola

#### 1. Stricchiola's Click-Through-Rate Opinions Satisfy FRE 702.

Jessie Stricchiola is an expert in digital platforms with decades of experience in search engines and related internet technologies, including web analytics. Ex. 73 [Stricchiola-NYT/DNP-Opening-Rpt.], ¶9. Defendants' challenges to her click-through-rate ("CTR") analysis and opinions provide no basis for exclusion.

Defendants allege Stricchiola "misleadingly" opines on economic harm. She does not. She opines about the lower CTR associated with Copilot and ChatGPT, relative to search CTRs, and the resulting reduction in opportunities to monetize user engagement on News Plaintiffs' websites. *Id*., ¶¶379-449. These opinions fall squarely within her expertise and are relevant to the issues of substitution and fair use. *See Associated Press v. Meltwater U.S. Holdings, Inc.*, 931 F.Supp.2d

537, 553-55 (S.D.N.Y. 2013) (low CTR associated with accused copyright infringer's news clipping service indicated substitution). Defendants identify no flaws in her CTR analysis that would preclude other experts from relying on it to support their own opinions about economic or substitutional harm.

Defendants also allege Stricchiola made admissions showing that her CTR opinion lacks significance. Mem.28-29. The opposite is true. She testified that lower CTRs for Copilot and ChatGPT users result in fewer opportunities for The Times to monetize user engagement. Malhotra Ex. 15 [Stricchiola-Depo.-Tr.], 80:19-81:4. She offers unchallenged opinions about how Defendants' GenAI products substitute for News Plaintiffs' products and content, based partially on admissions made by ***Defendants'*** employees and experts. Ex. 73, ¶¶423-427; Ex. 74 [Stricchiola-Rebuttal-Rpt.], ¶¶111-114, 127-131, 133-135; Ex. 97 [Stricchiola-Reply-Rpt.], ¶¶69, 71 & n.84. Stricchiola also did not admit that "GenAI tools are 'complementing – not replacing search engine behavior.'" Mem.29. That statement came from a third-party document Stricchiola cited for different purposes; she neither endorsed nor agreed with it. Malhotra Ex. 15, 85:7-18, 90:21-91:6, 93:8-95:5, 96:21-97:9; Malhotra Ex. 16.

Defendants criticize Stricchiola for not analyzing the amount of traffic directed to News Plaintiffs' websites after Defendants introduced their GenAI products. Mem.29-30. But Stricchiola never purports to perform such analyses, nor are they required. Moreover, Malackowski did these analyses and reached no conclusion because so many factors affect traffic. *See infra* at § IV.A. Defendants rely on their experts' traffic analyses to try to counter Stricchiola, Mem.29, but that provides no basis for exclusion, and these analyses are flawed and unreliable regardless, ECF2067 at 13-16, 33-34.

Finally, Defendants complain that Stricchiola did not analyze OpenAI's CTR data and extrapolated from her Copilot CTR analysis to ChatGPT without explanation. Mem.30. Not so. Stricchiola thoroughly examined OpenAI's CTR data and found it incomplete and unreliable. Ex. 73, ¶¶428-436; Ex. 74, ¶¶117-126; Ex. 97, ¶¶86-114. She separately concluded that the Copilot/Bing CTR data was a reasonable proxy to use for OpenAI, including because it is consistent with third party studies showing CTR differences between GenAI products and search. Ex. 73, ¶¶428-436; Ex. 74, ¶¶117-126.

2. Stricchiola's Meta Tag Opinions Satisfy FRE 702.

Microsoft seeks summary judgment on an implied license defense, arguing News Plaintiffs should have used the NOARCHIVE meta tag on their web pages to instruct Copilot not to use their content for grounding. Microsoft-News-MSJ.31-34. Microsoft's license defense fails for multiple independent reasons, News-MSJ-Opp.45-51, and one is that Microsoft continued to access and ground on content marked NOARCHIVE. Microsoft seeks to exclude Stricchiola's analysis and opinions related thereto. Its criticisms are cross-examination points at best.

Stricchiola analyzed web logs Microsoft provided via Bing Webmaster Tools showing many instances—***after*** News Plaintiffs applied NOARCHIVE meta tags to their web pages—where Microsoft disregarded those tags and used the content on those pages for grounding. Ex. 73, ¶¶280-297. Microsoft's sole challenge to Stricchiola's analysis is that she failed to rule out a purported "alternative explanation or cause"—*i.e.*, that Microsoft never knew about the tags because Bing does not re-crawl older web pages as frequently and Plaintiffs did not use the "IndexNow" feature to trigger re-crawl. Mem.30-32.

