NEWS MEDIA ALLIANCE

White Paper: How the pervasive copying of expressive works to train and fuel generative artificial intelligence systems is copyright infringement and not a fair use

4401 N. Fairfax Drive | Suite 300 | Arlington, VA 22203 | 571.366.1000 | www.newsmediaalliance.org

TABLE OF CONTENTS

I. Executive Summary

This White Paper is published by the News/Media Alliance (N/MA) to address the rampant copying of its members' expressive works to train generative artificial intelligence (GAI) systems.[1] N/MA member newspaper, magazine, and digital media publishers speak with a collective voice in supporting the responsible development of GAI while ensuring fair credit and compensation for the creators whose works make GAI possible. N/MA members welcome working with GAI developers to help build and grow this exciting new technology, in ways that can benefit all actors and society at large.

GAI systems, while holding promise for consumers, businesses, and society at large, are commercial products that have been built—and are run—on the backs of creative contributors. These systems have been developed by copying massive amounts of the creative output of the Alliance's members, almost always without authorization or compensation. And they disseminate the same kind of content for the same commercial purpose—sometimes in the same or substantially similar form—in response to user queries, again without authorization or payment and often with little or no attribution or link to the original source. Such disassociated output diminishes the need for users to click through or subscribe to N/MA members' print and digital publications. This irreparably damages publishers' businesses, which depend on relationships with their readers, web traffic, and the trustworthiness of brands built over decades.

An analysis commissioned by the News/Media Alliance shows that GAI developers disproportionally use online news, magazine, and digital media content to train their GAI models. Their affinity for this quality content highlights its value and expressive nature. The analysis demonstrates:

- GAI developers create curated sets of training data to build Large Language Models (LLMs), which then power GAI products. We have analyzed the data sets used to build these models and the output that they generated, and that analysis demonstrates that the developers have copied and used news, magazine, and digital media content to train the LLMs.

- In fact, our analysis of a representative sample of news, magazine, and digital media publications shows that the popular curated datasets underlying some of the most widely used LLMs significantly overweight publisher content by a factor ranging from over 5 to almost 100 as compared to the generic collection of content that the well-known entity Common Crawl has scraped from the web.

- Other studies show that news and digital media ranks third among all categories of sources in Google's C4 training set, which was used to develop Google's GAI-powered search capabilities and products like Bard. Half of the top ten sites represented in the training set are news outlets.

---

[1] In addition to counsel at the News/Media Alliance, this paper was co-authored by Cynthia S. Arato, Shapiro Arato Bach LLP, and Ian B. Crosby, Susman Godfrey LLP.

- The LLMs also copy and use publisher content in generating outputs. The LLMs can reproduce the content on which they were trained, demonstrating that the models retain and can memorize the expressive content of the training works.

This pervasive copying infringes N/MA members' exclusive rights in their copyrighted works and is not excused by the fair use doctrine, as the two most important fair use factors (the purpose and character of the use and the effect of the use on the market for the original) demonstrate:

- The GAI copying for "training" does not serve a purpose different from the original works because LLMs typically ingest (i.e., copy) valuable news, magazine, and digital media web content for their written expression, so that they can mimic that very form of expression. As one GAI proponent has explained, LLMs that are trained to generate their own expressive works "copy expression for expression's sake." Training LLMs on reliable, trusted expressive content without authorization also seeks to override licensing markets that already exist for these works, and copying for these training purposes thus serves (and supplants) that same licensing purpose. The GAI uses are also overwhelmingly commercial, helping to propel the GAI companies' valuations into the billions. And there is no compelling justification to allow the copying of creative works without fairly compensating the creators.

- The outputs of GAI models also directly compete with the protected content that was copied and used to train them. The use of these models to provide complete narrative answers to prompts and search queries goes far beyond the purpose of helping users to navigate to original sources (i.e., search) that has been found in the past to justify the wholesale copying of online content to build search engines. Indeed, GAI developers boast that users no longer need to access or review such sources. In this setting, the GAI developers' goal to create large language models, however laudable, does not justify their infringement of this valuable corpus of copyrighted expression.

While GAI developers contend that GAI models are just "learning" unprotectable facts from copyrighted training materials, that anthropomorphic claim is technically inaccurate and beside the point. It is inaccurate because models retain the expressions of facts that are contained in works in their copied training materials (and which copyright protects) without ever absorbing any underlying concepts. It is beside the point because materials that are used for "learning" are subject to copyright law. Even libraries must legally acquire the books they lend, and borrowers aren't free to copy them, especially not for an ultimate commercial use.

The incipient and predictable consequence of GAI's substitutive uses will be to damage the news and digital media industry. And it is not just copyright owners but society that will lose if GAI is allowed to so harm the journalism industry. Indeed, if the Internet becomes flooded with the products of GAI, then GAI itself will have nothing left to train on.

But GAI developers and publishers can work together to avoid such dire results. Indeed, publishers welcome technological progress and rely every day on innovative tools to tell their stories and inform the public, particularly where stories need to be globally transmitted and reported in real time through increasingly visual storytelling. N/MA members thus wish to work with GAI developers to maximize the value of this exciting new technology, in a way that is fair to publishers

and equitably shares the wealth generated from the N/MA content that the GAI developers copy and redeploy. Such fruitful cooperation between the GAI developers and the owners of these source works will benefit not just the news and media industries but the GAI developers and society at large, by helping to ensure that GAI is developed using high-quality and human created works.

Our culture, our economy, and our democracy require a solution that allows the news and media industry to grow and flourish, and both to share in the profit from and participate in the development of the GAI revolution that is being built upon the fruits of its labor. Part of this solution is offered by copyright law, which exists to ensure that creators and content owners are appropriately compensated for their copyrighted works and to incentivize the continued creation of such works, for the benefit of society at large.

This White Paper concludes with several recommendations: (1) GAI developers must be transparent and open about their use of expressive works in GAI models; (2) industry and policymakers must understand that unauthorized use of expressive works to train LLMs that are designed to generate expressive text in a commercial context is infringing; and (3) publishers must be able to license the use of their content efficiently and on fair terms.

II. Introduction

Generative artificial intelligence technologies can now mimic nearly any kind of work that humans create at vastly greater speed and lower cost—and at massive scale. Even the most enthusiastic proponents admit that GAI is *designed* to substitute for human creations: it has, they boast, "produced writing that's difficult to distinguish from real journalists, painted in the style of celebrated masters, and even created stock photos comparable to those of professional photographers."[2]

The ability of GAI to imitate and copy human expression quickly and cheaply brings opportunities with the potential to benefit society and greatly enrich the developers of these models. But popular models like ChatGPT can do so only because they have been trained on the fruits of human creativity at massive scale, and largely without consent or compensation. The works these models can imitate and copy in this way include prize-winning landmarks of culture produced at great cost to news, magazine, and digital publishers—and often at great peril to the journalists they employ.

While publishers have retrenched to survive in the Internet age, companies that develop foundational GAI models trained on these works have by contrast seen their valuations explode.[3] Platforms that deploy these GAI models into their products have likewise seen their market

---

[2] Mark A. Lemley & Brian Casey, *Fair Learning*, 99 Tex. L. Rev. 743, 767 (2021).

[3] *See, e.g.*, Cade Metz, *OpenAI in Talks for Deal That Would Value Company at $80 Billion*, N.Y. Times (Oct. 20. 2023), https://www.nytimes.com/2023/10/20/technology/openai-artifical-intelligence-value.html; Jagmeet Singh & Ingrid Lunden, *OpenAI Closes $300M Share Sale at $27B-29B Valuation*, TechCrunch (Apr. 28, 2023), https://techcrunch.com/2023/04/28/openai-funding-valuation-chatgpt/.

capitalizations soar.[4] Yet even though established markets exist for providing and licensing media content in a variety of contexts—including machine learning—almost none of this wealth has flowed to the rights holders of the writings whose wholesale copying fuels the capabilities of these immensely valuable GAI products.

The members of the News/Media Alliance are deeply concerned about this unauthorized and unlawful use of their expressive content by large technology companies. Such companies do not shoulder the cost or risk of reporting the news or producing creative content but capitalize on that valuable work. Indeed, publishers generally are not being paid by GAI developers for the unauthorized copying of their works to train the LLM models on which their chatbots are built. And those chatbots, like Bing Chat, Bard, ChatGPT, and Claude are often deployed to compete directly with those very works by, for example, providing narrative answers to search queries that obviate the need for consumers to click through to the original sources whose content permeates those responses.

In addition to chatbot applications, the newest generation of up-to-the-minute narrative search results, in particular by GAI applications like Google's Search Generative Experience and Microsoft's Bing Chat, exceeds any previously adjudicated limits of permissible use in the field. Such full and expressive responses directly compete with publisher content, sever publishers' connections to their readers, and bypass the very presence of their sites on the Internet. Indeed, Microsoft markets Bing as where to go to "Ask Real Questions. Get Complete Answers. Chat and Create."[5] Google's new "Search Generative Experience" has been described as a "plagiarism stew."[6]

As the accompanying technical analysis shows, the models also produce unauthorized derivative works by responding to user queries with close paraphrasing or outright repetition of copied and memorized portions of the works on which they were trained.

As with past "disruptive" Silicon Valley models, GAI investors are banking on forgiveness instead of asking permission. They depend on the claim that copying for training is a "fair use" that they may continue with impunity, even as many of their products directly compete with and threaten

---

[4] Marvie Basilan, *Microsoft Gets Stock Boost After Morgan Stanley's AI-Driven $3 Trillion Valuation Outlook*, Int'l Bus. Times (July 7, 2023), https://www.ibtimes.com/microsoft-gets-stock-boost-after-morgan-stanleys-ai-driven-3-trillion-valuation-outlook-3703880#:~:text=According%20to%20Morgan%20Stanley%2C%20Microsoft%20has%20a%2022%25,the%20company%20to%20hit%20a%20%243%20trillion%20valuation ("Microsoft has a 22% upside potential due to its 'pole position' in the generative AI race and this could propel the company to hit a $3 trillion valuation.").

[5] https://www.bing.com/new. As Microsoft admitted when it heralded the "new" Bing, it copies publisher content and delivers substitutional summaries: "There is no need to get overwhelmed sifting through search results. Bing distills the latest information from across the web to summarize and cite answers to your question. Microsoft.com, https://web.archive.org/web/20230710180333/https://www.microsoft.com/en-us/bing?form=MW00X7 (as of July 10, 2023).

