UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK

IN RE: OPENAI, INC., COPYRIGHT INFRINGEMENT LITIGATION This document relates to: All Actions 23-cv-08292 23-cv-11195 24-cv-01515 24-cv-03285 24-cv-04872 25-cv-04315	25-md-3143 (SHS) (OTW) Hon. Sidney H. Stein Hon. Ona T. Wang PUBLIC—REDACTED VERSION

OPENAI'S STATEMENT OF UNDISPUTED MATERIAL FACTS IN SUPPORT OF ITS MOTIONS FOR SUMMARY JUDGMENT

Pursuant to Rule 56.1 of the Local Rules of the United States District Court for the Southern District of New York, OpenAI Defendants ("OpenAI") set forth the following undisputed material facts in support of its Motions for Summary Judgment.

FACTS SHARED BY THE BOOKS AND NEWS CASES

I. OpenAI Before ChatGPT

1. OpenAI was founded in 2015 as a research lab. Declaration of Joseph C. Gratz ("J. Gratz Decl.") ¶¶ 2–3; Exs. 1, 2.

2. OpenAI has published academic papers on machine learning and "natural language processing" ("NLP"). J. Gratz Decl. ¶¶ 5–14; Exs. 4–13.

3. NLP researchers hypothesized that language follows statistical patterns that could be used to revise or generate new text. Declaration of Christopher Callison-Burch ("C. Callison-Burch Decl.") ¶ 11.

4. NLP researchers ran experiments to evaluate different architectures for capturing linguistic patterns in statistical systems called "language models," using books, news articles, and other texts to "train" the models. C. Callison-Burch Decl. ¶¶ 11–17.

5. Google Translate and Microsoft Word's grammar checker are NLP applications. C. Callison-Burch Decl. ¶ 17.

6. As of 2019, OpenAI researchers described "[c]urrent systems" as being "better characterized as narrow experts rather than competent generalists," and set a goal of moving "towards more general systems which can robustly perform many tasks." J. Gratz Decl. ¶ 8; Ex. 7.

7. OpenAI ran a series of experiments using a new architecture it developed called the "generative pretrained transformer" ("GPT"). J. Gratz Decl. ¶¶ 7, 84; Exs. 6, 83.

8. OpenAI trained its early models on self-published books (GPT-1) and other text from the Internet (GPT-2). C. Callison-Burch Decl. ¶¶ 52, 59; J. Gratz Decl. ¶¶ 7–8; Ex. 6 at OPCO_MDL_000000385; Ex. 7 at OPCO_MDL_000200506.

9. OpenAI has observed a direct relationship between the size and diversity of its pre-training data and model performance. C. Callison-Burch Decl. ¶¶ 62, 67–73, 74–76; J. Gratz Decl. ¶¶ 8–9, 21, 26; Ex. 7 at OPCO_MDL_000200506; Ex. 8; Ex. 20 at 49:4–14, 110:8–14; Ex. 25 at 76:4–6.

10. For GPT-3 and GPT-3.5, OpenAI used broader mixtures of data from the public Internet than GPT-1 and GPT-2, including compilations of books from Library Genesis. C. Callison-Burch Decl. ¶¶ 52, 60, 62, 117, 119; J. Gratz Decl. ¶¶ 7–8, 16, 19, 23, 26, 29, 54; Ex. 6 at OPCO_MDL_000000385; Ex. 7 at OPCO_MDL_000200506; Ex. 15 at 134:20–22; Ex. 18 at 10; Ex. 22 at 234:20–235:6; Ex. 25 at 67:17-68:4; Ex. 28 at OPCO_MDL_000000446; Ex. 53 ¶ 254.

11. No other LLMs at issue in this case besides GPT-3 and GPT-3.5 were trained on material downloaded from Library Genesis. J. Gratz Decl. ¶¶ 23–24; Ex. 22 at 235:23–236:8; Ex. 23 at Resp. to Interrog. No. 15.

12. OpenAI developed a suite of "post-training" techniques. C. Callison-Burch Decl. ¶¶ 90–98; J. Gratz Decl. ¶ 30; Ex. 29 at Resp. to Interrog. No. 16.

13. OpenAI created an interface for GPT-3.5 called "ChatGPT" and published a blog post on November 30, 2022 inviting people to try it out as a free "research preview." J. Gratz Decl. ¶¶ 23, 39; Ex. 22 at 235:20–22; Ex. 38 at OPCO_MDL_001003615.

14. OpenAI stated that it did so to "get users' feedback and learn about its strengths and weaknesses." J. Gratz Decl. ¶ 39; Ex. 38 at OPCO_MDL_001003615.

15. OpenAI's "original intent was to wind down the ChatGPT demo after [it] had the necessary learning and feedback." J. Gratz Decl. ¶ 40; Ex. 39 at 101:24–102:2.

II. OpenAI After ChatGPT

16. Within days of ChatGPT's November 30, 2022 release, more than a million people signed up to use it. J. Gratz Decl. ¶¶ 39, 41; Ex. 38; Ex. 40 at Resp. to RFA No. 66.

17. The Harvard Business Review described ChatGPT as a "tipping point for AI" and noted that the model had "crossed a threshold" to become "genuinely useful for a wide range of tasks, from creating software to generating business ideas to writing a wedding toast." J. Gratz Decl. ¶ 43; Ex. 42.

18. The New York Times ("The Times") published an article stating that "ChatGPT is, quite simply, the best artificial intelligence chatbot ever released to the general public" and praising OpenAI's "commendable steps to avoid the kinds of racist, sexist and offensive outputs that have plagued other chatbots." J. Gratz Decl. ¶ 42; Ex. 41 at 1, 3.

19. OpenAI developed techniques to enable language models to understand audio, images, and video, including by engaging in a verbal dialogue with users. J. Gratz Decl. ¶ 44; Ex. 43 at OPCO_MDL_001934128.

20. OpenAI created a mechanism for language models to browse the Internet for factual information relevant to a user's query. J. Gratz Decl. ¶ 45; Ex. 44 at OPCO_MDL_000000522.

21. Over a billion people around the world use OpenAI's LLMs. J. Gratz Decl. ¶ 46; Ex. 45 at 4.

22. ChatGPT is used for many purposes, including education, healthcare, and scientific research. J. Gratz Decl. ¶¶ 47–52; Ex. 46 at 11, Fig. 8; Ex. 47 at OPCO_MDL_001950137; Ex. 48 at 31:1–8; Exs. 49–51.

III. How OpenAI Transforms Training Data to Build New Products

23. Since early 2020, training OpenAI's LLMs has required hundreds of billions to tens of trillions of tokens from diverse sources. C. Callison-Burch Decl. ¶¶ 62, 67–73, 74–76; J. Gratz Decl. ¶¶ 17–18; Ex. 16 at OPCO_MDL_002071701; Ex. 17 ¶ 7; Ex. 20 at 110:8–14.

24. GPT-3 was pre-trained on approximately 300 billion tokens. J. Gratz Decl. ¶ 18; Ex. 17 ¶ 4.

25. GPT-3.5 was pre-trained on [redacted] tokens. J. Gratz Decl. ¶ 18; Ex. 17 ¶ 6.

26. GPT-4 was pre-trained on approximately [redacted] tokens. J. Gratz Decl. ¶ 18; Ex. 17 ¶ 7.

27. GPT-3.5 Turbo was pre-trained on approximately [redacted] tokens. J. Gratz Decl. ¶ 9; Ex. 17 ¶ 8.

28. GPT-4o was pre-trained on approximately [redacted] tokens. J. Gratz Decl. ¶ 18; Ex. 17 ¶ 9.

29. GPT-4o mini was pre-trained on approximately [redacted] tokens. J. Gratz Decl. ¶ 18; Ex. 17 ¶ 10.

30. The Internet provides a singularly diverse, global source of text for use in LLM training. C. Callison-Burch Decl. ¶¶ 14–17, 60, 76; J. Gratz Decl. ¶¶ 8, 53, 55; Ex. 7 at OPCO_MDL_000200506; Exs. 52, 54.

31. OpenAI model developers gather data using the same web crawling techniques that search engines like Google have used for decades to gather broad swaths of Web content. C.

Callison-Burch Decl. ¶ 14, 15–17; J. Gratz Decl. ¶¶ 26, 85; Ex. 25 at 78:7–22, 160:19–161:5; Ex. 84 at 37:10–38:6.

32. OpenAI used materials from publicly available websites, including materials from Common Crawl. J. Gratz Decl. ¶¶ 19, 27, 54; Ex. 18 at 29, 32; Ex. 26 at 283:13–17; Ex. 53 ¶ 229.

33. Common Crawl is a non-profit organization that created a publicly available index of billions of webpages. C. Callison-Burch Decl. ¶ 16; J. Gratz Decl. ¶ 96; Ex. 95 ¶ 112.

34. OpenAI does not make its training data accessible to the public. J. Gratz Decl. ¶ 158; Ex. 157 ¶¶ 76–81.

35. Once collected, web text data is de-duplicated and processed to extract language data. C. Callison-Burch Decl. ¶¶ 78–89; J. Gratz Decl. ¶¶ 17, 30; Ex. 16 at OPCO_MDL_002071701; Ex. 29 at Supp. Resp. to Interrog. No. 16.

36. Language data is also "tokenized," a process by which text or other inputs are transformed into numbers or "tokens." C. Callison-Burch Decl. ¶¶ 23–24; J. Gratz Decl. ¶¶ 31–35; Ex. 30 at Resp. to Interrog. No. 17; Ex. 31 at Resp. to Interrog. No. 24; Ex. 32 at Resp. to Interrog. No. 19; Ex. 33 at Resp. to Interrog. No. 25; Ex. 34 at Resp. to Interrog. No. 24.

37. OpenAI splits tokenized data into excerpts large enough for the model to derive the semantic and linguistic meaning of the words therein based on surrounding context. J. Gratz Decl. ¶ 21; Ex. 20 at 56:14–57:9.

38. OpenAI then uses tokenized excerpts for "pre-training," an iterative process whereby the model learns broad linguistic patterns that it can use to comprehend and generate new text—*i.e.*, to *generalize* broader ideas and principles of language from specific examples.

C. Callison-Burch Decl. ¶¶ 29–32, 33–36, 37–39; J. Gratz Decl. ¶¶ 21, 23, 26–27; Ex. 20 at 21:9–23, 56:15–18, 104:15–17; Ex. 22 at 24:18–22; Ex. 25 at 218:20–219:6; Ex. 26 at 308:1–5.

39. The model "learns" linguistic patterns by encoding them in a network of up to trillions of *parameters* or *weights*—*i.e.*, numerical values that function like coefficients in a highly complex math equation that generates predictions. C. Callison-Burch Decl. ¶¶ 20, 22–28, 29–32; J. Gratz Decl. ¶¶ 36, 54, 57; Ex. 35 at 83:25–84:11; Ex. 53 ¶ 74; Ex. 56 ¶¶ 46, 71.

40. LLM parameters begin the training process as a set of random numbers. C. Callison-Burch Decl. ¶ 20; J. Gratz Decl. ¶ 21; Ex. 20 at 19:17–18.

41. During each step of pretraining, the model takes a randomly selected excerpt from the training corpus and generates a *probability distribution* reflecting predictions of different words that might come next, along with scores indicating the model's level of confidence as to each prediction. C. Callison-Burch Decl. ¶ 30; J. Gratz Decl. ¶¶ 19, 36; Ex. 18 at 19; Ex. 35 at 82:21–83:13.

42. The algorithm compares that probability distribution against the *reference value*—*i.e.*, the actual next word in the excerpt—and uses the results of that comparison (*train loss*) to update the model's parameters. C. Callison-Burch Decl. ¶¶ 30–31, 70; J. Gratz Decl. ¶¶ 7, 29, 36, 58; Ex. 6 at OPCO_MDL_000000384–OPCO_MDL_000000386; Ex. 28 at OPCO_MDL_000000469; Ex. 35 at 74:13–19; Ex. 57 at 86:1–12.

43. This cycle repeats millions of times across many different excerpts, and the same model uses those excerpts to make a series of incremental adjustments to its parameters. C. Callison-Burch Decl. ¶ 31.

44. As training progresses, the model's parameters eventually encode basic linguistic rules that improve the quality of its predictions—leading to a steadily narrowing gap and a lower

train loss. C. Callison-Burch Decl. ¶ 31; J. Gratz Decl. ¶¶ 29, 36; Ex. 28 at OPCO_MDL_000000469; Ex. 35 at 76:25–77:15.

45. During pre-training, OpenAI periodically uses a subset of collected data set aside for testing purposes (the *test set*). C. Callison-Burch Decl. ¶ 45; J. Gratz Decl. ¶¶ 8, 21, 58; Ex. 7 at OPCO_MDL_000200510; Ex. 20 at 181:7–10; Ex. 57 at 57:2–13.

46. OpenAI uses the *test set* to measure how accurately the model makes predictions on data it has never seen and never used to update its parameters, resulting in a metric called *test* or *validation loss*. C. Callison-Burch Decl. ¶¶ 45–46; J. Gratz Decl. ¶ 23; Ex. 22 at 24:18–25.