Microsoft's proffered "alternative explanation or cause" does not change or undermine Stricchiola's analysis or her opinion that Microsoft disregarded meta tags. Regardless of

Microsoft's excuse, the fact remains that News Plaintiffs added NOARCHIVE tags signaling to Microsoft that content could not be captured and used for grounding, but Microsoft did so nonetheless. Microsoft even admits to this in its Summary Judgment Opposition brief, alleging it had "bugs" to fix. ECF2217 at 18. Microsoft's disagreement with Ms. Stricchiola's conclusions is a basis for cross-examination, not exclusion. Moreover, Microsoft did not inform webmasters that they should use the IndexNow feature in the September 2023 blog post it relies on, ECF1782-26, or on any other web page Defendants cite, and Microsoft admits that even IndexNow does not require or ensure re-crawl.[11] Only Microsoft decides whether and when Bing will re-crawl pages, so its failures to identify newly tagged pages are Microsoft's responsibility.

Microsoft's case law similarly fails to support exclusion. *Nimely v. City of New York*, 414 F.3d 381, 396 (2d Cir. 2005), concerned expert opinions about witness credibility, which is not an issue here. In *R.F.M.A.S., Inc. v. So*, 748 F.Supp.2d 244, 273 (S.D.N.Y. 2010), the court excluded a damages causation opinion where an expert failed to investigate other causes; here, Microsoft was the cause of disregarded meta tags even under its "alternative explanation."

Microsoft asserts that Stricchiola's opinions were excluded in *Ryoo Dental v. Han*, 2016 WL 6138413, at *5 (C.D. Cal. April 25, 2016), but in that case she was defendant's expert, whereas the ***plaintiff*** wanted to offer one of her opinions on its own behalf to save its case. The court understandably held there was no foundation to admit the opinion given that Stricchiola had offered ***no*** opinions on plaintiff's behalf.

[11] *See* Ex. 98 [Bing Webmaster Tools URL Submission FAQ] ("Bing does not guarantee that all content submitted will be indexed."); Malhotra Ex. 15, 156:2-157:4 (Stricchiola explaining it was unnecessary to analyze indexation status because Microsoft never told web publishers the NOARCHIVE tag would not work if Microsoft did not re-index the page).

**B. Krein**

In discovery, OpenAI refused to produce internal scrape logs documenting OpenAI's collection of copyrighted text from ZD websites in contravention of robots.txt instructions. ECF1383 at 2, 1315 at 2, 1402 at 2. OpenAI later submitted a report from Feamster, who erroneously opines that OpenAI honors robots.txt directives. Krein's rebuttal shows OpenAI mass-scrapes in violation of robots.txt disallows. Krein is a highly-credentialed, PhD computer scientist and still, OpenAI's unfounded criticisms (Mem.32-33) do not go to admissibility.

***First***, OpenAI criticizes Krein's reliance on internal revision history in conjunction with Wayback Machine records. But Krein did so "in lieu of using the scrape logs," which OpenAI failed to produce and which "would have been the most direct evidence" of OpenAI's scraping. Ex. 104 [Krein-Depo.-Tr.], 227:12-20. Krein used the next best evidence—***ZD's*** internal revision history—supplemented with Wayback Machine records as "a further check" and where logs were unavailable. *Id*. at 114:22–115:12, 117:21-120:16; Ex. 72 [Krein-Amended-Rebuttal-Rpt.] ¶53. Contrary to OpenAI's assertion (Mem.32-33), Krein applied source code and Wayback records consistently. Ex. 104, 114:22–115:12, 116:7-12, 117:21-120:16. OpenAI simply disregards, for example, lines of code where Krein analyzes BabyCenter (Malhotra Ex. 52 at ll. 16,106-127). OpenAI cites neither evidence nor case law suggesting Krein's methodology was improper. Indeed, Krein's methodology agrees with OpenAI's expert at least 94.8% of the time. *Compare* Ex. 76 [Krein-Second-Supplemental-Rpt.] ¶7 ([redacted] scrapes applying Krein's methodology); *with id*. ¶8 ([redacted] scrapes according to Feamster's methodology). Further, OpenAI's conjecture that the internal ZD logs supposedly "precede public deployment" or that changes occurred "without …log entries" is a basis for cross-examination, not exclusion—particularly given OpenAI's refusal to produce its own logs. *Rutledge*, 182 F.4th at 184–85 (court does not "adjudicate which points in the debate are more persuasive; that is the role of the jury").