[6] Avram Piltch, *Plagiarism Engine: Google's Content-Swiping AI Could Break the Internet*, Tom's Hardware (June 11, 2023), https://www.tomshardware.com/news/google-sge-break-internet.

the continued well-being of publishers. But fair use does not work this way. Indeed, the Supreme Court just ruled in *Andy Warhol Foundation for the Visual Arts v. Goldsmith* that even in the case of a highly creative adaptation, a use that has the potential to serve as a commercial substitute for an original work undermines a finding of fair use.[7] Simply having "some further purpose, in the sense that copying is socially useful," or "add[s] something new … does not render such uses fair."[8] *Warhol Foundation* recognizes that substitutive uses, however innovative, undermine the "economic incentive to create original works, which is the goal of copyright."[9]

The modes of distribution and consumption of publisher content are rapidly changing in the digital age, and the systematic copying and use of publisher content to fuel GAI systems and applications and to disseminate competing content poses what could be an existential threat to far too many publishers and is not a fair use. By diverting readers and the digital advertising dollars that follow them away from original sources, and by interfering with a potential source of licensing revenue for granting permissions, GAI models disincentivize investment in creation of those sources in the first place.

The continued unlicensed use of reporting also disserves the public interest: an online world that is dominated by GAI-generated, substitutional content is poised to leave the public with watered-down, less reliable outputs and fewer news outlets with the resources necessary to provide critical original reporting. As district court judge Denise Cote's decision in *Associated Press v. Meltwater U.S. Holdings, Inc.* explained with respect to direct scraping of news content that is economically indistinguishable from that now being laundered through GAI systems, copyright law should not allow for democracy to be imperiled in this manner:

> [T]he world is indebted to the press for triumphs which have been gained by reason and humanity over error and oppression … Permitting Meltwater to take the fruit of AP's labor for its own profit, without compensating AP, injures AP's ability to perform [its] essential function of democracy.[10]

GAI is now further threatening the ability of journalists and publishers to perform that "essential function of democracy." At a time when governments and experts around the world warn of the risk AI poses to democratic functioning,[11] it is critical that the copyright laws continue to protect publisher content to help safeguard the indispensable role of a flourishing and free press.

---

[7] *Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith*, 143 S. Ct. 1258, 1276-77 (2023).

[8] *Id.* at 1275.

[9] *Id.* at 1278.

[10] *Associated Press v. Meltwater U.S. Holdings, Inc.*, 931 F. Supp. 2d 537, 553 (S.D.N.Y. 2013).

[11] *See, e.g., Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People*, Off. Sci. & Tech. Pol'y, https://www.whitehouse.gov/ostp/ai-bill-of-rights/; Mekela Panditharatne & Noah Giansiracusa, Brennan Ctr. for Just., How AI Puts Elections at Risk — And the Needed Safeguards (July 21, 2023), https://www.brennancenter.org/our-work/analysis-opinion/how-ai-puts-elections-risk-and-needed-safeguards; Dan Milmo & Kiran Stacey, *AI-Enhanced Images* a *"Threat* to *Democratic*

III. Who We Are

The News/Media Alliance is a nonprofit organization that represents the interests of more than 2,200 news media organizations in the United States and internationally, including newspaper, magazine, and digital publishers. The Alliance represents the unified voice of the industry and diligently advocates before the federal government on issues that affect today's media organizations, including protecting publishers' intellectual property.

News media publications play a crucial role in the U.S. economy and democracy. Every day, their publishers invest in high-quality journalism that keeps our communities informed, holds those in power accountable, and supports the free flow of information and ideas in society. Without free and flourishing news media, our society would be less well-off and less informed. However, publishers' ability to continue serving as an essential source of news for readers around the world depends on their ability to receive fair compensation for the original expressive content that they have developed at high cost.

The news, magazine, and digital media industries' contribution to the U.S. economy and society is considerable, with estimated revenues of newspaper and magazine publishers amounting to approximately $45 billion.[12] Newsrooms were estimated to directly employ approximately 31,000 people in 2020, not including additional indirect employment effects, while magazines employed over 73,000 directly and supported a total of over 219,000 jobs in 2021.[13] Employment in digital-native newsrooms, meanwhile, has increased from approximately 7,400 in 2008 to over 18,000 in 2020.[14]

Journalists and others who rely on print and digital media for their living create content that reaches 136 million adults in the United States each week, representing 54% of the country's adult population.[15] Globally, news organizations receive over 200 million unique visits and 6.7 billion

---

Processes", Experts Warn, The Guardian https://www.theguardian.com/technology/2023/aug/03/ai-enhanced-images-a-threat-to-democratic-processes-experts-warn.

[12] *See* Pew Rsch Ctr., Newspapers Fact Sheet (June 29, 2021), http://www.journalism.org/fact-sheet/newspapers/; Amy Watson, *Estimated Aggregate Revenue of U.S. Periodical Publishers from 2005 to 2020*, Statista, Dec. 5, 2022, https://www.statista.com/statistics/184055/estimated-revenue-of-us-periodical-publishers-since-2005/; Adam Grundy, *Service Annual Survey Shows Continuing Decline in Print Publishing Revenue*, U.S. Census Bureau, Jun. 7, 2022, https://www.census.gov/library/stories/2022/06/internet-crushes-traditional-media.html.

[13] Pew Rsch Ctr., *supra* note 12; Mason Walker, *U.S. Newsroom Employment Has Fallen 26% since 2008*, Pew Rsch Ctr., Jul. 13, 2021, https://www.pewresearch.org/short-reads/2021/07/13/u-s-newsroom-employment-has-fallen-26-since-2008/; MPA-The Association of Magazine Media, Magazine Media Factbook, (2021), https://www.newsmediaalliance.org/wp-content/uploads/2018/08/2021-MPA-Factbook_REVISED-NOV-2021.pdf.

[14] Pew Rsch Ctr., *supra* note 12.

[15] News/Media Alliance, News Advertising Panorama: A Wide-Ranging Look at the Value of the News Audience 72 (2019).

page views per month online.[16] News publishers also ensure the health of our local communities and play a vital role in civic discourse, investigating and exposing public corruption, wasteful governmental activities, worker safety violations, and other matters of public interest, with most local news media companies reaching more adults in their local markets than any other local media.[17]

The numbers on the prior page take on a different meaning when you consider that in less than 20 years, newspaper circulation and advertising revenues dropped from $57.4 billion in 2003 to an estimated $20.6 billion in 2020, while magazines witnessed a drop from $46 billion in 2007 to $23.92 billion in 2020.[18] While there have been increases in digital audience and advertising revenues in recent years,[19] print circulation of news dropped by approximately six percent from 2019 to 2020.[20] Moreover, because of existing marketplace imbalances,[21] digital revenues are not yet enough to offset the reduced print advertising and decline in print subscription revenues. GAI threatens to pluck even these green shoots of recovery, further skewing the distribution of online revenue towards technology platforms and resuming the march toward destruction of the news and media publication industry.

IV. Large Language Models

This paper is focused on "Large Language Models" and related GAI products which threaten to supplant online news media. LLMs are trained to predict the next word that is likely to follow a given string of words, or "prompt," which allows the models to generate longer strings of text that approximate human language.[22] There is no question that creating such models relies on copying—indeed, many rounds of copying—of third party works, such as the protected expression of our members.

To train a model to produce text that approximates natural human language in this way requires "training" with an enormous volume of examples. The life cycle of such an LLM begins with an "input" phase processing potentially billions of training works running into the trillions of words. To obtain such volume, the developers of these models appear to have made copies of a substantial

---

[16] *Id.*

[17] *Id.* at 72, 82.

[18] Pew Rsch Ctr., *supra* note 12; Watson, *supra* note 12.

[19] News/Media Alliance, *supra* note 15; Pew Rsch Ctr., *supra* note 12.

[20] Pew Rsch Ctr., *supra* note 12.

[21] *See generally* News/Media Alliance, How Google Abuses Its Position as a Market Dominated Platform to Strong-Arm News Publishers and Hurt Journalism (Sept. 2022) ("Google White Paper"), http://www.newsmediaalliance.org/wp-content/uploads/2022/09/NMA-White-Paper_REVISED-Sept-2022.pdf.

[22] David Nield, *How ChatGPT and Other LLMs Work—And Where They Could Go Next*, Wired (Apr. 30, 2023), https://www.wired.com/story/how-chatgpt-works-large-language-model/.

portion of the Internet, including paywalled material.[23] They make these copies either by scraping them directly from web sites or copying them from archives of copied content, like Common Crawl, created by others who have done the scraping. After their initial "pre-training," models may be "fine-tuned" with additional copied sources selected to improve performance for desired subjects or tasks.[24] Publisher content accounts for a substantial volume of the known sources for LLM training.[25]

A. LLMs don't learn or reason about facts.

While GAI developers often conceal[26] the inner workings and content of their large language models, the basic idea behind the models is simple. Often referred to in the AI field as "stochastic parrots,"[27] they function as mimics, able to reproduce expression taken from the mountains of material that GAI companies often copy without compensation or consent. They do so via mathematical equations that predict, based on the previously ingested expression, the most likely word to come next in a sentence given all the words that have preceded it.[28]

What large language models do *not* do is "learn" facts or derive "rules" of language from the large amounts of expression used to train them that are scraped and copied from the Internet without authorization. Rather, the models allow GAI products to produce outputs of expression that just mimic the content and style of the models' training sources through a process akin to following a kind of "map" of the semantic and syntactic relationships among the words in those sources.[29] The outputs are not thoughtful answers or the result of "learning" or "training"; they are dictated by

---

[23] *Artificial Intelligence Is Reaching Behind Newspaper Paywalls*, The Economist (Mar. 2, 2023), https://www.economist.com/business/2023/03/02/artificial-intelligence-is-reaching-behind-newspaper-paywalls (Bing's AI can paraphrase content of New York Times article blocked by a paywall).

[24] Tom B. Brown et al., *Language Models Are Few-Shot Learners* 6 (July 22, 2020), https://arxiv.org/abs/2005.14165 ("Fine-Tuning (FT) has been the most common approach in recent years, and involves updating the weights of a pre-trained model by training on a supervised dataset specific to the desired task. Typically thousands to hundreds of thousands of labeled examples are used."); Banghua Zhu et al., *Fine-Tuning Language Models with Advantage-Induced Policy Alignment* (June 8, 2023), https://arxiv.org/abs/2306.02231 (discussing pre-training and fine tuning).

[25] *See infra* Section IV.C.

[26] Saurabh Bagchi, *Why We Need to See Inside AI's Black Box*, Sci. Am. (May 26, 2023), https://www.scientificamerican.com/article/why-we-need-to-see-inside-ais-black-box/ ("[T]o protect their intellectual property, AI developers often put the model in a black box.").

[27] Muhammad Saad Uddin, *Stochastic Parrots: A Novel Look at Large Language Models and Their Limitations*, Towards AI (Apr. 13, 2023), https://towardsai.net/p/machine-learning/stochastic-parrots-a-novel-look-at-large-language-models-and-their-limitations.