47. Test loss, not train loss, is the core metric OpenAI uses to evaluate pretraining. C. Callison-Burch Decl. ¶¶ 46–47; J. Gratz Decl. ¶¶ 21, 23, 37; Ex. 20 at 101:18–102:12; Ex. 22 at 24:18–25; Ex. 36 at 439:1–8.

48. Test loss is a direct measure of the model's ability to *generalize*: it evaluates the degree to which the model has learned general patterns that it can apply to new, unseen data. C. Callison-Burch Decl. ¶¶ 46, 48; J. Gratz Decl. ¶¶ 21, 37; Ex. 20 at 51:8–13, 185:2–18; Ex. 36 at 412:7–413:7.

49. Once pretraining is complete, OpenAI can then use a pretrained model to generate entirely new text. C. Callison-Burch Decl. ¶ 35.

50. Autoregressive text generation occurs through a process called *inference,* whereby the model processes an input sequence and then uses a *probability distribution* provided by the model to generate a new word in the sequence. C. Callison-Burch Decl. ¶¶ 33–36; J. Gratz Decl. ¶ 19; Ex. 18 at 22.

51. OpenAI conducts evaluations to test model safety and engages human "red teams" to test the model's resilience to attempts to circumvent its safety features. C. Callison-

Burch Decl. ¶ 107; J. Gratz Decl. ¶¶ 36, 56, 75; Ex. 35 at 33:25–34:12, 144:18–145:6, 146:2–4; Ex. 55 at 13:7–15; Ex. 74 at 172:16–20.

52. OpenAI trains its models to minimize potential bias from its training data. C. Callison-Burch Decl. ¶ 98, 105; J. Gratz Decl. ¶¶ 17, 30, 36, 59–64; Ex. 16 at OPCO_MDL_002071701; Ex. 29 at Resp. to Interrog. No. 16; Ex. 35 at 149:19–150:8; Ex. 58 at Supp. Resp. to Interrog. No. 16; Ex. 59 at 1, 9; Ex. 60 at 21–23; Ex. 61 at 1; Ex. 62 at OPCO_MDL_000999954; Ex. 63.

53. The Times has praised OpenAI's safety features. J. Gratz Decl. ¶ 42; Ex. 41.

54. OpenAI deploys models via one of two interfaces. J. Gratz Decl. ¶¶ 36, 65; Ex. 35 at 122:6–123:3; Ex. 64 at 116:6–10.

55. ChatGPT is a consumer-facing interface, available for free via website or mobile app. J. Gratz Decl. ¶¶ 66–67; Exs. 65, 66.

56. Users submit "prompts" to ChatGPT, and ChatGPT generates text in response. C. Callison-Burch Decl. ¶ 101.

57. OpenAI's application programming interface ("API") is used by software developers to power new applications or features in their products. C. Callison-Burch Decl. ¶ 100; J. Gratz Decl. ¶¶ 68–70; Ex. 67 at OPCO_MDL_001001906; Ex. 68 at 14:24–15:4; Ex. 69 at 99:21–100:6.

IV. How OpenAI Prevents Regurgitation

58. The purpose of pretraining is to teach the model generalizable rules of language, including facts about the world. J. Gratz Decl. ¶¶ 21, 23, 30; Ex. 20 at 51:8–13; Ex. 22 at 24:18–22, 61:14–15; Ex. 29 at Resp. to Interrog. No. 16.

59. The models learn these features by analyzing existing texts and identifying the *commonalities* between them. C. Callison-Burch Decl. ¶¶ 37–39; J. Gratz Decl. ¶¶ 21, 36; Ex. 20 at 49:4–24, 104:15–17; Ex. 35 at 76:25–77:15, 80:4–13.

60. A model that encodes the commonalities between texts will encode patterns corresponding to linguistic rules or common facts. C. Callison-Burch Decl. ¶ 39; J. Gratz Decl. ¶¶ 21, 71; Ex. 20 at 50:14–19; Ex. 70 at 51:2–23.

61. A model may encode common idioms or turns of phrase or even public-domain texts. C. Callison-Burch Decl. ¶ 39; J. Gratz Decl. ¶ 21; Ex. 20 at 50:25–51:18.

62. Researchers sometimes refer to this encoding of text sequences as *memorization*. C. Callison-Burch Decl. ¶ 44; J. Gratz Decl. ¶¶ 21, 37; Ex. 20 at 182:20–183:23; Ex. 36 at 352:13–23.

63. If an LLM encounters the same oft-repeated book passage or news article excerpt hundreds of times from different sources during training, it might discern that the sequence of words therein constitutes a common pattern. C. Callison-Burch Decl. ¶ 39.

64. *Regurgitation* refers to the phenomenon where a model given enough of a book passage or news article as context and tasked with completing it, may do so by generating the next portion verbatim. C. Callison-Burch Decl. ¶ 44; J. Gratz Decl. ¶ 23; Ex. 22 at 63:10–22.

65. OpenAI has invested significant time and resources in designing models that do not regurgitate copyrighted material. C. Callison-Burch Decl. ¶¶ 79, 108–15; J. Gratz Decl. ¶¶ 17, 29, 37, 71, 73; Ex. 16 at OPCO_MDL_002071701, OPCO_MDL_002071709; Ex. 28 at OPCO_MDL_000000446, OPCO_MDL_000000481; Ex. 36 at 356:6–13; Ex. 70 at 17:8–14; Ex. 72 at OPCO_MDL_002138931.

66. OpenAI devised and implemented techniques to prevent regurgitation. C.

Callison-Burch Decl. ¶¶ 79, 108–15; J. Gratz Decl. ¶¶ 17, 29, 37, 71, 73, 75; Ex. 16; Ex. 28 at OPCO_MDL_000000446, OPCO_MDL_000000481; Ex. 36 at 356:6–13; Ex. 70 at 17:8–14; Ex. 72 at OPCO_MDL_002138931; Ex. 74 at 20:5–13.

67. One of the most significant drivers of memorization is duplication in the training data. C. Callison-Burch Decl. ¶¶ 79, 109; J. Gratz Decl. ¶¶ 23, 36, 76; Ex. 22 at 61:14–62:1; Ex. 35 at 107:20–108:15; Ex. 75 at 8424.

68. When a text sequence occurs multiple times in the training set, the model may memorize it, rather than simply learning the broad patterns the sequence shares with other text. C. Callison-Burch Decl. ¶ 79; J. Gratz Decl. ¶¶ 23, 36; Ex. 22 at 61:4–62:1; Ex. 35 at 107:20–108:15.

69. De-duplicating training data is one way to minimize problematic memorization and, by extension, regurgitation. C. Callison-Burch Decl. ¶¶ 79, 103, 109; J. Gratz Decl. ¶¶ 23, 36, 76; Ex. 22 at 61:4–62:1; Ex. 35 at 208:13–209:9; Ex. 75 at 8424.

70. OpenAI began de-duplicating its training data at least as early as 2019 via its work on the GPT-2 model. J. Gratz Decl. ¶ 8; Ex. 7 at OPCO_MDL_000200506.

71. For GPT-3, OpenAI developed, and has since used, "fuzzy deduplication" techniques to capture not only exact but also near-verbatim matches. C. Callison-Burch Decl. ¶ 79; J. Gratz Decl. ¶ 29; Ex. 28 at OPCO_MDL_000000446.

72. For GPT-4, OpenAI implemented, and has since used, a technique to identify repeated phrases in the training data (such as longer passages frequently quoted in book reviews) to avoid training disproportionally on the same text sequence. C. Callison-Burch Decl. ¶ 109; J. Gratz Decl. ¶¶ 17, 38, 71, 77; Ex. 16; Ex. 37 at 54:25–55:17; Ex. 70 at 17:8–14; Ex. 76 at 158:7–159:12.

73. OpenAI designs its training algorithm to avoid memorization that might lead to unwanted regurgitation. C. Callison-Burch Decl. ¶¶ 42, 45, 103, 109; J. Gratz Decl. ¶¶ 23, 26; Ex. 22 at 316:6–24; Ex. 25 at 71:20–72:5.

74. OpenAI implements mechanisms to avoid training on the same training data excerpts until all of the data within an entire training set have been depleted. J. Gratz Decl. ¶¶ 29, 74; Ex. 28 at OPCO_MDL_000000481; Ex. 73 at 195:4–9.

75. OpenAI implements *regularization* techniques like drop-out and weight decay, which encourage the model to focus on broad, generalizable patterns and to ignore or deemphasize idiosyncratic features of specific training examples. C. Callison-Burch Decl. ¶ 42; J. Gratz Decl. ¶¶ 7, 29; Ex. 6 at OPCO_MDL_000000386; Ex. 28 at OPCO_MDL_000000481.

76. Since GPT-4, OpenAI conducts copyright-specific post-training to train models to avoid regurgitating any memorized training data, including techniques that train the models to recognize and refuse requests for copyrighted content. C. Callison-Burch Decl. ¶¶ 112–13; J. Gratz Decl. ¶¶ 23, 65, 75, 78; Ex. 22 at 316:13–317:15; Ex. 64 at 101:13–24; Ex. 74 at 20:5–13, 30:16–31:4; Ex. 77 at 31:6–17.

77. OpenAI trained the model during post-training to identify and refuse requests to output copyrighted content. C. Callison-Burch Decl. ¶ 110; J. Gratz Decl. ¶¶ 36, 72, 75; Ex. 35 at 152:14–153:8; Ex. 71 at 43:17–44:3; Ex. 74 at 30:16–31:4, 31:8–12.

78. OpenAI continued to refine post-training techniques throughout 2023 and 2024. C. Callison-Burch Decl. ¶¶ 95, 109–115; J. Gratz Decl. ¶¶ 75, 81–82; Ex. 74 at 20:5–13; Exs. 80–81.

79. OpenAI designed and implemented a filter to assess and screen model outputs as they are generated. C. Callison-Burch Decl. ¶ 111; J. Gratz Decl. ¶¶ 75, 77; Ex. 74 at 229:2–11;

Ex. 76 at 166:12–19.

80. The filter OpenAI employed to screen model outputs for potential regurgitation used a probabilistic data structure called a *bloom filter* to evaluate each model output against a subset of training data sequences. C. Callison-Burch Decl. ¶ 111; J. Gratz Decl. ¶¶ 83, 86–89; Ex. 82 at 30:3–4, 190:12–23; Ex. 85 at 265:17–21, 301:11–302:4; Ex. 86 at 43:1–14; Ex. 87 at 143:22–144:10; Ex. 88 at 72:2–15.

81. OpenAI ran a series of internal tests to identify text sequences that were more susceptible to regurgitation and used those sequences to populate the bloom filter's index. C. Callison-Burch Decl. ¶ 111; J. Gratz Decl. ¶¶ 86–87; Ex. 85 at 265:17–21, 277:11–18, 301:11–302:4; Ex. 86 at 82:2–12.

FACTS SPECIFIC TO THE CONSOLIDATED CLASS ACTION

82. OpenAI's expert searched for regurgitation of the Asserted Works in a sample of 95 million outputs from the 20 million ChatGPT conversation logs that OpenAI produced in this case and found 14 instances in which Class Plaintiffs allege regurgitation of content from their books—*i.e.*, an alleged regurgitation rate of 0.00007%. Declaration of Dr. Taylor Berg-Kirkpatrick ("T. Berg-Kirkpatrick Decl.") ¶ 30.

83. Of these 14 instances, all but two were 32 words or less (the two outliers were 67 and 122 words). T. Berg-Kirkpatrick Decl. ¶ 29, Tables 2, 3.

84. Of these 14 instances, none represented more than 0.017% of a Class Asserted Work. T. Berg-Kirkpatrick Decl. ¶ 29, Tables 2, 3.

85. Class Plaintiffs' expert, Dr. Shawn Shan, entered more than 5.3 million prompts to force regurgitations using prompts identifying the book, characters, plot, or requested continuation. J. Gratz Decl. ¶¶ 94–95; Ex. 93 at 138:19–140:12, 216:22–24; Ex. 94 ¶ 17.

86. The longest contiguous text Dr. Shan could generate was a 1,899-word excerpt from "A Game of Thrones." J. Gratz Decl. ¶ 95; Ex. 94 ¶ 22.

87. "A Game of Thrones" has 305,507 words. T. Berg-Kirkpatrick Decl. ¶ 70.

88. Thus, the longest contiguous text that Dr. Shan could generate was 0.62% of A Game of Thrones. T. Berg-Kirkpatrick Decl. ¶ 70.

89. The contiguous text that Dr. Shan identified is available on the Internet, including in the Amazon Kindle sample for "A Game of Thrones." T. Berg-Kirkpatrick Decl. ¶ 71.

90. OpenAI's expert found that based on the sample of 20 million conversation logs, verbatim excerpts of named Class Plaintiffs' Asserted Works appear in less than 0.00001% of ChatGPT outputs. T. Berg-Kirkpatrick Decl. ¶ 25, Tables 2, 3.

91. Dr. Shan conducted a study of conversation logs (using a different methodology and a different, but partially overlapping, sample) where he "identified reproductions of 30 or more near-verbatim words . . . from 7 Plaintiff works across 31 conversations and 33 chat responses." 31 instances out of 108 million ChatGPT conversation logs is a regurgitation rate of 0.00003%. J. Gratz Decl. ¶ 96; Ex. 95 ¶¶ 37, 154.