***Second***, OpenAI distorts purported "errors." Mem.33. Krein was initially provided with an incorrect dataset, Ex. 104, 30:13-31:4, 45:8-20, which had nothing to do with Krein or his methodology, and he promptly issued a corrected report. *Id*. at 29:7-46:4. Krein likewise corrected other bugs that he identified in his script, in accordance with Fed. R. Civ. P. 26(e). Ex. 75 [Krein-Supplemental-Rpt] ¶¶3-8,12-17; Ex. 76, ¶4. Such corrections are part of a robust scientific process, not a basis for exclusion. *Campbell ex rel. Campbell v. Metro. Prop. & Cas. Ins. Co*., 239 F.3d 179, 186 (2d Cir. 2001) ("gaps or inconsistencies" go to weight, not admissibility). [12]

While OpenAI complains that Krein (properly) supplemented his report, OpenAI fails to disclose that its own expert (Berg-Kirkpatrick) postponed his deposition to supplement his report. Ex. 79 [7/14/26-Email-from-J.Barretto]. Another OpenAI expert (Krishnan) supplemented after his deposition.[13] Here, ZD notified OpenAI of Krein's forthcoming supplement, and offered to postpone Krein's deposition, but OpenAI decided to proceed.[14]

***Third***, OpenAI wrongly criticizes Krein for counting multiple "scrapes" on one page. Mem.33. But his scripts properly count each scrape individually. Ex. 72, ¶74; Ex. 76, ¶6. If OpenAI believes Krein mischaracterizes the scrapes, OpenAI can raise that via cross-examination. *Micro Chem., Inc. v. Lextron, Inc*., 317 F.3d 1387, 1392 (Fed. Cir. 2003) (court does not evaluate "correctness of facts underlying" expert's testimony).

---

[12] *Daubert v. Merrell Dow Pharms., Inc*., 509 U.S. 579, 597 (1993) ("conclusions are subject to perpetual revision"); *Cassava Scis., Inc. v. Bredt*, No. 1:22-cv-9409-GHW, 2024 WL 1347362, at *8 (S.D.N.Y. Mar. 28, 2024) ("conclusions of empirical research are tentative and subject to revision"); *Malletier v. Dooney & Bourke, Inc*., 525 F.Supp.2d 558, 674–75 (S.D.N.Y. 2007) ("possible oversight does not justify exclusion").

[13] *See* Ex. 104, 214:11-215:6 (noting Krishnan's supplement); Ex. 80 [Krishnan-Supplemental-to-Wenger] ¶6; Ex. 81 [8/6/26-Email-from-T.Gorman].

[14] Ex. 105 [8/6/26-Email-from-C.Naggar] (notifying counsel of Dr. Krein's supplement); Ex. 104, 140:5-141:14 (offering inspection of incorrect dataset); 56:1-5, 214:23-216:4; 217:5-12, 231:13-24 (discussing forthcoming calculations and potential to continue deposition).

***Fourth***, OpenAI also raises an incoherent four-line criticism of Krein's substitution analysis. Krein "appl[ies] the formula provided by [OpenAI expert] Dr. Berg-Kirkpatrick." Ex. 72, ¶¶86-87. And he does so based on his unchallenged expertise in the field of large language models and AI systems. Ex. 104, 205:18-207:21. OpenAI fails to explain why Krein should be prohibited from doing so.

**C. Bell, Owen, and Cross**

1. Market Harm and Substitution Opinions.

Professors Bell, Owen, and Cross—industry experts in journalism and media—offer sufficient bases for their opinions on market harm and substitution, which are relevant to factor four of the fair use defense. Defendants criticize them for not conducting a "data driven" analysis. This is both wrong and not required.

The fourth fair use factor considers "the effect of the use upon the potential market for or value of the copyrighted work." 17 U.S.C. § 107(4). The Supreme Court has stated the objective of *Daubert* is not to ensure each expert relies on empirical data, but to ensure the reliability and relevancy of expert testimony, whether it is based on professional studies or personal experiences. *Kumho Tire Co., Ltd. V. Carmichael*, 526 U.S. 137, 152 (1999); *accord Hachette* 115 F.4th at 193 (rightsholder need not offer "empirical data" and may rely on "logical inferences."). Plaintiffs' experts properly rely on data and inferences.