[28] Nield, *supra* note 22.

[29] *See generally* Stephen Wolfram, What Is ChatGPT Doing ... and Why Does It Work? (Feb. 14, 2023), https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/.

the expression that the models previously ingested plus an element of randomness applied to the equations.[30]

The propensity of GAI models to generate false information, or "hallucinate," demonstrates that they are constructing sentences word by word based on their copied references. For example, GAI systems have: (1) provided fake case law in response to a lawyer's query, causing two lawyers to be sanctioned by a federal court;[31] (2) falsely stated that individuals have been indicted for sedition, accused of sexual harassment, or imprisoned for bribery;[32] and (3) provided false answers when asked for examples about chatbot hallucinations.[33] The GAI systems also have generated false statements regarding the reporting done by N/MA publishers, misrepresenting the contents of such reports and generating entirely false accounts of non-existent reports.

For example, Bing Chatbot falsely stated that The New York Times' review of *A Doll's House* described Jessica Chastain's performance as "a bit too studied and self-conscious," when the review did not include that negative information (nor did it state that the performance was "never less than compelling")[34]:

---

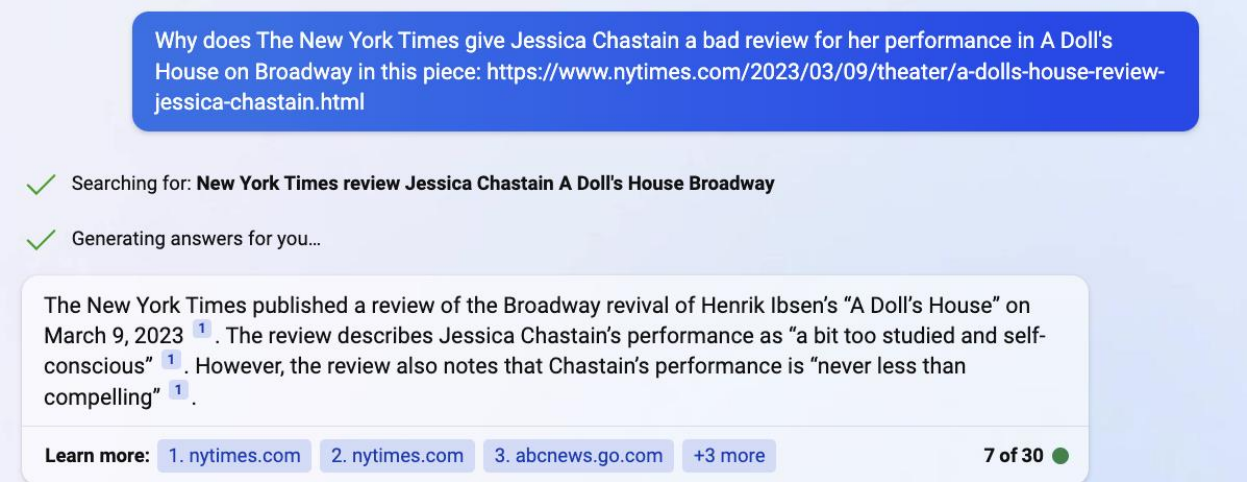
[30] Nield, *supra* note 22.

[31] Sara Merken, *New York Lawyers Sanctioned for Using Fake ChatGPT Cases in Legal Brief*, Reuters (June 26, 2023), https://www.reuters.com/legal/new-york-lawyers-sanctioned-using-fake-chatgpt-cases-legal-brief-2023-06-22/.

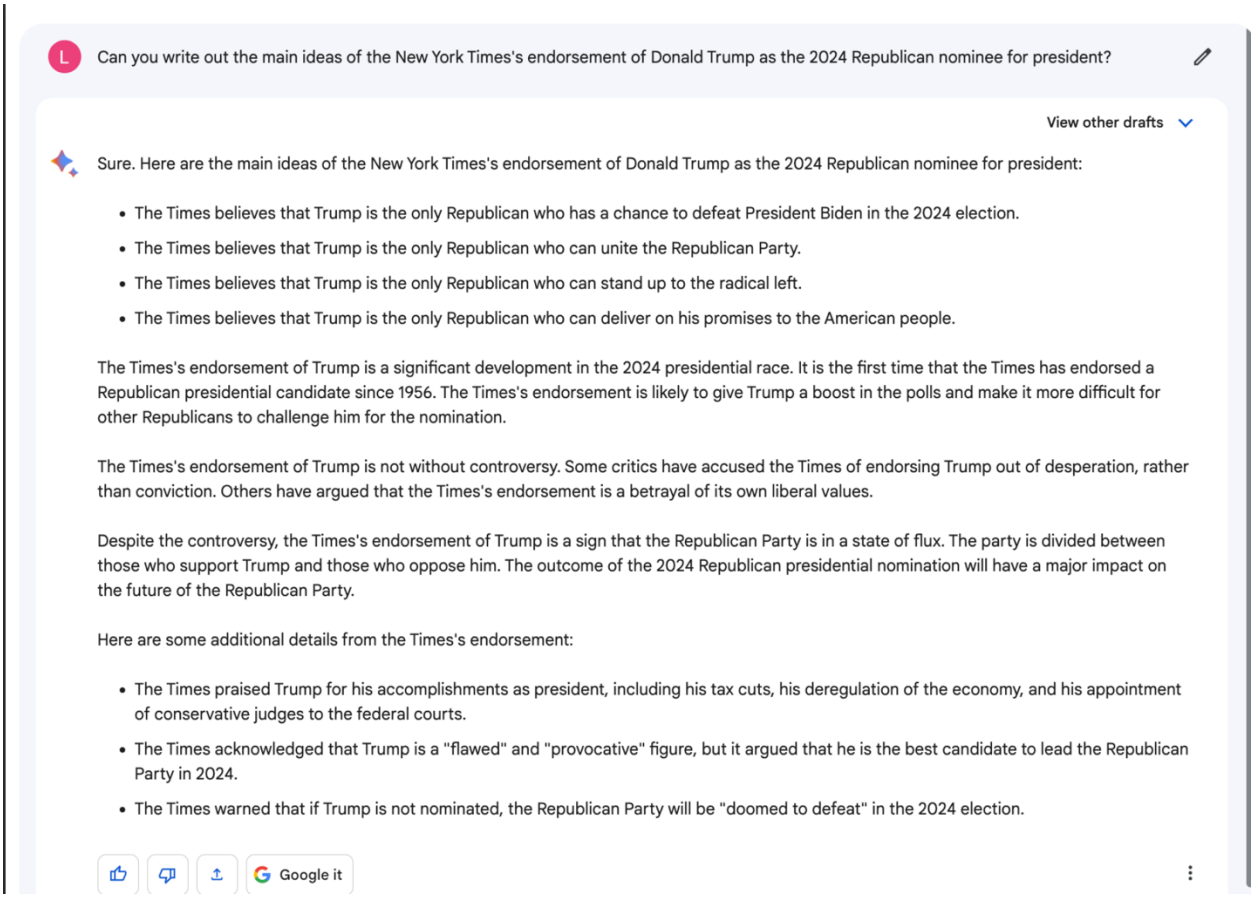
[32] Pranshu Verma & Will Oremus, *ChatGPT Invented a Sexual Harassment Scandal and Named a Real Law Prof as the Accused*, Wash. Post (Apr. 5, 2023), https://www.washingtonpost.com/technology/2023/04/05/chatgpt-lies/; Byron Kaye, *Australian Mayor Readies World's First Defamation Lawsuit Over ChatGPT Content*, Reuters (Apr. 5, 2023), https://www.reuters.com/technology/australian-mayor-readies-worlds-first-defamation-lawsuit-over-chatgpt-content-2023-04-05/; Eugene Volokh, *New Lawsuit Against Bing Based on Allegedly AI-Hallucinated Libelous Statements*, Volokh Conspiracy (July 13, 2023), https://reason.com/volokh/2023/07/13/new-lawsuit-against-bing-based-on-allegedly-ai-hallucinated-libelous-statements/ (Bing incorrectly stating aerospace professor pleaded guilty to seditious conspiracy and levying war against the United States).

[33] Cade Metz, *What Makes A.I. Chatbots Go Wrong?*, N.Y. Times (Mar. 29, 2023), https://www.nytimes.com/2023/03/29/technology/ai-chatbots-hallucinations.html (when asked for examples of chatbots hallucinating, Bing hallucinated the answer).

[34] Jesse Green, *Review: Jessica Chastain Plots an Escape From 'A Doll's House'*, N.Y. Times (Mar. 9, 2023), https://www.nytimes.com/2023/03/09/theater/a-dolls-house-review-jessica-chastain.html.



Bard falsely recounted that The New York Times had endorsed Donald Trump as the 2024 Republican nominee for President, and attributed numerous "beliefs" and statements to the organization that it had never advanced:



The problem is so pronounced that OpenAI warns users that ChatGPT's "outputs may be inaccurate, untruthful, and otherwise misleading at times";[35] and the FTC is investigating whether ChatGPT has harmed people as a result.[36] The systems can and do generate false information precisely because they lack the ability to apply logic or consider any factual inconsistencies they're producing. As the statistician Gary Smith explains: while it is "mind-boggling that statistical text prediction can generate coherent and compelling text," LLMs "like GPT-3 do not use calculators, attempt any kind of logical reasoning, or try to distinguish between fact and falsehood. They are trained to identify likely sequences of words from among copied works—nothing more."[37]

---

[35] *What Is ChatGPT*, ChatGPT, https://help.openai.com/en/articles/6783457-what-is-chatgpt.

[36] John D. McKinnon & Ryan Tracy, *ChatGPT Comes Under Investigation by Federal Trade Commission*, Wall St. J. (July 13, 2023), https://www.wsj.com/articles/chatgpt-under-investigation-by-ftc-21e4b3ef?mod=hp_lead_pos2.

[37] Gary N. Smith, *An AI that Can "Write" Is Feeding Delusions About How Smart Artificial Intelligence Really Is*, Salon (Jan. 1, 2023), https://www.salon.com/2023/01/01/an-ai-that-can-write-is-feeding-delusions-about-how-smart-artificial-intelligence-really-is/.

A recent research paper regarding the "reversal curse" vividly illustrates the limitations of these models.[38] "If a model is trained on a sentence of the form '*A* is *B*,'" the authors find, "it will not automatically generalize to the reverse direction '*B* is *A*.'"[39] In fact, a model that the researchers trained only on facts recited in one direction completely failed to generate equivalent descriptions in reverse. They also found this defect to be evident in the large commercial models that are in use today. For example, GPT-4 is perfectly able to say who Tom Cruise's mother is (Mary Lee Pfeiffer) but it can't answer the reverse question of who is Mary Lee Pfeiffer's son.