V. The Class Plaintiffs, the Class Asserted Works, and Class Allegations in This Case

92. The Class Plaintiffs include international bestselling authors like David Baldacci, John Grisham, George R.R. Martin, and Jodi Picoult, and several other prominent authors. ECF 183 ¶¶ 171, 182, 185, 196, 207, 218, 224, 229, 240, 250, 262, 273, 278, 283.

93. Collectively, Plaintiffs have won dozens of awards and sold hundreds of millions of books. ECF 183 ¶¶ 171, 182, 185, 196, 207, 218, 224, 250, 262, 273.

94. Several of the Class Plaintiffs have made millions of dollars or more over the course of their careers. J. Gratz Decl. ¶¶ 97–101; Ex. 96 at 28:14–19; Ex. 97 at 86:4–7; Ex. 98 at 174:6–8, 204:19–205:3; Ex. 99 at 98:8–99:5; Ex. 100 at 187:5–13.

95. Class Plaintiffs wrote their books to be read and enjoyed by a human audience. J. Gratz Decl. ¶¶ 99–107; Ex. 98 at 26:2–15, 129:4–10; Ex. 99 at 19:20–24; Ex. 100 at 187:5–13; Ex. 101 at 260:12–17; Ex. 102 at 17:4–8; Ex. 103 at FICTION-0021308; Ex. 104 at 101:16–18; Ex. 105 at 22:13–16; Ex. 106 at 162:20–21.

96. Class Plaintiffs have continued to write and sell new books after ChatGPT was released. J. Gratz Decl. ¶¶ 97, 99, 101, 103, 105, 107–116; Ex. 96 at 73:17–19; Ex. 98 at 171:21–25; Ex. 100 at 166:19–167:4; Ex. 102 at 17:16–18; Ex. 104 at 15:5; Ex. 106 at 30:14–20; Ex. 107 at Resp. to RFA No. 44; Ex. 108 at RFA No. 44; Ex. 109 at Resp. to RFA No. 44; Ex. 110 at Resp. to RFA No. 44; Ex. 111 at Resp. to RFA No. 50; Ex. 112 at 21:4–6, 22:8–11; Ex. 113 at 209:1–3; Ex. 114 at 25:19–21; Ex. 115 at 46:1–2, 112:23–113:1.

97. The Class Asserted Works are "trade books," meaning they are sold via bookstores and other standard channels for general readership. J. Gratz Decl. ¶¶ 99–100, 102, 105–107, 117; Ex. 98 at 26:2–15, 129:4–10; Ex. 99 at 19:20–24; Ex. 101 at 207:21–208:11; Ex. 104 at 101:16–18; Ex. 105 at 22:13–16; Ex. 106 at 162:20–21; Ex. 116 at 52:24–53:4.

98. The Asserted Works span different genres, including mystery, fantasy, and romance. ECF 183 ¶¶ 250, 282–83; J. Gratz Decl. ¶¶ 97, 99, 107; Ex. 96 at 30:15–17; Ex. 98 at 79:9–12; Ex. 106 at 20:11–17.

99. Several of the Asserted Works are bestsellers. ECF 183 ¶ 273; J. Gratz Decl. ¶¶ 97, 99; Ex. 96 at 19:25–20:1; Ex. 98 at 21:8–13.

100. The Asserted Works are published (some many years ago). J. Gratz Decl. ¶¶ 105, 118–130; Ex. 104 at 95:7–10; Ex. 117 at Resp. to RFA No. 60; Ex. 118 at Resp. to RFA No. 60; Ex. 119 at Resp. to RFA No. 59; Ex. 120 at Resp. to RFA No. 59; Ex. 121 at Resp. to RFA No. 60; Ex. 122 at Resp. to RFA No. 60; Ex. 123 at Resp. to RFA No. 60; Ex. 124 at Resp. to RFA No. 59; Ex. 125 at Resp. to RFA No. 60; Ex. 126 at Resp. to RFA No. 60; Ex. 127 at Resp. to RFA No. 60; Ex. 128 at RFA General Obj. No. 10; Ex. 129 at RFA General Obj. No. 10.

101. The average purchase price of a new book is roughly $30. J. Gratz Decl. ¶ 117; Ex. 116 at 79:18–80:5.

102. Excerpts of the Asserted Works are available on public websites. J. Gratz Decl. ¶¶ 101, 107, 131; Ex. 100 at 175:6–13; Ex. 106 at 165:24–166:2; Ex. 130 at Supp. Resp. to Interrog. No. 14.

103. Book reviews and summaries of the Asserted Works are available for free online from sources like Wikipedia and Goodreads. J. Gratz Decl. ¶¶ 97–99, 105; Ex. 96 at 117:19–118:5; Ex. 97 at 188:25–189:17, 203:3–20; Ex. 98 at 259:5–15, 259:25–260:7; Ex. 104 at 135:9–14.

104. To the extent that OpenAI downloaded any of Books Plaintiffs' Asserted Works, it was for training, not for other purposes. J. Gratz Decl. ¶ 27; Ex. 26 at 412:3–17.

105. Class Plaintiffs admit that when ChatGPT is prompted to return quotations of text from copyrighted books, ChatGPT generally does not do so. ECF 183 ¶¶ 105–107; J. Gratz Decl. ¶¶ 97, 99–101, 103, 107, 114, 117, 129, 132; Ex. 96 at 152:20–153:1; Ex. 98 at 246:1–7; Ex. 99 at 74:21–24; Ex. 100 at 26:7–10, 118:12–18; Ex. 102 at 65:13–16; Ex. 106 at 176:5–11; Ex. 113 at 149:24–150:14; Ex. 115 at 79:22–25; Ex. 128 at Resp. to RFA No. 44; Ex. 131 at Resp. to RFA Nos. 44, 45.

106. John Grisham testified that a four-paragraph summary of *The Chamber* would not serve as a substitute for a consumer reading the book. J. Gratz Decl. ¶ 97; Ex. 96 at 84:17–22.

107. Mr. Grisham testified that he is not "aware of anyone choosing [a] short summary of 'The Chamber' instead of purchasing or reading [his] novel." J. Gratz Decl. ¶ 97; Ex. 96 at 84:23–85:2.

108. Jonathan Franzen was "not bothered by" summaries of his books on other websites. J. Gratz Decl. ¶ 106; Ex. 105 at 68:6–18.

109. Mr. Martin's personal website links to other websites that contain chapter-by-chapter summaries of some of his books. J. Gratz Decl. ¶ 98; Ex. 97 at 207:13–24.

110. Christopher Golden admitted that he is "currently unaware of any specific ChatGPT output that has caused [him] harm and/or damages, or any ChatGPT output that serves as a direct or indirect substitute for the Asserted Works." J. Gratz Decl. ¶¶ 133–134; Ex. 132 at Resp. to Interrog. No. 22; Ex. 133 at Resp. to Interrog. No. 22.

111. Andrew Greer admitted that he is "currently unaware of any specific ChatGPT output that has caused [him] harm and/or damages, or any ChatGPT output that serves as a direct or indirect substitute for the Asserted Works." J. Gratz Decl. ¶¶ 135–136; Ex. 134 at Resp. to Interrog. No. 22; Ex. 135 at Resp. to Interrog. No. 22.

112. David Henry Hwang admitted that he is "currently unaware of any specific ChatGPT output that has caused [him] harm and/or damages, or any ChatGPT output that serves as a direct or indirect substitute for the Asserted Works." J. Gratz Decl. ¶¶ 137–138; Ex. 136 at Resp. to Interrog. No. 22; Ex. 137 at Resp. to Interrog. No. 22.

113. Class Plaintiffs' damages expert, Mr. Michael Lasinski, testified that he "do[es] not have the data to show that one specific customer switched from" reading the Asserted Works to a ChatGPT output. J. Gratz Decl. ¶ 117; Ex. 116 at 236:13–23.

114. Mr. Golden is not "aware of any instance where [he] lost a sale of Ararat because ChatGPT produced a copy of it" or where he "lost a sale of Ararat because of something else ChatGPT did." J. Gratz Decl. ¶ 114; Ex. 113 at 300:7–14.

115. Mr. Branch could not point to any evidence which "suggests that [he has] any lost sales beyond any copies that were allegedly used to train the models." J. Gratz Decl. ¶ 103; Ex. 102 at 73:11–74:1.

116. John Grisham testified that he is not aware of "any instance where [he] lost a sale of a book in competition with a ChatGPT-generated book" and "can't identify a single reader who chose not to purchase one of [his] books because they had a ChatGPT generated book instead." J. Gratz Decl. ¶ 97; Ex. 96 at 175:18–176:10.

117. George R.R. Martin testified that he was not aware of "ever having lost a sale of a book because of OpenAI." J. Gratz Decl. ¶ 98; Ex. 97 at 216:24–217:5.

118. Jonathan Franzen, when asked if he "believe[d] that ChatGPT is an existing threat to your livelihood – as an author by basically replacing your work," responded, "No." J. Gratz Decl. ¶ 106; Ex. 105 at 46:5–8.

119. Jonathan Franzen testified that it "seems very unlikely" that someone who intended to buy *The Corrections* would choose to read a two-page summary of the book instead. J. Gratz Decl. ¶ 106; Ex. 105 at 68:24–69:17.

120. Multiple Class Plaintiffs admit that their own books were inspired by other books they have read. J. Gratz Decl. ¶¶ 98–101; Ex. 97 at 165:19–25; Ex. 98 at 78:13–21; Ex. 99 at 12:16–19, 89:9–25; Ex. 100 at 19:6–11.

121. None of the Class Plaintiffs have ever received a payment for selling or licensing their books to train an LLM. J. Gratz Decl. ¶¶ 97, 101, 103, 105, 107, 113–114, 117, 140; Ex. 96 at 184:13–24; Ex. 100 at 86:8–11; Ex. 102 at 108:21–24; Ex. 104 at 112:24–113:3; Ex. 106 at 85:22–86:9; Ex. 112 at 197:9–12; Ex. 113 at 103:7–11; Ex. 116 at 215:16–21; Ex. 139 at 169:2–17.

122. Mr. Golden did not contemplate receiving payment for licensing his book *Ararat* at the time he completed it. J. Gratz Decl. ¶ 114; Ex. 113 at 305:21–24.

123. Microsoft's expert, Dr. Catherine Tucker, prepared an exhibit reflecting the first publication dates of the Class Plaintiffs' Asserted Works. J. Gratz Decl. ¶ 90; Ex. 89 at E-1–E-5, Exhibit E.

124. Prior to filing this lawsuit, Class Plaintiffs had not been approached to license their books for LLM training. J. Gratz Decl. ¶¶ 97, 99, 101, 103, 105, 107, 109, 113–114, 129; Ex. 96 at 184:13–24; Ex. 98 at 214:16–215:2; Ex. 100 at 165:10–15; Ex. 102 at 108:25–109:6; Ex. 104 at 112:24–113:3; Ex. 106 at 157:16–18, 211:13–16; Ex. 108 at Resp. to RFA No. 47; Ex. 112 at 195:19–23; Ex. 113 at 306:6–11; Ex. 128 at Resp. to RFA No. 37.

125. OpenAI's economic expert, Dr. Gregory Leonard, analyzed sales data for the Asserted Works. Leonard Decl. ¶ 9.

126. Dr. Leonard did not find a systematic decrease in sales trends for the Asserted Works after the launch of OpenAI's products. Leonard Decl. ¶¶ 11, 12, 16, 17.

127. Dr. Leonard analyzed pricing data for certain Asserted Works that met his criteria and found that OpenAI's products had no statistically significant impact. Leonard Decl. ¶ 32.

128. Dr. Leonard constructed a control group of books that were not used to train OpenAI's LLMs and compared sales trends for those books to the sales trends for over 100 of the Asserted Works. Leonard Decl. ¶ 18.

129. Dr. Leonard's control-group study found no statistically significant difference in sales trends between the Asserted Works he analyzed and the books in his control group. Leonard Decl. ¶¶ 26, 27.

130. Class Plaintiffs' expert, Mr. Michael Lasinski, did not conduct an empirical analysis of his own to demonstrate that Plaintiffs lost sales of the Asserted Works due to OpenAI's alleged use of the Asserted Works. Leonard Decl. ¶ 34; J. Gratz Decl. ¶ 117; Ex. 116 at 82:13–15, 236:13–23.

131. Mr. Lasinski stated that he did not have the "data necessary to identify a specific, just-prompted-by-ChatGPT book that would have caused another author to lose a sale." J. Gratz Decl. ¶ 117; Ex. 116 at 99:21–100:17.

132. Mary Rasenberger, CEO of the Authors Guild, testified that "[n]obody is disputing that ChatGPT is a very useful tool," and she "understand[s] from authors that they find it useful." J. Gratz Decl. ¶ 148; Ex. 147 at 51:23–52:10.

133. Umair Kazi, the Director of Policy at the Authors Guild, testified that at least some authors "clearly find [AI tools] useful"; that there is a community of authors who use AI in their writing process; and "that authors [could] generally benefit from AI assistance" with proofreading and grammar. J. Gratz Decl. ¶ 149; Ex. 148 at 91:1–12, 97:23–98:20, 106:13–20, 112:12–25, 122:16–17, 146:21–147:11, 148:7–149:1.