Defendants do not attack the experts' qualifications, which include decades of experience working in journalism and studying technology's impact on the news industry. Ex. 99 [Cross-Rpt.] at 3-5; Ex. 83 [Bell-Opening-Rpt.] ¶¶10-23; Ex. 84 [Owen-Opening-Rpt.] ¶¶10-20; Malhotra Ex. 13 [Owen-Depo.-Tr.], 27:22-28:2. Each expert properly applied this expertise to show how existing studies and data demonstrate that Defendants' GenAI products and those like them substitute for News Plaintiffs' journalism. *Rutledge*, 182 F.4th at 184 ("Experts need not conduct

studies of their own in order to opine on a topic; a review of other studies and scientific literature can be enough to qualify experts to testify and to make that proposed testimony reliable.") (quoting *In re Mirena IUD Prods. Liab. Litig.*, 169 F.Supp.3d 396, 412 (S.D.N.Y. 2016)); Fed. R. Evid. 703 ("expert may base an opinion on facts or data. . . that the expert has been made aware of").

Defendants incorrectly claim that Plaintiffs' experts "cite no studies specifically addressing the impact of Defendants' products" and did not look at "Plaintiffs' audience metrics or financial data." Mem.34-35. Bell cites studies by the Tow Center indicating that ChatGPT and Copilot fail to accurately cite the sources of news materials they reproduce. *See* Ex. 83, ¶¶25, 154-55, 160 n.6. Bell cites other studies on ChatGPT, including the Times's audience and brand surveys. *See id.* ¶¶93, 108, 116-118; Ex. 86 [Bell-Reply-Rpt.], ¶39. Owen similarly cites studies that reference ChatGPT and Copilot. Ex. 84, ¶¶116-117, 120, 140 (citing Tow Center/CJR studies that included ChatGPT when evaluating GenAI's failure to cite sources and propensity for mispresenting publishers' content); Ex. 85 [Owen-Rebuttal-Rpt.], ¶33; Ex. 87 [Owen-Reply-Rpt.], ¶36. Likewise, Cross cites several studies discussing the impact of ChatGPT. *See, e.g.,* Ex. 99 [Cross-Rpt.] at 18, n.31; *id*. at 19 n.33-34, 27 n.47, 29 n.48. Cross also considered CIR's and The Intercept's user data and financial data, including web traffic data, trend analyses, revenue and expense data, audience analyses, and consolidated financial statements. *See* Ex. 100 [Cross-Materials-Reviewed] (citing, e.g., CIR0013767; CIR0085125; CIR0676966-CIR0676964; CIR0676969; CIR0676970; INT0007389; INT0225997-INT0225974; INT0215300-INT0215332; INT0018789-INT0018790).

These experts' reliance on data about third-party GenAI tools like Google AI Overviews is not "irrelevant." Mem.34. Such evidence is relevant to the market harm resulting from

"unrestricted and widespread conduct" like Defendants'. *Hachette*, 115 F.4th at 189.[15] *See* ECF1728 at 72-74. Defendants' experts opine that News Plaintiffs have been harmed by Google AI Overviews, relying on some of the same studies. *See, e.g.*, Ex. 101 [Goldfarb-NYT-Opening-Rpt.], ¶169. Fair use law confirms the relevance of Plaintiffs' experts' consideration of impacts of technologies like Google AI Overview—which, like Defendants' products, uses an LLM and the RAG process built on publishers' copyrighted works to deliver news to users —in addition to ChatGPT and Copilot. Defendants' inapposite cases, Mem.34, do not support their arguments to the contrary.

2. Bell and Owen's Reliance on Audience Surveys.

Bell and Owen's reliance on Times and *Daily News* internal audience surveys as one of a number of inputs is appropriate and relevant. These surveys corroborate the many other studies and sources Bell and Owen rely on to conclude that people are increasingly using GenAI for news. *See, e.g.*, Ex. 83, ¶¶113-116 (citing outside studies showing growing use of GenAI for news), ¶117-118 (The Times's internal surveys corroborate those studies); Ex. 84, ¶44 n.40 (ChatGPT's newsletter states they receive 1 million prompts per week related to local news), ¶98 n.136, ¶104 n.147-48; Ex. 87, ¶42 (*Daily News* GenAI survey corroborates other evidence of harm). The challenged surveys are particularly relevant because they examine the behavior and intentions of actual Times and *Daily News* readers. *See* Ex. 86, ¶56; Ex. 87, ¶42. If anything, Defendants' complaints about these surveys go to weight, not admissibility. *See infra* at § IV.A.