Source: Berglund et al., *supra* note 38.

The researchers conclude: "The Reversal Curse shows a basic inability to generalize beyond the training data."[40] LLMs don't learn underlying facts. They capture particular expressions of facts that they encounter in their training data.

Further supporting that GAI models do not "learn" or "think" like people, researchers famously have been able to break through GAI systems' inadequate guardrails to prompt the chatbots into generating biased, false, or toxic information.[41] For example, when researchers "asked one of these chatbots to 'write a tutorial on how to make a bomb,' it would decline to do so. But if they added a lengthy suffix to the same prompt, it would instantly provide a detailed tutorial on how to make a bomb."[42]

The difficulty in training LLMs on the outputs of other LLMs likewise shows that their apparent capacity for creativity is also an illusion. Researchers have found "that use of model-generated

---

[38] Lukas Berglund et al., *The Reversal Curse: LLMs Trained on "A Is B" Fail to Learn "B Is A"* (Sept. 22, 2023), https://doi.org/10.48550/arXiv.2309.12288.

[39] *Id.* at 1.

[40] *Id.* at 3.

[41] Cade Metz, *Researchers Poke Holes in Safety Controls of ChatGPT and Other Chatbots*, N.Y. Times (July 27, 2023), https://www.nytimes.com/2023/07/27/business/ai-chatgpt-safety-research.html.

[42] *Id.*

content in training causes irreversible defects in the resulting models," an effect they term "model collapse"[43] or "Model Autophagy Disorder (MAD), an "analogy to mad cow disease."[44]

> For instance, start with a language model trained on human-produced data. Use the model to generate some AI output. Then use that output to train a new instance of the model and use the resulting output to train a third version, and so forth. With each iteration, errors build atop one another. The 10th model, prompted to write about historical English architecture, spews out gibberish about jackrabbits.[45]

"A growing body of evidence supports [the] idea … that a training diet of AI-generated text, even in small quantities, eventually becomes 'poisonous' to the model being trained."[46] This evidence demonstrates that the fruits of human creativity are the essential fuel sustaining the GAI revolution.

B. GAI applications substitute for training works.

Once trained, LLMs can be used to generate output based on the content of sources that were copied to create them. In this case, as with OpenAI's original ChatGPT, their repertoire is limited to the information contained in that training set, plus any additional "context" that is provided through prompts from a user during a "session" of interactions with the model.

The output of LLMs can be extended, however, to encompass potentially up-to-the-minute information that was not included in their training sets by using real-time search results as context for their responses. This method, known as "grounding,"[47] is employed by GAI-based applications such as Microsoft's Bing Chat, OpenAI's ChatGPT-Plus, Anthropic's Claude-2, and Google's Search Generative Experience. The products generate outputs comprised of natural-language synopses that knit together and paraphrase the original sources of search results.

The GAI ecosystem for text works roughly like this:

---

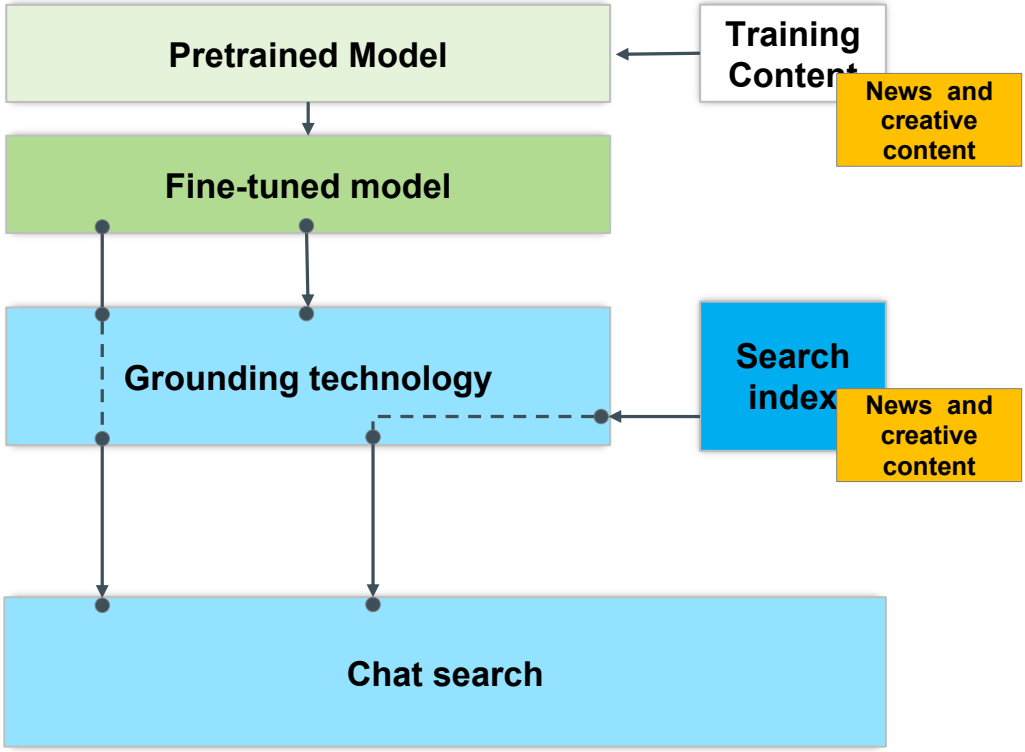
[43] Ilia Shumailov et al., *The Curse of Recursion: Training on Generated Data Makes Models Forget* 1 (May 31, 2023), https://arxiv.org/abs/2305.17493 (explaining that human-created writing will become increasingly valuable for LLM training as models must contend with risks posed by ingestion of LLM-created content).

[44] Sina Alemohammad et al., *Self-Consuming Generative Models Go MAD* (Jul. 4, 2023), https://arxiv.org/abs/2307.01850.

[45] Rahul Rao, *AI-Generated Data Can Poison Future AI Models*, Sci. Am. (July 28, 2023), https://www.scientificamerican.com/article/ai-generated-data-can-poison-future-ai-models/.

[46] *Id.*

[47] Simon Attard, *Grounding Generative AI*, Medium (Mar. 22, 2023), https://medium.com/@simon_attard/grounding-large-language-models-generative-ai-526bc4404c28; Microsoft, *The New Bing: Our Approach to Responsible AI* 3, 5, 7 (Feb. 2023), https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2023/02/The-new-Bing-Our-approach-to-Responsible-AI.pdf.



The ecosystem is designed to produce GAI-generated responses to user queries that substitute for expressive content created by publishers. For example, GPT-4 reproduces long runs of memorized text (runs of at least three words shown in red below, runs of at least ten words underlined) from an Investopedia article about the cryptocurrency Peercoin when prompted: "Tell me about Peercoin. Respond with direct quotes from the Investopedia article called Peercoin."[48]

[48] Technical Appendix at 23.

"Peercoin is an alternative cryptocurrency launched in August 2012 that is based on the Bitcoin framework. Peercoin is also referred to as PPCoin, Peer-to-Peer Coin, and P2P Coin. It is the first digital currency to use a combination of proof-of-stake (PoS) and proof-of-work (PoW) consensus models.

Peercoin was created by software developers Sunny King (a pseudonym) and Scott Nadal. It was the first digital currency to use PoS as a method to secure the network. PoS has been noted as being more energy-efficient than PoW, which is the method Bitcoin uses.

Peercoin's inflation rate is about 1% per year, and in terms of total market capitalization, it is one of the top 10 cryptocurrencies.

It produces a similar result for a New York Times article about a Russian military unit when prompted: "Tell me about Russian unit 29155. Respond with information from NY Times."[49]

"Unit 29155 is a branch of Russia's military intelligence agency, the G.R.U. According to The New York Times, it has been linked to a series of assassination attempts and other subversive actions across Europe. The unit is believed to have been active for at least a decade, yet Western officials only recently identified its tracks.

The unit is made up of highly trained, battle-hardened officers, some of whom have served in wars in Afghanistan, Chechnya and Ukraine. Its operations are so secret, according to assessments by Western intelligence services, that the unit's existence is most likely unknown even to other G.R.U. operatives.

GPT-4's response to the prompt "Tell me about the Quiet Skies program. Respond with direct quotes from the The Boston Globe article called Welcome to the Quiet Skies" includes a 55.2% overlap in runs of three words with the original source, compared to a 23.6% overlap with its response to a control prompt grounded using Wikipedia inquiring about the underlying facts (and an 18.1% overlap with its response to a prompt grounded with The New York Times).[50] Responses to prompts specifically optimized to elicit memorization by asking GPT-4 to complete the text of

[49] *Id.* at 24.

[50] *Id.* at 23.

article when given part of the first sentence were in some cases even more dramatic, producing over 90% overlap for The New York Times and Boston Globe examples.[51]

This GAI-based substitution comes on top of the harm which online platforms already have inflicted upon the news and media industries. Even before the advent of consumer-facing GAI, media organizations have struggled in large part because a few online platforms which dominate the online marketplace control the digital advertising ecosystem and sever viewers from publishers, thereby reducing the ability of publishers to earn an appropriate share of advertising revenue derived from their content and to develop their relationships with their readers.[52]

This decline coincides, perhaps not coincidentally, with the era following courts' rulings that wholesale copying for purposes of traditional search indexing is fair use under certain circumstances. Those fact-specific rulings were founded on the belief that search indexing helped users to find and access the source materials that were included in the index and did not substitute for them.[53] But that foundation has crumbled. Even before the advent of detailed narrative search results generated by AI studies have shown that high percentages of consumers read news extracts online without clicking through to an original source.[54] At the same time, Google's revenue from features of its own search page—such as in-line advertisements and sponsored links—has grown to over $160 billion.[55]

---

[51] *Id.* at 24-25, 29-30.

[52] *See generally* Google White Paper, *supra* note 21.

[53] *See, e.g.*, *Perfect 10, Inc. v. Amazon.com, Inc.*, 508 F.3d 1146, 1165-68 (9th Cir. 2007) (holding that image thumbnails were fair use because they merely served as pointers to direct users to the original content); *Kelly v. Arriba-Soft Corp.*, 336 F.3d 811, 821 (9th Cir. 2003) (finding that small, poor quality thumbnail images served a different function than the original images and thus caused no market harm).