134. Multiple Class Plaintiffs have used ChatGPT for research. J. Gratz Decl. ¶¶ 105, 148, 150; Ex. 104 at 146:7–11; Ex. 147 at 45:13–46:7, 51:8–11; Ex. 149 at 74:13–75:24.

135. ChatGPT is "commonly used to contribute to and edit the creation of books, which can allow novice writers or individuals who may never attempt to write a book to compete in the book marketplace, lowering the barriers to entry in the traditional book market." J. Gratz Decl. ¶ 151; Ex. 150 ¶368.

136. A book contains about 70,000 to 100,000 tokens. J. Gratz Decl. ¶ 4; Ex. 3 at 85:18–20.

FACTS SPECIFIC TO THE NEWS CASES

137. The News Plaintiffs are news or news-related organizations that include: (1) The New York Times Company ("The Times"), (2) eight regional newspapers ("The Daily News Plaintiffs" or "DNP"),[1] (3) digital media conglomerate Ziff Davis, Inc. ("Ziff Davis"), (4) The Center for Investigative Reporting ("CIR"), and (5) The Intercept Media, Inc. ("The Intercept"). *See The New York Times Co. v. Microsoft Co. et al.*, 23-cv-11195, Dkt. 73; *Daily News LP v. Microsoft Corp. et al.*, 24-cv-03285, Dkt. 777; *Ziff Davis, Inc. v. OpenAI, Inc. et al.*, 25-md-03143, Dkt. 1676; *The Center for Investigative Reporting, Inc. v. OpenAI, Inc. et al.*, 24-cv-04872, Dkt. 88; *The Intercept Media, Inc. v. OpenAI, Inc. et al.*, 24-cv-01515, Dkt. 155.

VI. Technical background relevant to pretraining

138. OpenAI sought to create "generalist" LLMs that could "seamlessly mix together or switch between many tasks and skills" with "fluidity and generality." J. Gratz Decl. ¶ 29; Ex.

[1] The Daily News Plaintiffs are Chicago Tribune Company, LLC (the Chicago Tribune), DP Media Network, LLC (the Denver Post), Daily News, L.P. (the New York Daily News), Northwest Publications, LLC (the Pioneer Press), ORB Publishing, LLC (the Orange County Register), Orlando Sentinel Communications Company, LLC (the Orlando Sentinel), San Jose Mercury-News, LLC (the San Jose Mercury News), and Sun-Sentinel Company, LLC (the Sun Sentinel).

28 at OPCO_MDL_000000442.

139. OpenAI used a filtered version of the publicly available index of webpages created by Common Crawl. C. Callison-Burch Decl. ¶¶ 16, 62; J. Gratz Decl. ¶ 26; Ex. 25 at 211:19–212:3.

140. Common Crawl included news content. J. Gratz Decl. ¶ 27; Ex. 26 at 263:1–8.

141. OpenAI also built new datasets that contained text from a wide variety of other websites, some of which included news articles. J. Gratz Decl. ¶ 8; Ex. 7 at OPCO_MDL_000200506; Ex. 54.

VII. Technical background relevant to Browse

142. When ChatGPT launched in November 2022, it lacked the ability to help users find current factual information on the Internet in real-time. J. Gratz Decl. ¶¶ 79, 155; Ex. 78 at 39:15–24; Ex. 154 at OPCO_MDL_001693113.

143. In March 2023, OpenAI introduced an experimental web-browsing feature for ChatGPT. J. Gratz Decl. ¶¶ 155–156; Ex. 154 at OPCO_MDL_001693118; Ex. 155 at 15:25–16:7.

144. OpenAI's web-browsing feature has been known at various times as "Browse with Bing," "Browse," or "ChatGPT Search."[2] Declaration of Dr. Nick Feamster ("N. Feamster Decl.") ¶ 10 n.1; J. Gratz Decl. ¶ 157; Ex. 156.

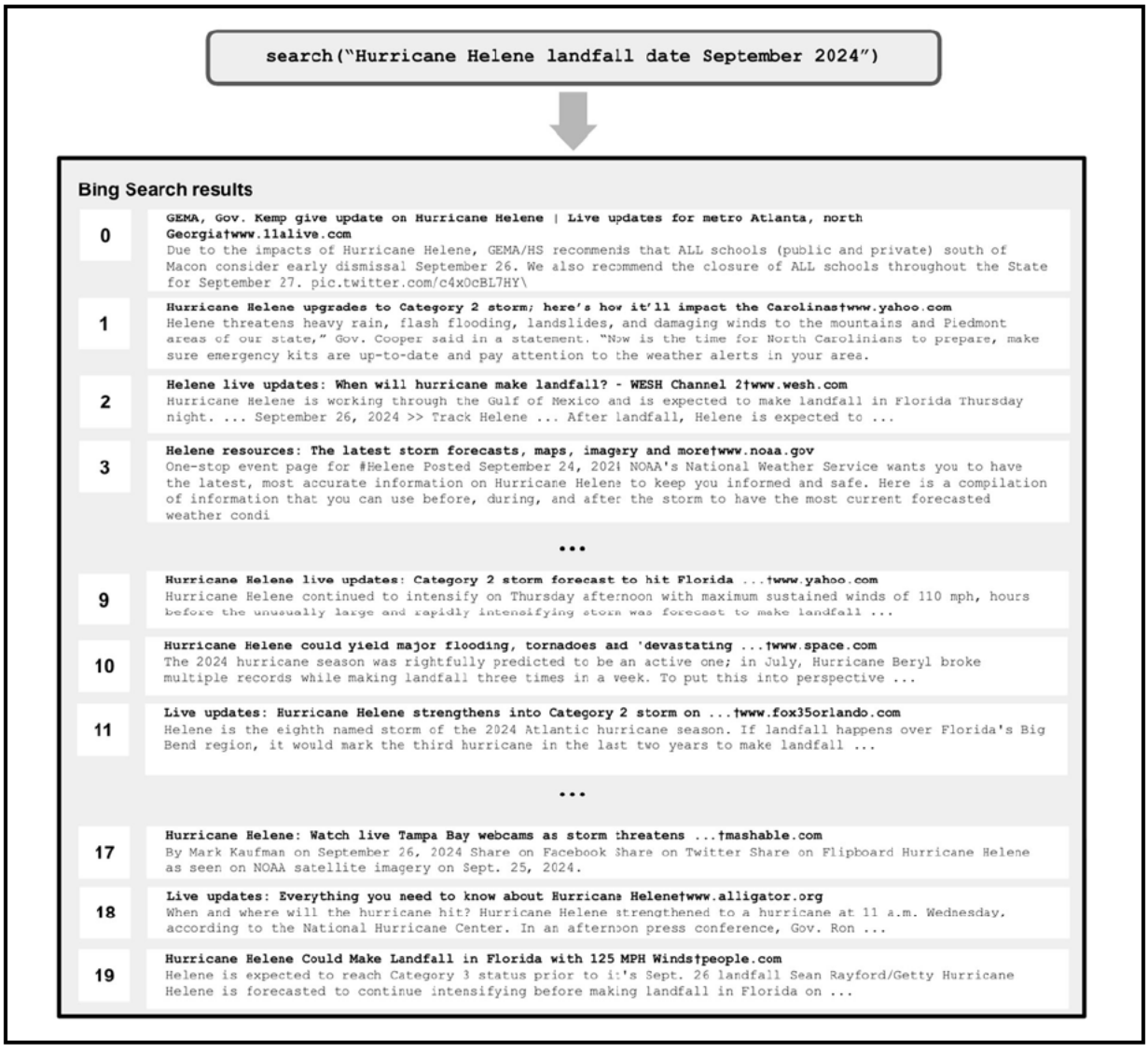
145. When a user asked ChatGPT a question, ChatGPT's Browse feature would convert the user prompt into a search query, adding contextual information like the date or the location. N. Feamster Decl. ¶¶ 12–17; J. Gratz Decl. ¶¶ 57, 157, 159; Ex. 56 ¶ 104; Ex. 156; Ex. 158.

[2] For ease of reference, OpenAI's web-browsing service is referred to throughout as "Browse."

146. ChatGPT would send that query to Microsoft Bing, which maintains an index of billions of web pages. N. Feamster Decl. ¶¶ 12–17; J. Gratz Decl. ¶ 159; Ex. 158.

147. Website operators can opt out of having their websites indexed by Bing. J. Gratz Decl. ¶¶ 160–161; Ex. 159; Ex. 160.

148. Bing returned a list of results with links to relevant webpages, along with short "snippets" of text. An example of this step is shown in the figure below. N. Feamster Decl. ¶¶ 12–17; J. Gratz Decl. ¶¶ 57, 157, 159, 163; Ex. 56 ¶ 105; Ex. 156; Ex. 158; Ex. 162 at 78:17–24.



149. The Bing-returned snippets in this step are not visible to the user. N. Feamster Decl. ¶ 15.

150. ChatGPT analyzed the search results returned by Bing to determine which webpages might contain relevant information and then—subject to opt-out limitations described below—sent an automated request to websites to obtain the content of those webpages. N. Feamster Decl. ¶ 16, Fig. 3.

151. In some instances, ChatGPT used the search results returned by Bing to inform its response to the user without accessing the associated webpages. N. Feamster Decl. ¶ 54.

152. OpenAI's expert Dr. Feamster found 149 such instances in the 11,015 conversations News Plaintiffs identified that invoke Browse. N. Feamster Decl. ¶¶ 54, 68–69.

153. When a webpage returned its content to ChatGPT, Browse identified a limited excerpt of "relevant information" from each source to help inform its response to the user. N. Feamster Decl. ¶¶ 16–17; J. Gratz Decl. ¶ 162; Ex. 161 ¶ 48.

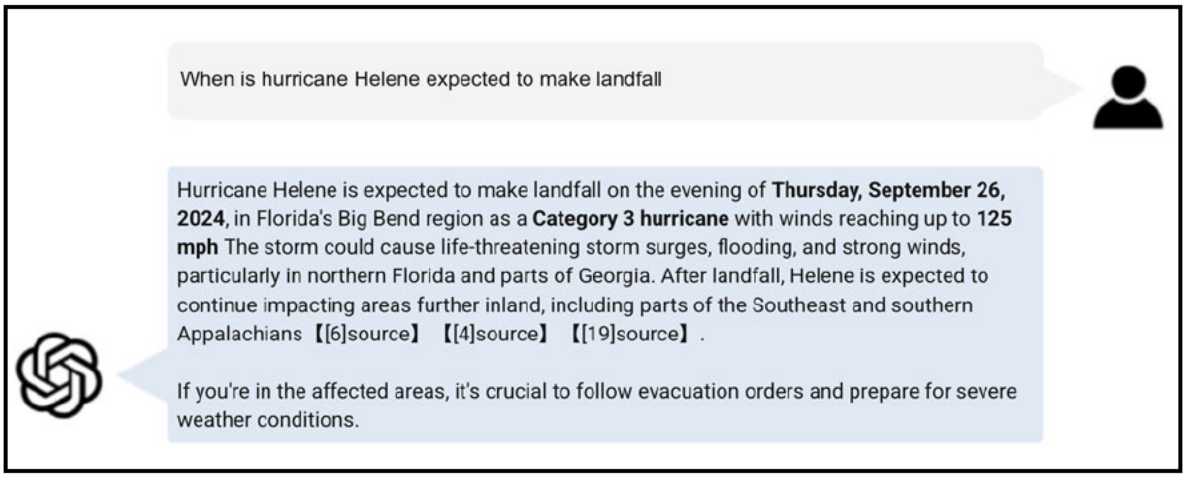
154. After receiving the content, ChatGPT would extract the relevant factual information it received from multiple webpages and synthesize it into a response to the user. N. Feamster Decl. ¶ 12, Ex. 1, ¶¶ 16–19; J. Gratz Decl. ¶¶ 19, 162, 163, 164; Ex. 18 at 23–25; Ex. 161 ¶¶ 48, 51; Ex. 162 at 151:14–19; Ex. 163 ¶ 47.

155. OpenAI designed Browse to derive and triangulate factual information from a variety of sources to generate a user response. N. Feamster Decl. ¶ 20.

156. OpenAI designed Browse not to display verbatim content from the webpages used to inform its response. [redacted] N. Feamster Decl. ¶¶ 18–20; J. Gratz Decl. ¶ 156; Ex. 155 at 18:15–21.

157. ChatGPT displayed links to the underlying referenced webpages or "sources" used to inform its response. T. Berg-Kirkpatrick Decl. ¶ 77; N. Feamster Decl. ¶ 17, Fig. 4; J. Gratz Decl. ¶¶ 156, 162, 164, 165; Ex. 155 at 127:14–17; Ex. 161 ¶¶ 337–338; Ex. 163 ¶ 206; Ex. 164, Apps. C, E.

158. The image below is a user chat from Sept 26, 2024. N. Feamster Decl. Fig. 4.



159. On July 3, 2023, OpenAI stated: "We've learned that the browsing beta can occasionally display content in ways we don't want, e.g. if a user specifically asks for a URL's full text, it may inadvertently fulfill this request." J. Gratz Decl. ¶ 155; Ex. 154 at OPCO_MDL_001693116.