[15] *See, e.g.*, Ex. 88 [Malackowski-Reply-Rpt.] App.3 (comparing Defendants' generative AI products to Google AI Overviews); "How Google's AI Overviews Work, and How to Turn Them Off (You Can't)," Wired, May 16, 2024, https://www.wired.com/story/google-ai-overviews-how-to-use-how-to-turn-off/; "Ways to Search (AI Overviews)," Google, https://search.google/ways-to-search/ai-overviews/.

Defendants' cited cases are inapposite. In *Hayden v. Int'l Bus. Machs.. Corp.*, 2025 WL 1697021, at *8 (S.D.N.Y. June 17, 2025), the expert's sole basis for his opinion about products he never personally inspected were PowerPoint slide graphics depicting the products. In *Munoz v. Orr*, 200 F.3d 291, 300-302 (5th Cir. 2000), an expert's statistical analysis was excluded because he relied *solely* on the plaintiffs' data compilations. In *Forte v. Liquidnet Holdings, Inc.*, 675 Fed.App'x 21, 23-24 (2d Cir. 2017), a statistical analysis expert "simply input" the plaintiff's payment data without validation and failed to perform a regression analysis before forming his unreliable opinion.

3. Bell's News and "Pink Slime" Lists.

Defendants incorrectly claim that Bell failed to apply a reliable methodology or disclose her criteria for preparing her news and pink slime domain lists. *See* Mem.36. She explained her methodology: she compiled her list of news websites using three outside sources that professionals in her industry regularly rely on, manually reviewed every publisher cited in more than five unique conversations from Defendants' samples to verify that they do original reporting and employ journalists, and relied on these outside sources, her decades of experience and industry knowledge, and her own independent research into publisher staff lists and reporting. Ex. 89 [Bell-Rebuttal-Rpt.] ¶207; Ex. 90 [Bell-Depo.-Tr.] at 240:20-241:16, 242:23-248:18, 249:1-252:1. To compile her list of pink slime websites, Bell relied on independent sources including the list her team at the Tow Center for Digital Journalism has compiled and maintained since 2019 using various forensic tools identified in her report. *See* Ex. 89, ¶¶208-210. There is no basis to exclude these lists.

**D. Tease**

Tease opines that a reasonable copyright lawyer would conclude OpenAI had reason to know removing CMI would further or conceal infringement. Under the DMCA, such reason to know involves facts, not legal issues or mental states. *See* 17 U.S.C. §1202(b)(1); *see also Middle*

*Mkt. Fin. Corp. v. D'Orazio*, No. 96-cv-8138, 2002 WL 31108260, at *8 (S.D.N.Y. Sept. 23, 2002) (allowing opinion by legal expert); *Highland Cap. Mgmt., L.P. v. Schneider*, 551 F.Supp.2d 173, 180-83 (S.D.N.Y. 2008) (allowing testimony on industry customs and practices but not mental state).

Defendants' case, *Deere & Co. v. Kinze Mfg., Inc.*, 2024 WL 1235569 (S.D. Iowa Feb. 8, 2024), ***undermines*** their position. Mem.37. *Deere* barred Tease from opining on whether counsel committed inequitable conduct, which "would simply tell[] the court what conclusion to reach," but permitted her opinion on how a reasonable attorney would advise on certain issues. *Deere*, 2024 WL 1235569 at *8. Thus, Tease may permissibly testify about how a reasonable copyright lawyer would advise given factual circumstances surrounding the second scienter requirement.

## IV. NEWS PLAINTIFFS' HARM EXPERTS

### A. Malackowski

James Malackowski is a co-founder and Senior Managing Director of Ocean Tomo and a Certified Licensing Professional. He has testified as an IP economics expert more than 100 times. In 2022, he was inducted into the IP Hall of Fame.