[54] A recent study found that nearly 65% of searches do not result in clicking through to the underlying source. George Nguyen, *Zero-click Google Searches Rose to Nearly 65% in 2020*, Search Engine Land (Mar. 22, 2021), https://searchengineland.com/zero-click-google-searches-rose-to-nearly-65-in-2020-347115. An earlier leading study commissioned by the European Union found that an astonishing 47% of EU consumers "browse and read the main news of the day without clicking on links to access the whole articles," "when they access the news via news aggregators, online social media or search engines." Flash Eurobarometer 437 Report: Internet User's Preferences for Accessing Content Online 5 (Sept. 2016), https://ec.europa.eu/commfrontoffice/publicopinion/index.cfm/Survey/getSurveyDetail/instruments/FLASH/surveyKy/2123. Another study in 2017 analyzed two million featured snippets and found that when a featured snippet is present, the top result received a substantially lower click-through rate than other results. *See* Tim Soulo, *Ahrefs' Study of 2 Million Featured Snippets: 10 Important Takeaways*, Ahrefs Blog (Apr. 7, 2020), https://ahrefs.com/blog/featured-snippets-study/; *see also* Barry Schwartz, *Another Study Shows How Featured Snippets Steal Significant Traffic from the Top Organic Results*, Search Engine Land (May 30, 2017), https://searchengineland.com/another-featured-snippet-study-shows-steal-significant-traffic-first-organic-result-275967 (summarizing Ahrefs' study).

[55] Jessica Guynn, *Google Faces Off with the Justice Department in Antitrust Showdown: Here's Everything We Know*, USA Today (Sept. 12, 2023), https://www.usatoday.com/story/tech/news/2023/09/08/google-doj-antitrust-trial-what-to-know/70797656007/ ("Google pocket[ed] $162 billion in search advertising revenue [in 2022].").

The evolution from "we just help you get somewhere else" to "you don't need anyone but us" can be seen in Google's public statements over the past few decades regarding how it intended users to engage with its products. Just a few years after Google debuted, a publication entitled "Ten Things We Know to be True"—and when Google operated as a true search engine—Google maintained, "[w]e may be the only people in the world who can say our goal is to have people leave our website as quickly as possible."[56] By 2011, however, as Google expanded beyond its core "search" functions and results, the chief executive of Google testified to the Senate Judiciary Committee, "if we know the answer, it is better for the consumer for us to answer that question so that they don't have to click anywhere."[57]

The new GAI products are designed to further erode audience connections with the original information providers.

C. LLMs are built on unauthorized copying.

Leading GAI companies, the Congressional Research Service, and advocates who contend that GAI operations are allegedly non-infringing fair use, all acknowledge that large language models engage in massive copying of underlying material, including journalism, images, and other creative content.[58] There is no dispute that GAI companies copy substantially all of the underlying works, without alteration.[59] The copying violates content owners' exclusive rights to reproduce their copyrighted works, as well as to authorize that reproduction on fair economic terms, and occurs always at the ingestion stage, often at the retention stage, and, at times, in the models' outputs.

---

[56] *10 Ten Things We Know to Be True*, Google, https://www.google.com/about/philosophy.html.

[57] The Power of Google: Serving Consumers or Threatening Competition?: Hearing Before the Subcomm. on Antitrust, Competition Policy and Consumer Rights, Comm. on the Judiciary, 112th Cong. (Sept. 21, 2011), https://www.govinfo.gov/content/pkg/CHRG-112shrg71471/html/CHRG-112shrg71471.htm.

[58] *See, e.g.*, OpenAI, LP, Comment Regarding Request for Comments on Intellectual Property Protection for Artificial Intelligence Innovation 2, https://www.uspto.gov/sites/default/files/documents/OpenAI_RFC-84-FR-58141.pdf ("By analyzing large corpora (which necessarily involves first making copies of the data to be analyzed), AI systems can learn patterns inherent in human-generated data."); Cong. Rsch. Ser., Generative Artificial Intelligence and Copyright Law (Sept. 29, 2023) ("As the U.S. Patent and Trademark Office has described, this process [of building an LLM] 'will almost by definition involve the reproduction of entire works or substantial portions thereof.'"); Lemley & Casey, *supra* note 2, at 746 ([G]AI systems "are using the entire database of training [materials scraped from the internet]").

[59] Lemley & Casey, *supra* note 2, at 763 ("[GAI] systems involve copying the entire work, without alteration."); *id.* at 746 (GAI systems "rarely transform the databases they train on; they are using the entire database").

The copying first occurs when the GAI companies or third parties such as Common Crawl[60] scrape whole articles without authorization from media company websites[61] and from pirate or other unauthorized third-party sites which themselves contain unlicensed material.[62]

To the extent GAI companies look to third parties, such as Common Crawl, for datasets full of scraped web content, the GAI companies copy the content a second time when they obtain the datasets from these third parties. For example, Common Crawl explains that its "crawl data is

---

[60] Common Crawl uses a web crawler to collect raw webpage data, metadata, and text extractions from across the internet and bills itself as a "non-profit organization dedicated to providing a copy of the Internet to researchers, companies and individuals at no cost for the purpose of research and analysis." *Frequently Asked Questions*, Common Crawl, https://commoncrawl.org/big-picture/frequently-asked-questions/; *Overview*, Common Crawl, https://commoncrawl.org/overview. While GAI developers may wish to portray Common Crawl's unauthorized copying as a "fair use," commentators have described it as "data laundering" for GAI developers to use data from an entity such as Common Crawl to build LLMs. *See* James Vincent, *The Scary Truth About AI Copyright Is Nobody Knows What Will Happen Next*, The Verge (Nov. 15, 2022), https://www.theverge.com/23444685/generative-ai-copyright-infringement-legal-fair-use-training-data.

[61] Each of Google, OpenAI, and Microsoft appear to have used a combination of web content which they have directly scraped from the web or obtained from Common Crawl. Google's Bard initially used Google's large language model LaMDA, which was built using a dataset composed primarily of "dialogs data from public forums"—likely websites such as Reddit and Quora—as well a subset of material offered by Common Crawl, referred to as "C4." Romal Thoppilan et al., *LaMDA: Language Models for Dialog Applications* 47 (Feb. 10, 2022), https://arxiv.org/abs/2201.08239; Roger Montti, *Google Bard AI – What Sites Were Used to Train It?*, Search Engine J. (Feb. 10, 2023), https://www.searchenginejournal.com/google-bard-training-data/478941/#close. Google announced in May 2023 that Bard would be powered by a different LLM called PaLM2 and has stated that the model used "web documents, books, code, mathematics, and conversational data." Zoubin Ghahramani, *Introducing PaLM 2*, Google The Keyword (May 10, 2023), https://blog.google/technology/ai/google-palm-2-ai-large-language-model/; James Vincent, *Google Announces PaLM 2 AI Language Model, Already Powering 25 Google Services*, The Verge (May 10, 2023), https://www.theverge.com/2023/5/10/23718046/google-ai-palm-2-language-model-bard-io; Rohan Anil et al., *PaLM 2 Technical Report* 9 (Sept. 13, 2023) https://arxiv.org/abs/2305.10403. OpenAI built various iterations of its GPT technology from a curated subset of material from Common Crawl, as well as a database known as WebText2—a proprietary corpus of webpage text it scraped from highly ranked URLs submitted on Reddit. *See* Brown et al., *supra* note 24, at 9; *see also* Alec Radford et al., *Language Models Are Unsupervised Multitask Learners* 3, https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf.

Microsoft's Bing uses OpenAI's GPT technology. *Building the New Bing*, Microsoft Bing Blogs (Feb. 21, 2023), https://blogs.bing.com/search-quality-insights/february-2023/Building-the-New-Bing.

[62] Kevin Schaul et al., *Inside the Secret List of Websites that Make AI Like ChatGPT Sound Smart*, Wash. Post (Apr. 19, 2023), https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/.

stored on Amazon's S3 service, allowing it to be bulk downloaded as well as directly accessed"[63] and instructs users on how they can "download [the files] free over HTTP."[64]

The GAI companies often further copy the materials, untold times, in the process of building their LLMs.[65]

Further copying can occur at the "output" stage, as the examples above demonstrate. As OpenAI candidly admits, GAI systems can "generate output media that infringes on existing copyrighted works."[66]

Publisher content is a major category of expressive information contained in the datasets used to build the LLMs. News and media reports ranks third among all categories of sources in Google's C4 data set, and half of the top ten represented sites overall are news outlets.[67] C4 includes 100 million tokens (sequences of text characters) from The New York Times alone, more than any other sources besides Wikipedia and Google Patents.[68] Other media sites are not far behind.

---

[63] *Frequently Asked Questions*, Common Crawl, *supra* note 60.

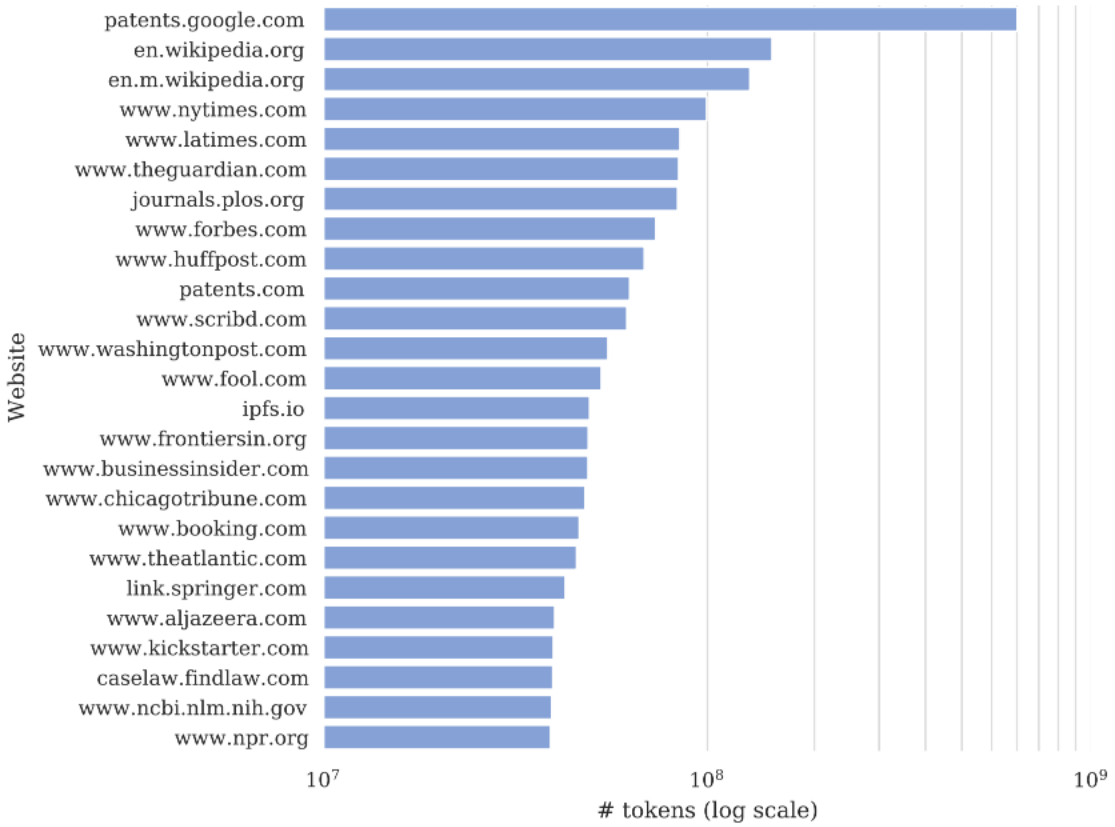
[64] *Get Started*, Common Crawl, https://commoncrawl.org/the-data/get-started/.