160. OpenAI "temporarily disabl[ed] Browse" while it fixed this issue. J. Gratz Decl. ¶ 155; Ex. 154 at OPCO_MDL_001693116.

161. OpenAI "roll[ed] [Browse] back out to Plus users on September 27, 2023." N. Feamster Decl. Fig. 5; J. Gratz Decl. ¶ 155; Ex. 154 at OPCO_MDL_001693113.

162. OpenAI used post-training and system-prompt measures for Browse to prevent the model from regurgitating significant portions of content from individual webpages. N. Feamster Decl. ¶ 19; J. Gratz Decl. ¶¶ 156, 166; Ex. 155 at 18:15–21; Ex. 165.

163. OpenAI's expert Dr. Feamster concluded that, in over 96% of the instances in

which News Plaintiffs claim that Browse made an infringing copy of their work, OpenAI only received a short snippet of webpage content as part of a Bing-returned search result during the first step in the Browse process. N. Feamster Decl. ¶¶ 61–62, Ex. 1, Examples 3–4; J. Gratz Decl. ¶ 158; Ex. 157 ¶¶ 224, 227.

164. Those snippets were used to determine which search results were relevant, and OpenAI did not send a request for content to the News Plaintiff's webpage. N. Feamster Decl. ¶¶ 15–16, 61.

165. OpenAI's expert Dr. Feamster confirmed that, for every Bing search snippet containing News Plaintiffs' content that appeared in the ChatGPT conversation records, the associated website permitted access to Bingbot—the web crawler Microsoft uses for its search index—at the time the snippet was returned. N. Feamster Decl. ¶¶ 63–67.

VIII. Technical background related to robots.txt

166. OpenAI designed Browse to comply with the Robots Exclusion Protocol, an industry-standard Internet protocol which is implemented via robots.txt files. N. Feamster Decl. ¶¶ 35–39, 48–51; J. Gratz Decl. ¶¶ 54, 156, 167; Ex. 53 ¶ 94; Ex. 155 at 48:3–5, 63:3–9, 76:24–77:4, 95:1–3, 96:9–14; Ex. 166 at 48:5–7, 70:16–22.

167. The Robots Exclusion Protocol allows website owners to express their preference to "opt-out" of receiving automated requests for webpage content from specific "user agents" commonly referred to as "bots." N. Feamster Decl. ¶¶ 36, 39–40; J. Gratz Decl. ¶¶ 19, 162, 168–170; Ex. 18 at 11; Ex. 161 ¶¶ 86, 94; Ex. 167; Ex. 168 at 221:1–16; Ex. 169 at 89:2–4.

168. Before sending a request to retrieve a webpage's content, OpenAI's system performs a robots.txt check to determine whether the webpage expresses preferences regarding access by OpenAI's bots. N. Feamster Decl. ¶¶ 50–51.

169. If the applicable robots.txt file directs OpenAI's user agents ChatGPT-User or OAI-SearchBot not to access a webpage, the system will not send a request to that webpage. N. Feamster Decl. ¶¶ 50–51.

170. For decades, search engines like Google have built web indexes by crawling webpages on the Internet. N. Feamster Decl. ¶¶ 30–32, 34; J. Gratz Decl. ¶ 171; Ex. 170.

171. Other products and services—including e-commerce platforms, market research tools, fraud monitoring services, financial services platforms, government agencies, accessibility tools, and academic researchers—crawl or scrape the Internet to collect information. N. Feamster Decl. ¶ 32; J. Gratz Decl. ¶¶ 172–175; Exs. 171–174.

172. The Times and Chicago Tribune have publicly released web scrapers of their own. N. Feamster Decl. ¶ 33; J. Gratz Decl. ¶¶ 176–178; Exs. 175–177.

173. The Times and the corporate owners of the Daily News Plaintiffs filed an amicus brief with the Supreme Court in which they stated that "web scraping" is an "increasingly common data-journalism technique that relies on automation to pull large amounts of information from websites." J. Gratz Decl. ¶ 179; Ex. 178 at 4.

174. OpenAI's expert Dr. Feamster analyzed 20 million ChatGPT conversation logs produced in this case and found that in every instance in which Browse requested and received content from one of the News Plaintiffs' websites, the relevant website's robots.txt file was not configured to "disallow" OpenAI's Browse user-agent. N. Feamster Decl. ¶¶ 50–51, 54–59.

175. Two of Books' Plaintiffs' experts, Dr. Ben Zhao and Dr. Shawn Shan, published an article in 2024 stating that ChatGPT's "built-in crawler respects the robots.txt file." J. Gratz Decl. ¶ 180; Ex. 179.

176. OpenAI has set up distinct user-agents so that publishers can communicate

preferences for how OpenAI's automated systems may interact with their content for different types of uses. N. Feamster Decl. ¶¶ 48–49; J. Gratz Decl. ¶¶ 45, 156, 162, 164, 181, 182; Ex. 44; Ex. 155 at 96:9–16; Ex. 161 ¶ 238; Ex. 163 ¶ 156; Ex. 180; Ex. 181.

177. The Times, New York Daily News, Chicago Tribune, Mercury News, OC Register, CNET, and ZD Net were aware of and adopted the REP in 1999. N. Feamster Decl. ¶¶ 41–42, Ex. 2, App'x F.

178. The Sun Sentinel, Pioneer Press, Orlando Sentinel, Mother Jones, and Baby Center were aware of and adopted the REP in 2001. N. Feamster Decl. ¶¶ 41, 43, Ex. 2, App'x F.

179. Other News Plaintiffs' domains including The Intercept, Lifehacker, Mashable, PC Mag, Diabetes Daily, Dice Breaker, and Everyday Health maintained robots.txt files before 2010. N. Feamster Decl. ¶¶ 41, 44, Ex. 2, App'x F.

180. CIR has maintained a robots.txt file for centerforinvestigativereporting.com since 2011 and for revealnews.org since 2015. N. Feamster Decl. ¶¶ 41, 45, Ex. 2, App'x F.

181. CIR admitted that it learned about robots.txt "[a] long time ago." J. Gratz Decl. ¶¶ 183; Ex. 182 at 168:4–18.

182. The majority of News Plaintiffs' robots.txt files have been continually updated and edited. N. Feamster Decl. ¶ 46, Ex. 2.

183. In its March 2023 blog post announcing Browse, OpenAI disclosed that its user-agent was "ChatGPT-User," and that "to adhere to the web's norms," it was "configured to honor websites' robots.txt files." J. Gratz Decl. ¶¶ 45, 182; Ex. 44 at OPCO_MDL_000000517; Ex. 181.

184. Websites that preferred to "opt out" of requests from "ChatGPT-User" could

insert the following text string into their existing robots.txt file:

> 'User-agent: ChatGPT-User'
> 'Disallow: /.'

N. Feamster Decl. ¶¶ 39–40, 48–51; J. Gratz Decl. ¶¶ 54, 187, 188; Ex. 53 ¶ 94; Ex. 186; Ex. 187.

185. OpenAI began using the user agent "OAI-SearchBot" for Browse in 2024. N. Feamster Decl. ¶¶ 48–49; J. Gratz Decl. ¶¶ 189, 190, 191; Ex. 188; Ex. 189 at OPCO_MDL_001695874; Ex. 190.

186. OpenAI's blog post announcing ChatGPT-User was widely reported, including by Ziff Davis and other news publishers. J. Gratz Decl. ¶¶ 192–197; Exs. 191–196.

187. On April 1, 2023, The Times employees discussed whether they should "block OpenAI's ChatGPT-User bot." J. Gratz Decl. ¶ 188; Ex. 187.

188. On April 7, 2023, The Times stated: "[W]e have not as of yet added a directive to robots.txt" to block ChatGPT-User but it "[m]ight be worth keeping an eye on if we decide to take any action there." J. Gratz Decl. ¶ 198; Ex. 197 at NYT_00582829.

189. On April 10, 2023, John Sechrist, VP of Corporate Product at Ziff Davis, informed Joey Fortuna, Chief Technology Officer at Ziff Davis, that Ziff Davis domains could "disallow plugins to access [a] site" by "add[ing] the ChatGPT-User to [the] site's robots.txt." J. Gratz Decl. ¶ 187; Ex. 186.

190. The Times did not disallow ChatGPT-User until February 27, 2024 and did not disallow OAI-SearchBot until July 25, 2024. J. Gratz Decl. ¶ 199; Ex. 198, Procas Ex. 5.

191. Ziff Davis did not disallow ChatGPT-User or OAI-SearchBot until December 2024. J. Gratz Decl. ¶¶ 164, 200; Ex. 163 ¶ 156; Ex. 199.

192. The Intercept did not disallow ChatGPT-User until January 18, 2024. As of

August 27, 2026, The Intercept's robots.txt file did not include a disallow for OAI-SearchBot. J. Gratz Decl. ¶¶ 202, 203, 204; Ex. 200; Ex. 201; Ex. 202 at 423:9–425:7; Ex. 203.

193. The Daily News Plaintiffs included a robots.txt disallow for ChatGPT-User by October 26, 2023 and for OAI-SearchBot by August 23, 2024. J. Gratz Decl. ¶¶ 25, 205; Ex. 24; Ex. 204.

194. As of February 3, 2026, CIR did not include the user agent ChatGPT-User or OAI-SearchBot in its robots.txt files for motherjones.com or revealnews.org. These websites included a robots.txt disallow for ChatGPT-User and OAI-SearchBot by August 2026. J. Gratz Decl. ¶¶ 206, 207; Ex. 205; Ex. 206.

195. Websites can opt out of OpenAI accessing their webpages by configuring servers to "hard block" OpenAI requests based on source IP address. N. Feamster Decl. ¶¶ 70–71; J. Gratz Decl. ¶¶ 162, 164; Ex. 161 ¶253; Ex. 163 ¶ 158.

196. OpenAI publicly disclosed IP address ranges for OAI-SearchBot and ChatGPT-User to enable websites to identify and block incoming requests. N. Feamster Decl. ¶¶ 72–73; J. Gratz Decl. ¶ 191; Ex. 190.

197. The Times implemented a hard block against requests attributed to ChatGPT-User on April 16, 2024. J. Gratz Decl. ¶ 199; Ex. 198, Procas Ex. 5.

198. The Times implemented a hard block against requests attributed to OAI-SearchBot on July 25, 2024. J. Gratz Decl. ¶ 199; Ex. 198, Procas Ex. 5.

199. Ziff Davis implemented a "hard block" against requests attributed to OpenAI on July 1, 2025. J. Gratz Decl. ¶¶ 169, 208–209; Ex. 168 at 191:14–22, 192:21–25; Ex. 207; Ex. 208.

200. The Daily News Plaintiffs have not implemented a "hard block" against requests

attributed to OpenAI. N. Feamster Decl. ¶ 74; J. Gratz Decl. ¶ 210; Ex. 209 at 427:1–3, 442:21–444:6.

201. CIR has not implemented a "hard block" against requests attributed to OpenAI. N. Feamster Decl. ¶ 74; J. Gratz Decl. ¶ 211; Ex. 210 at 456:6–11, 458:1–6.

202. The Intercept has not implemented a "hard block" against requests attributed to OpenAI. N. Feamster Decl. ¶ 74; J. Gratz Decl. ¶ 202; Ex. 201.

IX. News Plaintiffs, the Asserted Works, and the Allegations in the News Cases

A. News Plaintiffs' Asserted Works

203. News Plaintiffs have identified over 10.8 million Asserted Works in this case. Declaration of Dr. Avi Goldfarb ("A. Goldfarb Decl.") ¶¶ 16, 34, 50, 65, 78; J. Gratz Decl. ¶¶ 19, 54, 158; Ex. 18 at 40; Ex. 53 ¶ 140; Ex. 157 ¶¶ 221–227.

204. According to Plaintiffs' experts, 6.3 million of the Asserted Works were used to train OpenAI's LLMs. J. Gratz Decl. ¶¶ 212–214; Ex. 211 ¶ 15; Ex. 212 at 6; Ex. 213; Ex. 157 at 90, Table 4.

205. News Plaintiffs wrote their works to be read by a human audience. J. Gratz Decl. ¶¶ 183, 215, 220–223; Ex. 182 at 183:25–184:4; Ex. 214 at 12:9–14; Ex. 219 at 49:1–13; Ex. 220 at 39:16–22; Ex. 221 at 41:11–17; Ex. 222 at 99:20–100:19.

206. News Plaintiffs' experts opine that ChatGPT substitutes for news content because it "satisf[ies] user demand for information." J. Gratz Decl. ¶ 224; Ex. 223 at 31.

207. The Times' Asserted Works were published between 1928–2024. A. Goldfarb Decl. ¶ 16, Fig. 1.

208. The Times identified 6,030,928 articles in its July 17, 2026 Asserted Works list. Only 2.9% of the articles were published in or after 2018. Approximately 82.73% were

published before 2000, 95.86% before OpenAI was founded in November 2015, 98.38% before GPT-3 was launched in May 2020, and 99.47% before ChatGPT was released in November 2022. Of the list, 3,993,611 articles, or 66.2%, were published before 1981. A. Goldfarb Decl. ¶ 17, Fig. 1.