Malackowski opines on economic issues related to fair use, including market harm and substitution, on behalf of The Times, *Daily News* Plaintiffs, and Ziff Davis. He considered documents and deposition testimony of Defendants, Plaintiffs, and third parties; third party reports and studies; financial, subscription and traffic data; and opinions of other experts. Ex. 91 [Malackowski-DNP/NYT-Opening-Rpt.], pp.4-10, App.2; Ex. 102 [Malackowski-DNP/NYT-Rebuttal-Rpt.], pp.4-5, App.2.

Defendants challenge Malackowski's harm opinions on two bases: (1) "he has no "data-driven basis" for finding substitution or harm"; and (2) "instead of presenting data to support his

market-harm opinions, Malackowski relies on the report of another expert (Stricchiola)." Mem.38-39. Both arguments are incorrect.

Malackowski ***did*** analyze data. He analyzed survey data from Plaintiffs, data showing that CTR to News Plaintiffs' content was substantially lower from Microsoft Copilot than from (non-GenAI) Bing Chat, and data and testimony on search referral declines. *E.g.* Ex. 91, pp.46-56. He also analyzed data from numerous third-party studies and data regarding harm and substitution caused by Google's AI Overviews, including Defendants' experts' admissions that [redacted]. Ex. 102, pp.19-20, 54-55. This data supports Malackowski's opinions that Defendants' infringement—and widespread conduct of a similar sort—is causing substitution and market harm. *See Hachette,* 115 F.4th at 189.

Defendants seemingly argue that Malackowski's analysis is unreliable because he did not conduct the same type of "empirical analysis" as Defendants' experts to assess the impact of ChatGPT and Copilot on News Plaintiffs' revenue, subscriptions, or traffic. Mem.39. But Malackowski did analyze the same data. He just did not reach the same conclusion as Defendants' experts, because "[t]here were too many countervailing facts within the marketplace that mask the ability to parse that out." Malhotra Ex. 11 [Malackowski-7/16/26-Depo.Tr.], 343:12-345:13; Ex. 92 [Malackowski-7/16/26-Depo.-Tr.], 413:22-414:1.16.

Defendants are also wrong that an expert analyzing substitution or market harm must evaluate "information… such as sales, revenue, or consumer survey data" (Mem.39), as the Second Circuit made clear recently in *Hachette*. 115 F.4th at 194.

[16] The large number of factors affecting revenues, subscriptions, and traffic, and the inability of any expert to correctly "parse out" their effects on the overall trend, are precisely why News Plaintiffs seek to exclude the opinions of Drs. Goldfarb and Tucker. ECF2067 at 13-16, 33-34. Malackowski recognized this problem with the "trends" data and consequently turned to other evidence to assess market harm and substitution. Ex. 102, pp.69-73.

Defendants cite inapposite antitrust cases with distinguishable facts. *See supra* at § I.B. (distinguishing *Laumann*). In *Berlyn,* the expert offered an opinion on an economic concept despite lacking economic expertise. *Berlyn, Inc. v. Gazette Newspapers, Inc.*, 214 F.Supp.2d 530, 539 (D. Md. 2002). Here, Defendants do not dispute Malackowski's expertise or that his opinions are tethered to the facts of the case. They merely argue he should have analyzed the same data Defendants' experts analyzed. But Malackowski did analyze this data. He just found it inconclusive.

Defendants' second challenge alleges that none of the other materials Malackowski relies on "purport to show substitution between ChatGPT and Copilot and News Plaintiffs' materials." Mem.40. But this is a criticism of Malackowski's conclusions, not his analysis or methodology, and is not a proper basis for a *Daubert* challenge. *Daubert*, 509 U.S. at 595 ("The focus, of course, must be solely on principles and methodology, not on the conclusions that they generate"). Defendants are also wrong. They argue that Malackoski's reliance on Stricchiola's CTR analysis does not provide evidence of substitution, but similar evidence was found in *Associated Press v. Meltwater* to constitute evidence of substitution. *See supra* at § III.A.1. Defendants also argue that plaintiff-administered surveys are "inadmissible," and "neither measured whether ChatGPT or Copilot changed respondents' behavior." Mem.40. Plaintiffs disagree, and neither argument is a basis for excluding the testimony. *U.S. v. Locascio*, 6 F.3d 924, 938 (2d Cir. 1993) (facts forming basis for expert's opinions or inferences need not be admissible "[i]f of a type ***reasonably relied*** upon by experts in the particular field"); *Schering Corp. v. Pfizer Inc.*, 189 F.3d 218, 227 (2d Cir. 1999) (surveys are sometimes best admitted as bases of expert testimony pursuant to Rule 703).