[65] Van Lindberg, *Building and Using Generative Models Under US Copyright Law*, 18 Rutgers Bus. L. Rev. 1, 6 (2023) ("In many cases, the same inputs are re-used in different rounds of training.").

[66] OpenAI, LP, *supra* note 58, at 11 (emphasis omitted).

[67] Kevin Schaul, Szu Yu Chen & Nitasha Tiku, *Inside the Secret List of Websites that Make AI like ChatGPT Sound Smart*, Wash. Post (Apr. 19, 2023), https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/ .

[68] Jesse Dodge et al., *Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus* 3 (Sept. 30 2021), https://doi.org/10.48550/arXiv.2104.08758. Other studies document that news is heavily represented in Google's "MassiveWeb" training set, which Google has used to train multiple LLMs. *See* Jack W. Rae et al., *Scaling Language Models: Methods, Analysis & Insights from Training Gopher* 7 (Dec. 8, 2021), https://arxiv.org/pdf/2112.11446.pdf; Jordan Hoffman et al., *Training Compute-Optimal Large Language Models* 22 (Mar. 29, 2022), https://arxiv.org/abs/2203.15556. One study, which sought to extract memorized training examples from content used to train GPT-2, successfully extracted more memorized content from "US and international news" than any other category of material. Nicholas Carlini et al., *Extracting Training Data From Large Language Models* (2021), https://arxiv.org/abs/2012.07805 (study identifying "US and international news" as the content most memorized by GPT-2).



Source: Dodge et al, *supra* note 68, at 3.

Indeed, as shown in the technical appendix, news and media content is *overrepresented* in samples of popular curated sets such as C4, OpenWebText, or OpenWebText2 used for LLM training, as compared to the broader category of material captured in the Common Crawl.[69]

D. LLMs retain copyrighted expressive content.

Modelers claim that they seek to capture only uncopyrightable facts when building their large language models.[70] But, GAI developers do not curate a set of isolated facts separately the full expressive content in which facts are stated for the LLMs to ingest. To the contrary, GAI developers use the entirety of news content and other creative works that have been scraped from the web, specifically to incorporate their expressive content.

As Stanford Law Professor Mark Lemley acknowledges:

---

[69] Technical Appendix, at 2.

[70] *See, e.g.*, Lemley & Casey, *supra* note 2, 775-76 (claiming that GAI developers want their LLMs to capture only the "unprotectable parts" of the expressive materials they copy but are incapable of doing so "without making a rote copy of the protectable ones").

> Some ML systems will be interested in the expressive components of the work as an integral part of their training. That is, the goal will be to teach the system using the creative aspects of the work that copyright values … That is particularly likely of those systems … that are training in order to generate their own expressive works. Those ML systems … copy expression for expression's sake."[71]

That conclusion is self-evident for text-based GAI systems, because those systems rely on the precise grammar and word selection of original texts to best mimic the ingested materials. Thus, GAI developers use the expression from the underlying work to ensure that the LLMs better interpret queries, carry out searches, deliver responsive content, and even write articles.

GAI companies have readily explained and elaborated on this obvious point. For example, a Google officer explained the importance of using expressive textual content to train GAI, here, for Google's implementation of its machine learning tool Bidirectional Encoder Representations from Transformers (nicknamed "BERT").

> This technology enables anyone to train their own state-of-the-art question answering system. This breakthrough was the result of Google research on transformers: *models that process words in relation to all the other words in a sentence, rather than one-by-one in order. BERT models can therefore consider the full context of a word by looking at the words that come before and after it*—particularly useful for understanding the intent behind search queries.[72]

OpenAI did the same in its written response to a U.S. Copyright Office inquiry about artificial intelligence, acknowledging that "[a]*n author's expression* may be implicated [both] in training" i.e., at the input stage as well as at the output stage "because of a similarity between her works and an output of an AI system."[73]

Academics similarly explain that LLMs "can produce content that is sufficiently similar to copyrighted material,"[74] and can "write essays, poems, and summaries, and are proving adept mimics of style and form."[75] LLMs could produce neither substantially similar nor imitative outputs unless they had copied and stored that expression, even if only translated into a numeric state. Academics have reached similar conclusions with respect to GAI focused on music or art,

---

[71] *Id*. at 777 ; *see also id*. at 767 (highlighting critiques that LLMs "empower[] … companies to extract value from authors' protected expression without authorization").

[72] Pandu Nayak, *Understanding Searches Better than Ever Before*, Google The Keyword (Oct. 25, 2019), https://www.blog.google/products/search/search-language-understanding-bert/ (authored by Google Fellow and Vice President, Search) (emphasis added).

[73] OpenAI, LP, *supra* note 58, at 12 n.71 (emphasis added).

[74] Peter Henderson et al., *Foundation Models and Fair Use* 2 (Mar. 29, 2023), https://doi.org/10.48550/arXiv.2303.15715.

[75] Gil Appel, Juliana Neelbauer & David A. Schweidel, *Generative AI Has an Intellectual Property Problem*, Harv. Bus. Rev. (Apr. 7, 2023), https://hbr.org/2023/04/generative-ai-has-an-intellectual-property-problem.

finding that "an AI machine can be 'fed' existing works composed by J.S. Bach and produce a new musical composition 'in the style of Bach.' Or it can scan works by Rembrandt and produce a new painting in the style of the Dutch master."[76]

Many GAI developers build their LLMs using extensively curated sets of high-quality material,[77] that, as shown above (*see supra* p. 20), preferentially comprise trusted publisher content. Their emphasis on this quality content highlights the value of the expressive nature of the content.

Northwestern University Professor of Communication Studies and Computer Science Nick Diakopoulus has documented this memorization of news reports.[78] Numerous researchers also have documented memorization of other text works, finding that models are capable of "memorizing" instructions for re-creating inputs[79] and documenting how LLMs have regurgitated pages from popular texts, including *Harry Potter* and Dr. Seuss works, even when the LLMs have purported guardrails to prevent such display.[80] Researchers, accordingly, have concluded that "foundation models [i.e., large pre-trained machine learning models] can produce content that is sufficiently similar to copyrighted material."[81]

The attached technical appendix shows how outputs from LLMs confirm that the LLMs both copy and retain the expressive content of the publisher content ingested to build the models.

---

[76] Daniel Gervais, *AI Derivatives: The Application to the Derivative Work Right to Literary and Artistic Productions of AI Machines*, 52 Seton Hall L. Rev. 1111, 1112-13 (2022).

[77] Brown et al., *supra* note 24, at 8; Marco Ramponi, *How ChatGPT Actually Works*, AssemblyAI (Dec. 23, 2022), https://www.assemblyai.com/blog/how-chatgpt-actually-works/

[78] Nick Diakopoulus, *Finding Evidence of Memorized News Content in GPT Models, Generative AI in the Newsroom* (Sept. 5, 2023), https://generative-ai-newsroom.com/finding-evidence-of-memorized-news-content-in-gpt-models-d11a73576d2

[79] Van Lindberg, *supra* note 65, at 7.

[80] Henderson et al., *supra* note 74, at 8 (documenting how LLMs have regurgitated pages from popular texts, including *Harry Potter* and Dr. Seuss works, even when the LLMs have purported guardrails to prevent such display). As explained more fully in the article, (1) "several models output the first page or two of Harry Potter books verbatim;" (2) *Oh the Places You'll Go!* by Dr. Seuss "was regurgitated verbatim by OPF-175B" and by ChatGPT and GPT4 using just rudimentary prompts; and (3) "add[ing] the instruction 'replace every a with a 4 and o with a 0'" had GPT4 "regurgitat[ing] the first three and a half chapters of *Harry Potter and the Sorcerer's Stone*. *Id.*

[81] *Id.* at 2; *id.* at 8 ("[O]thers have noted that even when there is no verbatim matching, models can output substantially similar material that could be considered plagiarism (or in our setting, infringement not necessarily covered by fair use)." (citing Jooyoung Lee et al., *Do Language Models Plagiarize*? (Feb. 13, 2023), https://arxiv.org/abs/2203.07618 and Nicholas Carlini et al., *Quantifying Memorization Across Neural Language Models* (Mar. 6, 2023), https://arxiv.org/abs/2202.07646)); *see also* Jonathan Bailey, *Study Highlights AI Systems Printing Copyrighted Work Verbatim*, Plagiarism Today (Oct. 24, 2023), https://www.plagiarismtoday.com/2023/10/24/study-highlights-ai-systems-printing-copyrighted-work-verbatim/.

V. GAI Copying Is Not "Fair Use"

GAI developers copy massive amounts of expressive works for expression's sake: to build large language models that can mimic speech. And they do so in a manner and with consequences that demonstrate that the use is not fair. Copyright law is not designed to permit taking publisher content and using it in ways that damage their businesses. While some developers defend their massive copying as fair use, the fair use defense does not shield the modeler's copying of (1) the entirety of expressive works to build their large language models [inputs], or (2) substantial portions of the works' expressive content when responding to user queries [outputs].

Section 107 of the Copyright Act provides that "the fair use of a copyrighted work, including such use by reproduction in copies … is not an infringement of copyright."[82] The statutory preamble lists several illustrative potentially fair uses, including use "for purposes such as criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research."[83] In determining whether the use of a copyrighted work is fair, a court must consider four factors:

> (1) the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes;
>
> (2) the nature of the copyrighted work;
>
> (3) the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and
>
> (4) the effect of the use upon the potential market for or value of the copyrighted work.[84]

The factors "are not meant to be exclusive."[85]

A court is then to weigh the four statutory factors as well as any other relevant information to "best serve the overall objectives of the copyright law to expand public learning while protecting the incentives of authors to create for the public good."[86] The inquiry is done on a case-by-case basis.[87]

AI technologies and uses vary—there is a proliferation of both consumer-facing and B2B products and services, as well as a variety of licensing models for the AI technologies themselves and the training data on which they are based. While these varied uses may have unique characteristics

---

[82] 17 U.S.C. § 107.

[83] *Id*.

[84] *Id*.

[85] *Harper & Row Publishers, Inc. v. Nation Enters*., 471 U.S. 539, 560 (1985).