209. DNP's Asserted Works are composed of older content published before OpenAI's LLMs existed. 31.6% of the articles were published before 2000, 87.4% before OpenAI was founded in November 2015, and 99.2% before ChatGPT was released in November 2022; 1% were published in or after September 2022, and all works were published before GPT-4o was released in May 2024. A. Goldfarb Decl. ¶ 35, Fig. 11.

210. CIR's *Mother Jones* Asserted Works ranged from April 1979 through September 2024, and 92% were published in 2020 or earlier. A. Goldfarb Decl. ¶ 66, Fig. 34.

211. CIR's *Reveal* Asserted Works ranged from October 2002 through September 2024, and 92% were published in 2020 or earlier. A. Goldfarb Decl. ¶ 68, Fig. 36.

212. Dr. Goldfarb found, based on his traffic decay analysis for the Asserted Works, that ChatGPT outputs generated years after an Asserted Work's publication date have limited opportunity to divert meaningful traffic from an article whose demand had largely been realized. A. Goldfarb Decl. ¶¶ 20, 39, 70, 80.

213. The Intercept's Asserted Works list contained 9,511 works with publication dates ranging from February 2014 to June 2024. 11.3% of the articles were published before OpenAI was founded in November 2015, 88.4% before ChatGPT was released in November 2022, and 99.4% before GPT-4o was released in May 2024. A. Goldfarb Decl. ¶ 78, Fig. 46.

214. News Plaintiffs republish facts first published by others. J. Gratz Decl. ¶¶ 203, 217, 225–233; Ex. 202 at 360:22–361:1; Ex. 216 at 355:17–19; Ex. 224 at 93:2–8; Ex. 225 at

52:20–53:4; Ex. 226 at Resp. to RFA No. 33; Ex. 227 at Resp. to RFA No. 33; Ex. 228 at 190:1–3; Ex. 229 at Resp. to RFA No. 43; Ex. 230 at 49:5–9; Ex. 231 at 87:19–25; Ex. 232 at 16:3–8.

215. Adam Doree, Ziff Davis's Senior Vice President of Global Partners & International, testified that "almost all news-based reporting is reporting what other people have reported." J. Gratz Decl. ¶ 231; Ex. 230 at 49:5–9.

216. News Plaintiffs' Asserted Works include republished press releases or wire service reports. J. Gratz Decl. ¶¶ 301–303; Exs. 300–302.

217. News Plaintiffs' Asserted Works include "Sports Results" from July 13, 1941 and a "TV Schedule" for the 1976 Winter Olympics, published by The Times. J. Gratz Decl. ¶¶ 91–92; Exs. 90, 91.

218. News Plaintiffs' Asserted Works include "Colorado snow totals for Oct. 19, 2023" from the Denver Post. J. Gratz Decl. ¶ 236; Ex. 235.

B. Experts' Analyses of Conversation Logs and Models

219. OpenAI's expert analyzed 20 million ChatGPT conversation logs that OpenAI produced in this case and found 24 instances containing verbatim regurgitation—*i.e.*, an alleged regurgitation rate of 0.00012%. T. Berg-Kirkpatrick Decl. ¶ 28 (24 / 20M = 0.00012%).

220. Dr. Goldstein, expert for The Times and the Daily News Plaintiffs found 9 examples of regurgitation in the same 20 million conversation logs, 93 examples of regurgitation in 78 million conversation logs, and 18 examples of regurgitation in 9 million conversation logs, which constitutes an overall regurgitation rate of 0.00011%. J. Gratz Decl. ¶ 237; Ex. 236 ¶¶ 93–94, 102–103 (120 / (20M + 78M + 9M) = 0.00011%).

221. Dr. Wenger, expert for The Intercept and CIR, found similar regurgitation rates. For The Intercept, Dr. Wenger identified two instances of regurgitation across the 20 million

conversation logs, and no regurgitations for either the 78 million or 10 million conversation logs she later analyzed, which constitutes an overall regurgitation rate of 0.000002%. J. Gratz Decl. ¶¶ 238–240; Ex. 237 at 48–49; Ex. 238 at 122; Ex. 239 at 189:6–19 (2 / (20M + 78M +10M) = 0.000002%).

222. For CIR, even including examples where the user pasted text from a Plaintiff article as input, Dr. Wenger found only 9 instances of regurgitation across a set of 78 million and 10 million conversation logs, and none in the 20 million conversation logs, which constitutes an overall regurgitation rate of 0.000008%. J. Gratz Decl. ¶ 240; Ex. 239 at 185:11–186:1, 189:20–192:16 (9 / (20M + 78M + 10M) = 0.000008%).

223. The longest verbatim string of words identified by Dr. Wenger contained 29 words. J. Gratz Decl. ¶ 238; Ex. 237 at 48–49.

224. Mr. Delorey, expert for Ziff Davis, identified four instances of regurgitations from the same 20 million conversation logs, which constitutes a regurgitation rate of 0.00002%. J. Gratz Decl. ¶ 212; Ex. 211 ¶¶ 113–114 (4 / 20M = 0.00002%).

225. The longest verbatim string of words in these examples had a length of 43 words. J. Gratz Decl. ¶ 212; Ex. 211 at 95–98, Figs. 16.1–16.4.

226. OpenAI's expert, Dr. Berg-Kirkpatrick, conducted a study where he submitted prompts about the specific subject matter addressed in 710 randomly selected Asserted Works and found that in over 68,000 GPT model outputs, there were ***zero*** instances of regurgitation. T. Berg-Kirkpatrick Decl. ¶¶ 41–42.

227. Dr. Berg-Kirkpatrick conducted another study in which he used adversarial prompting techniques targeting 19 randomly sampled works (one from each of 19 outlets across the five News Plaintiffs). T. Berg-Kirkpatrick Decl. ¶¶ 47–48.

228. Dr. Berg-Kirkpatrick submitted over 9,400 prompts to attempt to induce regurgitation of the News Plaintiffs' Asserted Works and ChatGPT regurgitated on average 2.3% of the content from the selected works. T. Berg-Kirkpatrick Decl. ¶¶ 49–50.

229. The Times and Daily News Plaintiffs' expert Dr. Goldstein performed his own regurgitation study where he submitted 195 million total prompts. J. Gratz Decl. ¶ 28; Ex. 27 at 401:11–403:14.

230. Dr. Goldstein submitted these prompts through a technical interface and used non-public models including code-davinci-002 and gpt-4-base. T. Berg-Kirkpatrick Decl. ¶ 59; J. Gratz Decl. ¶ 158; Ex. 157 ¶ 314.

231. Dr. Goldstein prompted the models using "the first few sentences of an article" or the title of an article from The Times or Daily News Plaintiffs' Asserted Works and asked the model to continue or deliver the text. J. Gratz Decl. ¶ 158; Ex. 157 ¶¶ 310–314.

232. Dr. Goldstein also used a temperature setting of zero. J. Gratz Decl. ¶ 28; Ex. 27 at 400:14–401:4.

233. Based on Dr. Goldstein's studies, The Times and Daily News Plaintiffs claimed "regurgitations" of approximately 87,000 Times Asserted Works and 28,000 Daily News Asserted Works which on average reproduced under 4.4% of the work. J. Gratz Decl. ¶¶ 158, 241; Ex. 157 at 111, Table 7; Ex. 240, Table 4, ¶¶ 244–248.

234. Ziff Davis's expert, Mr. Delorey, performed a regurgitation study where he submitted 65 million total prompts. J. Gratz Decl. ¶ 16; Ex. 15 at 232:25–233:5.

235. Mr. Delorey submitted these prompts through a technical interface and used non-public models including code-davinci-002 and gpt-4-base. T. Berg-Kirkpatrick Decl. ¶ 59; J. Gratz Decl. ¶ 54; Ex. 53 ¶ 278 n.167.

236. Mr. Delorey prompted the models using text from Ziff Davis's Asserted Works and tiered prompt strategies asking the model to continue or reconstruct the text. J. Gratz Decl. ¶ 54; Ex. 53 ¶¶ 280–283 n.169, 418–420.

237. He also adjusted the model temperature to zero which "maximiz[es] deterministic results." J. Gratz Decl. ¶ 54; Ex. 53 ¶ 281 n.169.

238. Based on Mr. Delorey's study, Ziff Davis claimed "regurgitations" of approximately 172,952 Ziff Davis Asserted Works, which on average reproduced less than 4.3% of the work. J. Gratz Decl. ¶¶ 54, 212, 242; Ex. 53 ¶ 290; Ex. 211 ¶¶ 101; Ex. 241 ¶ 115.

239. Dr. Goldstein and Mr. Delorey collectively claimed "regurgitations" that contain at least 10 words from 288,000 New York Times, Daily News Plaintiffs, and Ziff Davis Asserted Works. This represents 2.6% of the roughly 11 million total works these Plaintiffs are asserting in this case. T. Berg-Kirkpatrick Decl. ¶ 66.

C. News Plaintiffs' Allegations of Harm

240. News Plaintiffs' damages expert, James E. Malackowski, testified that since ChatGPT's release, "the traffic for The New York Times has continued to grow. Traffic for the Daily News [redacted] And the traffic for Ziff Davis [redacted]." J. Gratz Decl. ¶ 243; Ex. 242 at 93:4–7.

241. Dr. Avi Goldfarb, an expert economist retained by OpenAI, used Comscore user-level traffic data to distinguish users who accessed ChatGPT from users who did not. A. Goldfarb Decl. ¶ 27.

242. Dr. Goldfarb compared Comscore user-level data for users' visits to The Times, Daily News Plaintiffs, and Ziff Davis websites before and after first observed ChatGPT use. A. Goldfarb Decl. ¶ 29, Figs. 8–9; ¶ 46, Fig. 20; ¶ 61, Fig. 32.

243. Dr. Goldfarb found no statistically significant decrease in users' visits to The Times, Daily News Plaintiffs, or Ziff Davis websites based on users' adoption of ChatGPT. A. Goldfarb Decl. ¶¶ 30–32, 47–49, 62–64.

244. Dr. Goldfarb analyzed aggregate and page-level traffic data for CIR and The Intercept. A. Goldfarb Decl. ¶¶ 74–75, 85–86.

245. Dr. Goldfarb did not find a new, sustained decline at ChatGPT's release or at later updates for aggregate and page-level traffic data as to CIR and The Intercept. A. Goldfarb Decl. ¶¶ 75–76, Figs. 42–44; ¶¶ 86–87, Figs. 50–51.

246. Dr. Goldfarb concluded that there was no measurable market harm attributable to ChatGPT distinct from preexisting trends and other publisher- or platform-specific factors for any of the News Plaintiffs. A. Goldfarb Decl. ¶¶ 9, 13–15, 89.

247. The Times's retained expert, Professor Emily Bell, testified that she did not "personally conduct" any empirical analysis showing that Copilot or ChatGPT users reduced their consumption of The Times content. J. Gratz Decl. ¶ 244; Ex. 243 at 168:20–170:3.

248. Professor Bell testified she was unaware of any empirical analysis showing that Copilot or ChatGPT users reduced their consumption of The Times content. J. Gratz Decl. ¶ 244; Ex. 243 at 168:20–170:3.

249. Professor Bell testified that she did not conduct any analysis of the Times' web traffic data showing a decline correlated with Copilot usage. J. Gratz Decl. ¶ 244; Ex. 243 at 169:3–170:3.

250. The Times' internal data identified "the top seven reasons" customers canceled subscriptions, and generative AI was not listed among them. J. Gratz Decl. ¶ 245; Ex. 244 at 164:3–18.

251. New York Times audience insights manager, Camilla Velasquez, testified that algorithm changes are the "best hypothesis" the Times has for "what is affecting the change" in Google referral traffic. J. Gratz Decl. ¶ 246; Ex. 245 at 138:10–17, 139:14–18.

252. The Times' Executive Vice President and Chief Financial Officer, William Bardeen, testified that Facebook's deprioritization of news content "led to less traffic" to The Times' website. J. Gratz Decl. ¶ 247; Ex. 246 at 47:18–24.

253. When asked if he "conducted any analysis to determine whether any declines in the Daily News Plaintiffs' traffic or revenue were, in fact, caused by generative AI," Daily News Plaintiffs' retained expert, Taylor Owen, testified that he had "not done that study." J. Gratz Decl. ¶ 248; Ex. 247 at 106:20–24.

254. Mr. Owen testified that "the addition of Google Overviews to Google Search has . . . probably, likely, a more significant impact on the business models of publishers than the emergence of stand-alone chatbots." J. Gratz Decl. ¶ 248; Ex. 247 at 157:2–9.

255. When asked whether the Chicago Tribune conducted "any analysis to determine any reduction in referral traffic based on the use of particularly [OpenAI's] generative AI products" or other generative AI tools, Mitch Pugh, Executive Editor of the Chicago Tribune, testified that he was not aware of any such study. J. Gratz Decl. ¶ 249; Ex. 248 at 103:15–104:3.

256. The Daily News's corporate designee, CJ Jacobs, testified that The Daily News's analysis of traffic declines indicated that "the problem was a number of factors, not just one factor." J. Gratz Decl. ¶ 170; Ex. 169 at 297:6–23.

257. The Daily News's corporate designee, Adel Ibrahim, did not recall any "specific customer" or customer record showing that a DNP subscriber canceled a subscription because of generative AI. J. Gratz Decl. ¶ 80; Ex. 79 at 79:14–19.