Finally, Defendants' motion cites a case for the proposition that expert testimony can be stricken if it only provides a summary of what other experts have said "without application of his

own expertise." Mem.40-41. But Defendants have not argued Malackowski did this. He relied on numerous data points—including opinions of other experts, such as Stricchiola (expert on digital platforms and web analytics) and Bell and Owen (journalism experts)—and applied his own economic, licensing, and market expertise, to conclude there was harm and substitution.

Defendants' challenges to Malackowski all fail.

**B. McDonald**

McDonald opines that GenAI products substitute for news. Defendants insist that McDonald identify individual diverted users or quantify lost subscriptions and advertising revenue, Mem.41-42, despite the fact that such "empirical data" is unnecessary. *See Hachette*, 115 F.4th at 193-95.

McDonald permissibly relied on TollBit click-through data. *In re Mirena IUS Levonorgestrel-Related Prods. Liab. Litig.*, 341 F.Supp.3d at 249-53 is distinguishable because it concerned expert testimony purporting to establish causation through studies that did not support that conclusion. Here, however, the TollBit data shows differences in referral behavior. Ex. 78 [McDonald-Rpt.], pp.13-14 (opining that, per TollBit, publishers' content is consumed at increasing rates while audience engagement is declining). He confirmed this by analyzing ChatGPT and Copilot responses to his news-related questions, concluding, "AI chats compete in the same market as news content." *Id*. Defendants' disagreement with this conclusion does not give them grounds to exclude it. *See Hachette*, 115 F.4th at 193-94.

Defendants wrongly allege that McDonald's regression analysis does not include GenAI usage as a variable. Mem.41-42. Omissions from regression analyses generally affect "probativeness, not [their] admissibility." *Floyd v. City of New York*, 861 F.Supp.2d 274, 288-90 (S.D.N.Y. 2012) (quoting *Bazemore v. Friday*, 478 U.S. 385, 400 (1986)). Defendants' cases are inapt. *Fed. Hous. Fin. Agency v. Nomura Holding Am., Inc.*, 2015 WL 539489, at *5 (S.D.N.Y.

Feb. 10, 2015), concerned the requirements for a valid control group in an analysis designed to isolate a particular causal effect; it did not exclude descriptive regression evidence concerning market trends. And McDonald relied on actual web traffic data, rather than "substitut[ing] mathematical assumptions for actual, readily-obtainable information." *Laumann*, 117 F.Supp.3d at 315.

## CONCLUSION

Defendants' Daubert motion should be denied.[17]

Dated: October 29, 2026

Respectfully submitted,

*/s/ Justin A. Nelson*
Justin A. Nelson (*pro hac vice*)
SUSMAN GODFREY L.L.P.
1000 Louisiana Street, Suite 5100
Houston, TX 77002
Tel.: 713-651-9366
jnelson@susmangodfrey.com

*Interim Lead Counsel for Class Plaintiffs*

[17] Defendants' identification of content impacted by their challenges is overbroad and disputed. If any of Defendants' challenges are granted, Plaintiffs reserve the right to confer with Defendants regarding impacted testimony and, if necessary, seek judicial relief.

*/s/ Davida Brook*
Davida Brook
SUSMAN GODFREY L.L.P.
1900 Avenue of the Stars, Suite 1400
Los Angeles, CA 90067
Tel.: 310-789-3100
dbrook@susmangodfrey.com

*Counsel for The New York Times Company*
*News Plaintiffs' Liaison Counsel*

*/s/ Steven Lieberman*
Steven Lieberman
ROTHWELL, FIGG, ERNST & MANBECK, P.C.
901 New York Avenue, N.W., Suite 900 East
Washington, DC 20001
Tel: (202) 783-6040
slieberman@rothwellfigg.com

*Counsel for The New York Times Company and*
*Daily News Plaintiffs*

**CERTIFICATE OF WORD COUNT COMPLIANCE**

Pursuant to Local Civil Rule 7.1(c), the foregoing Opposition complies with the Court's order of August 7, 2026 (ECF 1664), as it contains 11,678 words, according to the word processing program used to prepare it.

*/s/ Steven Lieberman*