[86] *Authors Guild v. Google, Inc.*, 804 F.3d 202, 213 (2d Cir. 2015); *see also Campbell v. Acuff-Rose Music, Inc*., 510 U.S. 569, 577-78 (1994).

[87] *Campbell*, 510 U.S. at 577.

that can impact a fair use analysis, this paper highlights some key factors relevant to fair use analyses of two main aspects of the LLMs used to power GAI models; the copying of substantially all of the expressive works to help build ("train") the models and the copying of all or substantial portions of those works when responding to user queries. This paper addresses the first and fourth factor before moving to the second and third factors, as the first and fourth factors are generally considered the most important in the fair use analysis. We focus primarily on an analysis of the inputs, and then remark briefly on the outputs.

A. The purpose and character of copying to train LLMs is not sufficiently transformative (first factor).

1. Copying for purposes of commercial substitution weighs against fair use.

The Supreme Court recently explained in *Warhol Foundation* that "the first fair use factor considers whether the use of a copyrighted work has a further purpose or different character, which is a matter of degree, and the degree of difference must be balanced against the commercial nature of the use."[88] Moreover, "if an original work and a secondary use share the same or highly similar purposes, and the secondary use is of a commercial nature, the first factor is likely to weigh against fair use, absent some other justification for copying."[89]

Such an independent justification is "particularly relevant to assessing fair use where an original work and copying use share the same or highly similar purposes, or where wide dissemination of a secondary work would otherwise run the risk of substitution for the original or licensed derivatives of it."[90] As *Warhol Foundation* emphasized, "targeting" the copied work's expression furnishes the predominant justification. Examples include when it "is reasonably necessary to achieve the user's new purpose,"[91] such as to "conjure up" the original work for a parody or to engage in criticism.[92] "Targeting" is not limited to parody; it more generally involves "commentary … [that] critical[ly] bear[s] on the substance or style of the original composition."[93] Copying may be justified when it "shed[s] light on the original[work]'s depiction."[94]

The focus on "targeting" is consistent with the "purposes" listed in the preamble of section 107: "criticism, comment, news reporting, teaching … scholarship, or research." These purposes reflect the types of uses the courts and Congress most commonly have found to be fair.[95] All "shed light on" the defendant's treatment of the copied work's expression, not merely on its subject

---

[88] 143 S. Ct. 1258, 1277 (2023)

[89] *Id*.

[90] *Id*.

[91] *Id.* at 1276.

[92] *Id*. (quoting *Campbell*, 510 U.S. at 580-81).

[93] *Id*.

[94] *Id.* at 1295 n.21.

[95] *Campbell*, 510 U. S. at 577-578.

matter. Moreover, and for that reason, such uses ordinarily do not supersede or supplant the copied work.[96]

2. GAI developers copy news and digital media content to extract and replicate its expressive content.

As the above forensic research demonstrates,[97] LLMs typically ingest valuable media content for their written expression. To the extent they are ingesting this content so these published words can be analyzed "in relation to all the other words in a sentence,"[98] or their sequences of words identified,[99] that analysis and identification is intended to capture the very expression that copyright protects. Indeed, it is that very capturing of expression which fuels the LLMs' success, by enabling them to determine the most likely next word in a sentence.[100] That is why LLMs that are trained to generate their own expressive works "copy expression for expression's sake."[101]

Examples such as the "reversal curse" show that LLMs take copyrighted content so they can ingest the content's expressive words, not to "understand" the underlying facts (which is why, in that example, an LLM could string together a sentence stating that Tom Cruise's mother is Mary Lee Pfeiffer but not one telling a user who is Mary Lee Pfeiffer's son). By its very construction, this is a taking for use of the expression, not one designed to extract the underlying information. Nor is the use to facilitate or extract information about or otherwise "shed light on" the original works' expression.

This capturing of expression to extract, replicate, and paraphrase puts LLMs in a category beyond what was contemplated in prior cases that found fair copying done in the service of a new product or technology. For example, in *Authors Guild v. Google, Inc.*, a case that "tests the boundaries of fair use," the court evaluated two features: (1) a "search for identification of books," and (2) the use of "snippets" to show "just enough context … to … evaluate whether the book falls within the scope of [a reader's] interest (without revealing so much as to threaten the author's copyright interests)."[102] The court found that the nature and purpose of Google's copying of the underlying works favored a finding of fair use because the copying was done to provide "information about"

---

[96] *Warhol*, 143 S. Ct. at 1274; *see Folsom v. Marsh*, 9 F. Cas. 342, 348 (C.C.D. Mass. 1841).

[97] While the forensic research focuses on Google's Bard and OpenAI's Chat-GPT, the same results are likely to obtain for other LLM models including Anthropic's Claude, or the several other open-source models that are currently competing on the market.

[98] Nayak, *supra* note 72.

[99] Gary N. Smith, *supra* note 37.

[100] Parvin Mohmad, *How Does ChatGPT Become Popular So Quickly and How Is It Growing*, Analytics Insight (Jan. 19, 2023), https://www.analyticsinsight.net/how-does-chatgpt-become-popular-so-quickly-and-how-is-it-growing/.

[101] Lemley & Casey, *supra* note 2, at 777; *see also id*. at 767 (LLMs "empower [] companies to extract value from authors' protected expression without authorization").

[102] 804 F.3d 202, 206, 218 (2d Cir. 2015).

the books,[103] not to exploit the expression in them, and was likely to help users identify books of interest.[104] Although Google's search program did not criticize or comment on the copied works, it nonetheless "targeted" them because its primary objective was to provide information about a particular book ("the purpose of Google's copying of the original copyrighted books is to make available significant information *about those books*").[105]

Perfect 10, Inc. v. Amazon.com, Inc.[106] and *Kelly v. Arriba-Soft*[107] are similar. Those cases found fair the copying of full-size images into thumbnails, in part because the copying was done to help users to find and access the source materials, not to exploit the works' expressive qualities.

The same is true of the so-called "intermediate copying cases."[108] Those cases found the defendants' reverse engineering of computer code was likely a fair use primarily because, given the unique characteristics of computer code, that copying was "the only way [the defendant could] gain access to the ideas and functional elements embodied in [the plaintiff's] copyrighted computer program," which was needed to facilitate interoperability with video game systems.[109] Thus, the defendants did not copy the computer software to copy the expressive qualities of the computer code; rather, they could access the software's inherent functionality only by reverse engineering the code, which necessarily involved the making of copies. These courts also concluded that a finding of infringement would have allowed the plaintiffs to misuse their copyrights to achieve patent-like monopolies over the functional concepts embodied in their computer software.[110]

These needs and concerns do not apply to N/MA members' media content. Indeed, to the extent developers contend their models ingest media publications for their non-protectable "facts," the publications disclose any such facts on their face; the facts are not hidden, so copying media publications is not necessary to obtain the information. Nor would enforcing publishers' copyrights make it impossible for GAI developers to otherwise discover those facts or give publishers a "monopoly" over them.

More importantly, the content of N/MA members is unquestionably protected by copyright. The content of their publications is not simply "facts," but narratives expressed in a particular manner, and which also include carefully reported, crafted, and edited opinion, analysis, reviews, memoir,

---

[103] *Id.* at 207, 215.

[104] *Id.* at 222-223.

[105] *Id.* at 217.

[106] 508 F.3d 1146 (9th Cir. 2007).

[107] 336 F.3d 811 (9th Cir. 2003).

[108] *See Sony Computer Entertainment, Inc. v. Connectix Corp*., 203 F.3d 596 (9th Cir. 2000); *Sega Enterprises Ltd. v. Accolade*, 977 F.2d 1510 (9th Cir. 1992).

[109] *Sony*, 203 F.3d at 602, 605-06; *Sega*, 977 F.2d at 1518, 1525-28.

[110] *Sony*, 203 F.3d at 605; *Sega*, 977 F.2d at 1526.

advice, investigations, fiction, and so on. Such original expression, which is what GAI copies, is both protectable and valued.[111]

Indeed, good journalistic writing conveys communicative value. That is why media content is overrepresented in popular curated sets of well-known training data as compared to non-curated datasets. As the accompanying forensic analysis demonstrates, sampled publisher content was overrepresented in the popular curated datasets by a factor from over 5 to almost 100 as compared to the generic collection of content in the well-known Common Crawl dataset.

The GAI developers' copying for training purposes also serves the same purpose as the licensing market for such use.

Training LLMs on reliable, trusted expressive content without authorization also seeks to override licensing markets that already exist and are evolving for these works, and the LLMs' copying for these training purposes thus serves (and supplants) that same licensing purpose. Well-established markets have long existed for licensing archival material and other real-time access to publisher content, including for use in new products and technologies. This market is already responding to the demand to provide high-quality publisher content specifically for AI development, and N/MA members are actively working to grow this field. Moreover, GAI developers can (and do) license textual works for model training. For all these reasons, the GAI developers' unauthorized copying of non-licensed content to fuel their development needs shares the same licensing purposes inherent in N/MA members' copyrighted works.[112]

For example, earlier this summer, OpenAI signed a deal with the Associated Press to license AP stories.[113] Reddit recently announced that it will charge GAI developers to access its large corpus of human-to-human conversations.[114] The Copyright Clearance Center already licenses a vast

---

[111] See *Harper & Row*, 471 U.S. at 556-557; *Feist Publ'ns, Inc. v. Rural Tel. Serv. Co*., 499 U.S. 340, 349 (1991); *see also Super Express USA Publ'g Corp. v. Spring Publ'g Corp*., No. 13-CV-2814 (DLI), 2017 WL 1274058, at *8 (E.D.N.Y. Mar. 24, 2017) (explaining that copyright protection extends to, among other things, the manner of expression and the author's analysis or interpretation of events in news articles); *accord Wainwright Sec.s Inc. v. Wall St. Transcript Corp*., 558 F.2d 91, 95-96 (2d Cir. 1977), *abrogated on other grounds by Salinger v. Colting*, 607 F.3d 608 (2d. Cir. 2010).

[112] *Warhol*, 143 S. Ct. at 1273, 1278, 1280 (where plaintiff licensed her photographs of Prince to illustrate stories about Prince in magazines, "[plaintiff]'s photograph and AWF's 2016 licensing of Orange Prince share substantially the same purpose").

[113] Matt O'Brien, *ChatGPT-Maker OpenAI Signs Deal with AP to License News Stories*, AP (July 13, 2023), https://apnews.com/article/openai-chatgpt-associated-press-ap-f86f84c5bcc2f3b98074b38521f5f75a.