258. Ziff Davis's retained expert, Keith Waehrer, testified that determining what portion of the alleged harm to Ziff Davis is attributable to OpenAI's conduct, as opposed to conduct by other generative AI companies, was "beyond the scope of" his analysis. J. Gratz Decl. ¶ 250; Ex. 249 at 85:14–19.

259. Ziff Davis's Chief Technology Officer, Joey Fortuna, testified that he had not seen any "quantitative" analysis of the impact of OpenAI's products on Ziff Davis's business. J. Gratz Decl. ¶ 251; Ex. 250 at 127:14–25.

260. Ziff Davis's CEO, Vivek Shah, testified that Ziff Davis is "[redacted]" what led to the [redacted]. J. Gratz Decl. ¶ 223; Ex. 222 at 76:4–11.

261. CIR and The Intercept's retained expert, Dr. Michael McDonald, testified that his regression model used to analyze The Intercept's traffic is "necessarily basic" and that "[d]efinitively proving harm, or establishing the magnitude of it, is outside the scope of [his] analysis." J. Gratz Decl. ¶ 253; Ex. 252 at 17 n.13.

262. CIR and The Intercept's retained expert, Dr. McDonald, testified that "it's difficult to suss out specifically what the magnitude of the harm is associated with artificial intelligence versus other possible factors[.]" J. Gratz Decl. ¶ 254; Ex. 253 at 94:12–95:4.

263. CIR and The Intercept's retained expert, Dr. McDonald, testified that "a variety of factors" may have caused harm to CIR and Intercept. Dr. McDonald testified that his regression analysis does not answer the question of how generative AI has caused harm to CIR and Intercept. J. Gratz Decl. ¶ 254; Ex. 253 at 92:6–13.

264. CIR and The Intercept's retained expert, Sue Cross, testified that she has not "attempted to do any causal analysis to isolate the impact of ChatGPT or Copilot from other

factors that might influence web traffic to CIR or The Intercept's websites in connection with this case." J. Gratz Decl. ¶ 255; Ex. 254 at 142:15–21.

265. The Intercept's board materials noted that "[redacted]." J. Gratz Decl. ¶ 144; Ex. 143 at 226:14–17.

266. The Intercept's board materials noted that "[redacted]." J. Gratz Decl. ¶ 144; Ex. 143 at 226:24–227:3.

267. The Intercept's corporate designee, Annie Chabel, testified that The Intercept's readership decline started before the introduction of ChatGPT and "for a number of complicated reasons." J. Gratz Decl. ¶ 144; Ex. 143 at 205:24–207:8.

268. Ms. Chabel testified that The Intercept does not have "any direct data" showing that ChatGPT diverted readers away from The Intercept's website. J. Gratz Decl. ¶ 144; Ex. 143 at 207:25–208:8.

269. Ms. Chabel testified that The Intercept does not "have an analysis showing who didn't come" to The Intercept's website "because they're using other platforms." J. Gratz Decl. ¶ 144; Ex. 143 at 46:7–16.

270. When asked whether "The Intercept [has] any data about the number of users that are going to social media or AI news sources instead of coming to The Intercept," Ms. Chabel testified "[n]o." J. Gratz Decl. ¶ 144; Ex. 143 at 46:17–22.

271. News Plaintiffs' expert, Jessie Stricchiola, asserts that News Plaintiffs "lost" clicks because ChatGPT's "synthesized response" is "sufficient on its own," and leaves users "feeling their information needs satisfied without consulting external sources such as the website

of the original source of the information." J. Gratz Decl. ¶ 162; Ex. 161 ¶ 438.

272. Ms. Stricchiola did not calculate click through rates on ChatGPT. A. Goldfarb Decl. ¶¶ 90–91.

273. Ms. Stricchiola testified that she did not "perform[] an analysis on traffic or quantifying traffic volumes in any way." J. Gratz Decl. ¶ 256; Ex. 255 at 82:24–83:10.

274. At the time of her deposition on October 24, 2025, when asked whether she was "aware of any evidence or any data showing that any negative impact to The Times' business can be traced back to conduct by OpenAI . . . as opposed to competitive pressures in the many markets in which The Times competes," Alex Hardiman, Executive Vice President and Chief Product Officer at The Times, testified that she was not aware of any. J. Gratz Decl. ¶ 257; Ex. 256 at 200:22–201:7.

275. The Times's revenue was approximately $2.825 billion in 2025, $2.038 billion in 2022, and $1.784 billion in 2020. J. Gratz Decl. ¶¶ 258–259; Ex. 257 at 104; Ex. 258 at 25.

276. The Times had over 12 million subscribers in December 2025 and is on track to reach its goal of 15 million subscribers by 2027. J. Gratz Decl. ¶¶ 258, 260, 304; Ex. 257 at 1; Ex. 259 at 5; Ex. 303.

277. In Dr. Goldfarb's combined traffic analysis for The Times, core news sessions showed no change at ChatGPT's November 30, 2022 release, while pageviews increased over time. A. Goldfarb Decl. ¶¶ 24–25, Fig. 5.

278. The Daily News Plaintiffs' reported digital-only subscriptions [redacted]. A. Goldfarb Decl. ¶ 40, Fig. 13.

279. Ziff Davis reported total revenue of approximately $1.451 billion for fiscal year

2025, an increase of 3.5% over fiscal year 2024. A. Goldfarb Decl. ¶¶ 52, Fig. 22; J. Gratz Decl. ¶¶ 261–262; Ex. 260 at 45; Ex. 261 at ZD0534683.

280. As of the end of fiscal year 2024, Ziff Davis did not identify threats posed by generative AI and related technologies as material to its businesses. J. Gratz Decl. ¶¶ 223, 252, 262–263; Ex. 222 at 24:8–13; Ex. 251 at 78:6–19; Ex. 261 at ZD0534656; Ex. 262 at 270:13–20.

281. Ziff Davis's 2025 Form 10-K did not identify generative AI as a material risk, but its Chief Executive Officer, Vivek Shah, agreed that "[redacted] it would have been [redacted]." J. Gratz Decl. ¶¶ 223, 261; Ex. 222 at 108:2–110:1; Ex. 260.

282. The Intercept is a "nonprofit organization" that "relies on donations from [its] readers and listeners." J. Gratz Decl. ¶ 300; Ex. 299 at INT0216064.

283. The Intercept's total contributions increased from approximately $4.0 million in 2022 to approximately $5.5 million in 2024, and its 2025 materials projected approximately $5.6 million. J. Gratz Decl. ¶¶ 264–266; Ex. 263 at INT0008100; Ex. 264 at INT0223423; Ex. 265 at INT0017403; Ex. 266; Ex. 267.

284. The Intercept's individual member contributions have been higher every year since the launch of ChatGPT than they were in the preceding years. A. Goldfarb Decl. ¶ 82, Fig. 48.

285. CIR is a "nonprofit journalism organization" in which donations are one of the main sources of revenue. J. Gratz Decl. ¶ 183; Ex. 182 at 45:9–12, 297:22–25.

286. CIR reported approximately $23.1 million in revenue for February through December 2024 and approximately $23.6 million for January through October 2025. A. Goldfarb Decl. ¶ 72, Fig. 38; J. Gratz Decl. ¶¶ 269–271; Exs. 268–270.

287. CIR recorded its highest monthly digital revenue to date in February 2025, which "set[] a new high watermark for [its] digital revenue efforts, driven by exceptional performance in one-time giving, subscriptions, sustainers, advertising, and audience growth." A. Goldfarb Decl. ¶ 72; J. Gratz Decl. ¶ 272; Ex. 271 at CIR0679029.

288. Dr. Goldfarb found that CIR's direct, search, and social referral traffic was declining before ChatGPT and did not show a clear break at ChatGPT's launch. A. Goldfarb Decl. ¶ 76, Fig. 44.

D. News Plaintiffs' Use of OpenAI's LLMs

289. The Times admitted that "at least fifteen tools [it] uses utilize, at least in part, OpenAI's Large Language Models." J. Gratz Decl. ¶ 273; Ex. 272 at Resp. to RFA No. 210; *see also id.* at Resp. to RFA Nos. 126, 186–194, 202–203, 205–209, 212–218.

290. The Times admitted to using "an internal tool called ChatExplorer that utilizes OpenAI's Large Language Models." J. Gratz Decl. ¶ 273; Ex. 272, at Resp. to RFA No. 126.

291. The Times admitted to using Echo, "which utilizes OpenAI's Large Language Models, for creating internal summaries of published New York Times content." J. Gratz Decl. ¶ 273; Ex. 272, at Resp. to RFA No. 186.

292. The Times admitted to using Cheatsheet, "which utilizes OpenAI's Large Language Models, to review spreadsheets and help reporters analyze data." J. Gratz Decl. ¶ 273; Ex. 272, at Resp. to RFA No. 187.

293. The Times admitted to using "a third-party tool called LibreChat, which utilizes OpenAI's Large Language models, as a chat interface." J. Gratz Decl. ¶ 273; Ex. 272, at Resp. to RFA No. 188.

294. The Times admitted to using the Manosphere Report, "which utilizes OpenAI's

Large Language Models, to analyze podcasts in the manosphere for reporting." J. Gratz Decl. ¶ 273; Ex. 272, at Resp. to RFA No. 189.

295. The Times admitted to using Cursor, "which utilizes OpenAI's Large Language models." J. Gratz Decl. ¶ 273; Ex. 272, at Resp. to RFA No. 190.

296. The Times admitted to using GitHub Copilot, "which utilizes OpenAI's Large Language Models." J. Gratz Decl. ¶ 273; Ex. 272, at Resp. to RFA No. 191.

297. The Times admitted to using Elvex, "which utilizes OpenAI's Large Language Models." J. Gratz Decl. ¶ 273; Ex. 272, at Resp. to RFA No. 192.

298. The Times admitted to using MindStudio, "which utilizes OpenAI's Large Language Models." J. Gratz Decl. ¶ 273; Ex. 272, at Resp. to RFA No. 193.

299. The Times admitted to using Graphics Bot, "which utilizes OpenAI's Large Language Models." J. Gratz Decl. ¶ 273; Ex. 272, at Resp. to RFA No. 194.

300. The Times admitted to using Connections Bot, "which utilizes OpenAI's Large Language Models, as a game companion to The New York Times Connections puzzle." J. Gratz Decl. ¶ 273; Ex. 272, at Resp. to RFA No. 202.

301. The Times admitted to using Switchboard, "which utilizes OpenAI's Large Language Models, for comment analysis"; "to ingest comments, cluster them, and then generate a title and summary of what commenters are talking about in that cluster"; and "for content summarization." J. Gratz Decl. ¶ 273; Ex. 272, at Resp. to RFA Nos. 207–209.

302. The Times admitted to using OpenAI's LLMs "for comment moderation"; "comment clustering"; and ".edu content matching." J. Gratz Decl. ¶ 273; Ex. 272, at Resp. to RFA Nos. 203, 205, 206.

303. The Times admitted to using OpenAI's LLMs "to generate a summary of major

themes contained in reader submissions collected through reader call outs"; "to generate a summary of the major themes contained in an article's comments and generate recommendations about which comments are best for reporter engagement"; "to answer reader questions about a specific topic and receive an answer"; "to generate 'What to Watch' content recommendations based on chat-style prompts"; "to research large text datasets"; "to generate draft alternative text to be edited by photo editors"; and "to generate editing recommendations for draft article text." J. Gratz Decl. ¶ 273; Ex. 272, at Resp. to RFA Nos. 212–218.

304. The Times "admits that it permits certain employees to use OpenAI's Large Language models to help with brainstorming and editing, and to summarize Times-owned materials and non-propriet[ar]y information." J. Gratz Decl. ¶ 274; Ex. 273, at Resp. to RFA No. 102.

305. The Times corporate designee Zach Seward testified that "Echo is an internal tool for summarizing The New York Times company articles" and which uses "a few different LLMs offered by Google and OpenAI." J. Gratz Decl. ¶ 275; Ex. 274 at 58:2–20.

306. According to Mr. Seward, Echo is "intentionally a broad tool" that can "help[] the team that writes [The Times's] morning newsletter." Mr. Seward also testified that teams at The Times "will rely on the first draft of [a] summary that's been generated by Echo and then edit that summary there before publication in the newsletter." J. Gratz Decl. ¶ 275; Ex. 274 at 58:2–20.

307. CIR's internal June 17, 2025 guidance states that AI tools "are being used by people [t]here right now to analyze data, make giant piles of information accessible, clean up audio, play around with vibecoding, draft emails, and lots more." The guidance identifies "[e]xamples of safe ways in which people at CIR have been using AI tools: To summarize or

search through documents for research (verifying any facts and reviewing/editing generated summaries before publication!); generate SEO material, check grammar, brainstorm headlines, and transcribe interviews (again with human review and verification for any passages used); brainstorm language for documents, memos, or emails; analyze spreadsheets (again with human review); write basic code (again with human review); and clean up audio and video (ditto)." J. Gratz Decl. ¶ 276; Ex. 275.

308. The CIR-affiliated article *40 Acres and a Lie*—published as a collaboration among Mother Jones, Public Integrity, and Reveal—states that the investigative team "used artificial intelligence to track down the people, places, and stories that had long been misunderstood and forgotten." J. Gratz Decl. ¶ 277; Ex. 276.

309. Bay Area News Group advertising content manager Rowan Briggs testified that ChatGPT generated the headline and text for a 350-word cannabis-themed filler article published by The Mercury News on December 24, 2024. Briggs noted that he had used generative-AI tools to generate ideas, including prompting ChatGPT to "generate 10 topics about cannabis and the new year." J. Gratz Decl. ¶¶ 218–219, 278; Ex. 217; Ex. 218; Ex. 277 at 53:8–54:6, 106:3–116:22, 119:6–22.

310. A Daily News Plaintiff document titled "ChatGPT for Content Creation: Opportunities and Limitations" identifies various ways to "[l]everage ChatGPT for [c]ontent [c]reation, including "[k]eyword generation"; "[t]opic ideation"; "[c]ontent briefing"; "[c]reative prompting"; "[s]chema markup"; and "[c]ontent drafts." It states that "ChatGPT can generate rough drafts" of short-form content such as "[s]ocial-media posts"; "[m]eta descriptions"; "[t]itle tags"; "[a]dvertising cop[ies]"; "[e]mail blasts"; and "[a]lt text." J. Gratz Decl. ¶ 279; Ex. 278.

311. A Daily News Plaintiff document titled "Enhancing Your Workflow with ChatGPT: An Introduction for BANG Employees" informs Bay Area News Group employees that ChatGPT can be used to "get a jump start on a project"; "[j]ump start research without having to scroll through search results"; [g]enerate [c]reative [c]ontent; and "[s]ummarize [d]ata." J. Gratz Decl. ¶ 280; Ex. 279 at DNP-00019574–76.

312. A Ziff Davis internal document states that [redacted] developed a [redacted]." According to the document, [redacted] and [redacted] Ziff Davis media brands. J. Gratz Decl. ¶ 281; Ex. 280 at ZD0094239.

313. The "[redacted]" in Ziff Davis spreadsheet reports that "[redacted]" and to generate "[redacted]" The sheet also records uses of ChatGPT for "[redacted] for presentation decks. J. Gratz Decl. ¶ 282; Ex. 281.

314. A Ziff Davis internal document titled "[redacted]" states that [redacted] The document states that ChatGPT may be used for "[redacted]" and [redacted] and "[redacted]." J. Gratz Decl. ¶ 283; Ex. 282 at ZD0008631.

315. Ziff Davis corporate designee Robert Parisi testified that Everyday Health Group Consumer uses a "virtual editorial assistant named Aiden, which is powered by custom GPTs" and which is "used for creating summaries" of Everyday Health Group Consumer articles or "takeaway" text appearing near the top of articles. Parisi further testified that Aiden is used to "create meta tags like meta titles and . . . meta descriptions about those articles that [they are] writing"; used for "newsletter headlines and introduction copy in the newsletter that [they] are sending out"; and used "for topic tags which helps [them] in the categorization within [their] taxonomy of the content that [they have] written." J. Gratz Decl. ¶ 284; Ex. 283 at 44:18–45:10.

316. In a February 25, 2025 Intercept email exchange, CEO Annie Chabel stated that she was "interested in growing membership contributions by . . . [u]sing AI to analyze growth opportunities in [The Intercept's] existing data." J. Gratz Decl. ¶ 286; Ex. 285 at INT0000803.

317. The Times uses OpenAI's LLMs through an internal chatbot called ChatExplorer, which accessed OpenAI's LLMs like GPT-3.5 Turbo and GPT-4 through the OpenAI API. J. Gratz Decl. ¶¶ 274, 296–297; Ex. 273 at Resp. to RFA Nos. 85–86; Ex. 295; Ex. 296 at Resp. to RFA No. 54.

318. One Times employee used ChatExplorer to assist in "brainstorming headline ideas." J. Gratz Decl. ¶ 287; Ex. 286.

319. One Times employee asked ChatExplorer to "cut a draft article to 500 words." J. Gratz Decl. ¶ 288; Ex. 287.

320. One Times employee asked ChatExplorer to "trim" another article "to roughly 500 words" without "alter[ing] any words within the article" and "chang[ing] any facts or meaning." J. Gratz Decl. ¶ 295; Ex. 294 at NYT_01066891.

321. One Times employee entered the following prompt into ChatExplorer: "It is clear

as a team and a company we are headed in a direction in which we will be utilizing GenAI more and more, my continued skills in this area will benefit both me and the team." J. Gratz Decl. ¶ 293; Ex. 292.

322. One Times employee entered the following prompt into ChatExplorer: "For selecting a top 10 list of events for specific news topic, for example Taylor Swift - what would be some criteria of significance that would generally apply to any news topic." J. Gratz Decl. ¶ 289; Ex. 288.

323. One Times employee entered the following prompt into ChatExplorer: "I work at the NYTimes, any follow up questions I should ask Harness about their answers to the following questions?" J. Gratz Decl. ¶ 290; Ex. 289 at NYT_01062157.

324. One Times employee entered the following prompt into ChatExplorer: "Compare and contrast Google's rollout of AI features within the Pixel 9 and Apple's rollout of Apple Intelligence [] summarize in 4-5 bullets." J. Gratz Decl. ¶ 292; Ex. 291.

325. One Times employee entered the following prompt into ChatExplorer: "[W]rite an interview question to discuss the interviewee's observations on the characteristics and demographics of US soccer fans based on their experience." J. Gratz Decl. ¶ 291; Ex. 290.

326. Rowan Briggs, corporate designee for DNP, testified that ChatGPT and Gemini were used "to create paid advertorial content for Best In spotlights and real estate content" for Mercury News. J. Gratz Decl. ¶¶ 216, 234, 278, 281; Ex. 215 at 452:5–22; Ex. 233 at CIR0003767; Ex. 277 at 74:18–75:8; Ex. 280 at ZD0094237.

327. The Times admitted to "develop[ing] a tool for targeted advertising that uses generative AI called BrandMatch." J. Gratz Decl. ¶ 274; Ex. 273 at Resp. to RFA No. 121.

328. The Times admitted that "at one point, BrandMatch used OpenAI's Large

Language Models." Alex Hardiman, Executive Vice President and Chief Product Officer at The Times, testified that "Embex" was a predecessor version of BrandMatch and that it "drove value for [The Times's] advertising business." J. Gratz Decl. ¶¶ 257, 274; Ex. 256 at 105:8–18; Ex. 273 at Resp. to RFA No. 122.

329. Christine Mattheis, Group Vice President of Content and Brand Solutions for Everyday Health, Pregnancy, and Parenting, testified that she used ChatGPT for "brainstorming things like . . . how to make our products more engaging." J. Gratz Decl. ¶ 305; Ex. 304 at 75:14–76:2.

330. Robert Parisi, corporate designee for Ziff Davis, stated that "plaintiffs use ChatGPT for paid e-mail acquisition forecasting," which is "financial analysis around determining how many e-mail addresses [they are] able to acquire through paid platforms." J. Gratz Decl. ¶ 284; Ex. 283 at 166:8–16.

331. The Times publicly stated in August 2025 that "AI will augment journalism and in no way replace it." J. Gratz Decl. ¶ 257; Ex. 256 at 141:6–19.

332. In a 2023 presentation made to The New York Times Board of Directors about Generative AI, The Times stated, "GenAI cannot compete on [The Times's] human beat reporting built on sources, presence, and expertise." J. Gratz Decl. ¶¶ 184, 299; Ex. 183 at 169:16–171:16; Ex. 298 at NYT_00871499.

333. Samuel Dolnick, Deputy Managing Editor for The Times, testified that he "believe[d] that [The Times's] human reporters can go out in the world and find out things that LLMs cannot." J. Gratz Decl. ¶ 306; Ex. 305 at 32:4–6.

334. Mr. Dolnick and Mr. Seward wrote an article that states: "The expertise and judgment of [The Times's] journalists are competitive advantages that machines simply can't

match, and [The Times] expect[s] that will become even more important in the age of AI." J. Gratz Decl. ¶ 306; Ex. 305 at 33:6–14.

335. Michael Turpin, Executive Vice President and Chief Revenue Officer of The Mercury News, testified that he "do[es] not view generative AI as a substitute for published news/advertorial environment content" and "do[es] not believe that the [generative AI] technology is a substitute for human involvement in the collection and producing of news content." J. Gratz Decl. ¶ 185; Ex. 184 at 33:2–8, 74:1–9.

X. Technical Background for DMCA claims

336. The raw data OpenAI obtains from the Internet does not look like what users see when they visit a website on a browser. C. Callison-Burch Decl. ¶¶ 80–89; J. Gratz Decl. ¶¶ 16, 19, 36; Ex. 15 at 194:10–22; Ex. 18 at 15; Ex. 35 at 243:25–244:6.

337. The raw data contains thousands of lines of HTML code, graphics specifications, and other, non-language data that is illegible to most users. C. Callison-Burch Decl. ¶¶ 78, 80; J. Gratz Decl. ¶¶ 16, 19, 23, 36, 139, 158; Ex. 15 at 193:25–194:4, 194:10–22; Ex. 18 at 15; Ex. 22 at 33:15–17; Ex. 35 at 243:25–244:6; Ex. 138; Ex. 157 ¶ 138.

338. Attribution information that might qualify as "CMI" is buried within hundreds if not thousands of fields and scattered throughout illegible raw data. C. Callison-Burch Decl. ¶¶ 83–88; Declaration of Barbara Frederiksen-Cross ("B. Frederiksen-Cross Decl.") ¶¶ 10–11.

339. Such fields of illegible raw data include hidden metadata fields that list title or author. C. Callison-Burch Decl. ¶¶ 82–87; B. Frederiksen-Cross Decl. ¶¶ 10–11.

340. As an example, such raw data fields may contain HTML code that, when executed by a web browser, will display a "byline" at the top of a browser. C. Callison-Burch Decl. ¶ 85; B. Frederiksen-Cross Decl. ¶¶ 10–11.

341. Training a language model on illegible, non-language data would result in a model that primarily generates illegible, non-language data. C. Callison-Burch Decl. ¶ 81; J. Gratz Decl. ¶¶ 19, 93, 141–143; Ex. 18 at 22; Ex. 92 ¶¶ 71–73; Exs. 140–142.

342. Avoiding the use of "unnatural" language data for training is a well-established and necessary part of NLP research. C. Callison-Burch Decl. ¶ 81; J. Gratz Decl. ¶¶ 53; Ex. 52 at 214–15.

343. OpenAI ran computer scripts to identify and extract usable language data for model training. C. Callison-Burch Decl. ¶¶ 82, 88; B. Frederiksen-Cross Decl. ¶¶ 12–22; J. Gratz Decl. ¶¶ 16, 19, 22–23, 54, 158; Ex. 15 at 195:15–23; Ex. 18 at 48–49; Ex. 21 at 190:17–191:6; Ex. 22 at 33:14–20; Ex. 53 ¶ 248; Ex. 157 ¶ 138.

344. The purpose of OpenAI's text-extraction scripts was to convert raw HTML files into chunks of natural language text suitable for model training—not to remove CMI. C. Callison-Burch Decl. ¶ 81; J. Gratz Decl. ¶¶ 16, 19, 23, 36, 54; Ex. 15 at 195:15–23; Ex. 18 at 48–49, 52, 54–56; Ex. 22 at 42:20–23; Ex. 35 at 245:6–12, 246:2–16; Ex. 53 ¶¶ 223, 248; Ex. 157 ¶ 138.

345. OpenAI's source code for text-extraction scripts uses a series of heuristics to identify and retain blocks of human language data on a page. B. Frederiksen-Cross Decl. ¶¶ 12–22.

346. As a result of OpenAI's use of text-extraction scripts, some data that might qualify as CMI for some works is not extracted from the raw files and used for model training. C. Callison-Burch Decl. ¶¶ 82–87; J. Gratz Decl. ¶¶ 19, 23, 54; Ex. 18 at 52, 54–56; Ex. 22 at 36:6–20; Ex. 53 ¶¶ 223, 248.

347. OpenAI's expert Dr. Berg-Kirkpatrick analyzed all 68 alleged "regurgitations"

identified in the parties' opening expert reports (across 108 million conversations) and found that in 63 of the 68 identified instances the source of the allegedly regurgitated text would have been obvious to the user. T. Berg-Kirkpatrick Decl. ¶ 77.

348. News Plaintiffs' experts did not conduct any experiment suggesting that retaining CMI in training data would enable a model to generate accurate CMI alongside regurgitated body text. J. Gratz Decl. ¶ 240; Ex. 239 at 170:10–18.

349. News Plaintiffs' experts did not cite any third-party study suggesting that retaining CMI in training data would enable a model to generate accurate CMI alongside regurgitated body text. J. Gratz Decl. ¶¶ 19, 93, 241; Ex. 18 at 110–116; Ex. 92 ¶¶ 250–253; Ex. 240 ¶¶ 171–172.

350. One of News Plaintiffs' experts acknowledged that, had OpenAI's models been trained with News Plaintiffs' CMI, CMI would not be output with every regurgitation and the models could misattribute information by generating false CMI. J. Gratz Decl. ¶ 240; Ex. 239 at 169:5–170:8.