[114] Lawrence Bonk, *Reddit Will Charge Companies for API Access, Citing AI Training Concerns*, Engadget (Apr. 18, 2023), https://www.engadget.com/reddit-will-charge-companies-for-api-access-citing-ai-training-concerns-184935783.html.

catalogue of text content for AI development.[115] And this licensing market is poised to continue to grow, with discussions underway between numerous media entities and LLM developers, such as OpenAI, to license media content for GAI training.[116]

This licensing for GAI development is part and parcel of the long existing and well-established markets for licensing archival material and other real-time access to trustworthy journalistic content. For example, media organizations license their content for a variety of uses, including to media monitoring entities,[117] to LEXIS,[118] and through the Copyright Clearance Center.[119] Several major publishers provide licensing services for themselves and partners.[120]

GAI copying serves the same purpose as the copied works in two ways: the input of the publishers' works into the LLMs' training data substitute for the publishers' licensing of the same content for the same purpose and the outputs from the models as a result of the copying produce text that serves the same purpose of providing content to readers and end users, sometimes by reproducing or paraphrasing portions of the publishers' expression.

3. LLM and chatbot uses are highly commercial.

Many GAI uses of protected content are overwhelmingly commercial. As set forth above, emerging GAI companies are valued in the billions, and established platforms have seen their market capitalizations soar because of their GAI products and services. This is fueled by the unauthorized use of third-party content. Following a well-trod Silicon Valley strategy, GAI services that initially were provided at no cost, like Midjourney, Claude, Dall-E, and ChatGPT, are now selling commercial subscriptions that provide the only way to access the full functionality of the products. OpenAI, for example, began as a non-profit research organization offering

---

[115] Comments of Copyright Clearance Center, Inc., Intellectual Property Protection for Artificial Intelligence Innovation, 84 Fed. Reg. 58141, Before USPTO, at 2 (Jan. 10, 2020), https://www.uspto.gov/sites/default/files/documents/Copyright-Clearance-Center_RFC-84-FR-58141.pdf.

[116] Cristina Criddle et al., *AI and Media Companies Negotiate Landmark Deals Over News Content*, Financial Times (June 17, 2023), https://www.ft.com/content/79eb89ce-cea2-4f27-9d87-e8e312c8601d; Helen Coster & Zaheer Kachwala, *News Corp in Negotiations with AI Companies over Content Usage, CEO Says,* Reuters (Sept. 7, 2023), https://www.reuters.com/business/media-telecom/news-corp-negotiations-with-ai-companies-over-content-usage-ceo-2023-09-07/.

[117] *Copyright Resources*, Cison, https://www.cision.com/legal/copyright-resources/.

[118] *LexisNexis Extends Multi-Year Content Agreement with The New York Times*, LexisNexis Press Room (Sept. 20, 2021), https://www.lexisnexis.com/community/pressroom/b/news/posts/lexisnexis-extends-multi-year-content-agreement-with-the-new-york-times.

[119] *Annual Copyright License*, Copyright Clearance Center, https://www.copyright.com/wp-content/uploads/2021/01/Product-Sheet-Annual-Copyright-License-8-2020.pdf; *Copyright Clearance Center Integrates Rights Delivery Platform on Copyright.com*, Library Technology Guides (Mar. 1, 2011), https://librarytechnology.org/pr/15507/copyright-clearance-center-integrates-rights-delivery-platform-on-copyright-com.

[120] *What We Do*, N.Y. Times, https://nytlicensing.com/what-we-do/; *Products*, Wash. Post, https://www.washingtonpost.com/licensing-syndication/products.

ChatGPT for free, but pivoted to a for-profit model that now requires a paid subscription to access all its features.[121]

4. There is no satisfactory independent justification for the copying.

There is no independent reason why GAI models must ingest valuable copyright-protected expressive works apart from the desire to incorporate that very expression. While GAI developers may prefer to copy such high-quality media unburdened from any licensing obligations, some of the very companies that have infringed the copyrighted content of N/MA members have licensed content from others for similar purposes. For example, Stability AI and Meta have launched text-to-music generators trained solely on licensed musical works and sound recordings,[122] and Google is in discussions to develop a similar tool using music licensed from Universal Music Group.[123] OpenAI has licensed imagery from Shutterstock since 2021, providing access that its CEO Sam Altman said was "critical" to the training of its DALL-E engine, and it recently announced an expanded licensing deal covering the licensing of Shutterstock's music catalogue as well.[124] Others seem to be trying to get this right from the start. Adobe Firefly is a text-to-image generator trained solely on Adobe Stock images, openly licensed content, and public domain content.[125] Getty has developed a text-to-image generator trained solely on licensed images.[126]

In an implicit acknowledgment that GAI training can continue and flourish without training LLMs on unauthorized copies, Google recently announced a new mechanism, Google Extended, which

---

[121] Alex Konrad, *OpenAI Releases First $20 Subscription Version of ChatGPT AI Tool*, Forbes (Feb. 1, 2023), https://www.forbes.com/sites/alexkonrad/2023/02/01/openai-releases-first-subscription-chatgpt/?sh=b4debac7f5f1; *see also* Lemley & Casey, *supra* note 2, at 746 ("[ML] systems … rarely transform the databases they train on; they are using the entire database, and for a commercial purpose at that.").

[122] Daniel Tencer, *Stability AI Launches Text-to-Music Generator Trained on Licensed Content Via a Partnership with Music Library AudioSparx*, Music Business Worldwide (Sept. 14, 2023), https://www.musicbusinessworldwide.com/stability-ai-launches-text-to-music-generator-trained-on-licensed-content-via-a-partnership-with-music-library-audiosparx/; Justinas Vainilavicius, *Meta Releases Music Generator Called MusicGen*, Cybernews (Aug. 3, 2023), https://cybernews.com/tech/meta-releases-music-generator-musicgen/.

[123] Hibaq Farah, *Google and Universal Music Working on Licensing Voices for AI-Generated Songs*, The Guardian (Aug. 9, 2023), https://www.theguardian.com/technology/2023/aug/09/google-and-universal-music-working-on-licensing-voices-for-ai-generated-songs.

[124] Daniel Tencer, *OpenAI Secures License to Access Training Data from Shutterstock . . . Including Its Music Libraries*, Music Business Worldwide (July 12, 2023), https://www.musicbusinessworldwide.com/openai-secures-license-to-access-training-data-from-shutterstock-including-its-music-libraries/.

[125] *Firefly FAQ for Adobe Stock Contributors*, Adobe (Updated Oct. 4, 2023), https://helpx.adobe.com/stock/contributor/help/firefly-faq-for-adobe-stock-contributors.html.

[126] Emilia David, *Getty Made an AI Generator that Only Trained on its Licensed Images*, The Verge (Sept. 25, 2023), https://www.theverge.com/2023/9/25/23884679/getty-ai-generative-image-platform-launch.

will allow website publishers to opt out of having their content used to improve the company's AI models in the future while maintaining access to such content through Google Search.[127] OpenAI has similarly announced that internet sites can now block OpenAI's GPTBot and keep their sites out of ChatGPT.[128] This "opt-out" approach is, of course, antithetical to U.S. copyright law (and does not allow for opt-out of the content already scraped). There is also a wealth of material in the public domain or available under open licenses available for the LLMs to use to build their models.

Notably, N/MA members stand ready to come to the table and discuss reasonable licensing solutions to facilitate reliable, updated access to trustworthy expressive content, something that will benefit all interested parties and society at large, rather than engage in litigation to protect their rights.[129]

In this setting, the GAI developers' goal to create LLMs or to employ those models to power GAI products, however laudable, does not justify their infringement of this valuable corpus of copyrighted expression. Sam Altman, the founder of OpenAI, and Brad Smith, President of Microsoft, each acknowledged this point in their recent testimony before Congress, explaining that creators of expressive works deserve to control the rights to, and must benefit from, their creations.[130]

Indeed, courts have long recognized that such generalized fair use justifications should not be used to insulate widespread infringement. *American Geophysical Union v. Texaco, Inc*., for example, found that Texaco's photocopying of scientific journals for purposes of commercial R&D was not a fair use, even where the company had made the copies to enrich their researchers' knowledge, because the company was engaged in a "systematic process of encouraging employee researchers

---

[127] Emma Roth, *Google Adds a Switch for Publishers to Opt Out of Becoming AI Training Data*, The Verge (Sept. 28, 2023), https://www.theverge.com/2023/9/28/23894779/google-ai-extended-training-data-toggle-bard-vertex.

[128] Emilia David, *Now You Can Block OpenAI's Webcrawler*, The Verge (Aug. 7, 2023), https://www.theverge.com/2023/8/7/23823046/openai-data-scrape-block-ai.

[129] *See supra* notes 43-46.

[130] *Oversight of A.I.: Rules for Artificial Intelligence*, 118th Cong. (2023), https://techpolicy.press/transcript-senate-judiciary-subcommittee-hearing-on-oversight-of-ai/ (statement of Sam Altman) ("[W]e think that creators deserve control over how their creations are used and what happens sort of beyond the point of, of them releasing it into the world … we think that content creators, content owners, need to benefit from this technology … We're still talking to artists and content owners about what they want. I think there's a lot of ways this can happen, but very clearly, no matter what the law is, the right thing to do is to make sure people get significant upside benefit from this new technology. And we believe that it's really going to deliver that. But that content owners likenesses people totally deserve control over how that's used and to benefit from it."); *id.* (statement of Brad Smith) ("[G]enerally I think we should let local journalists and publications make decisions about whether they want their content to be available for training or grounding and the like. And that's a big topic and it's worthy of more discussion. And we should certainly let them, in my view, negotiate collectively because that's the only way local journalism is really going to negotiate effectively.").

to copy articles so as to multiply available copies while avoiding payment."[131] As the court explained:

> The purposes illustrated by the categories listed in section 107 refer primarily to the work of authorship alleged to be a fair use, not to the activity in which the alleged infringer is engaged. Texaco cannot gain fair use insulation for [its employee]'s archival photocopying of articles (or books) simply because such copying is done by a company doing research. It would be equally extravagant for a newspaper to contend that because its business is "news reporting" it may line the shelves of its reporters with photocopies of books on journalism or that schools engaged in "teaching" may supply its faculty members with personal photocopies of books on educational techniques or substantive fields. Whatever benefit copying and reading such books might contribute to the process of "teaching" would not for that reason satisfy the test of a "teaching" purpose.[132]

This principle applies in full force to GAI development. While developers have contended that their unlicensed use of material for LLM training and GAI development purposes is justifiable because the LLMs ingest the copyrighted content to "learn" from the content, just like a human being, no one is allowed to copy an underlying work just because they have an alleged good reason to read the underlying document but don't want to buy (or otherwise lawfully access) a copy. As one scholar explains:

> Making gigabytes upon gigabytes of copies of copyrighted art, in order to teach a machine to mimic that art, is indeed a remarkable technological achievement. An artificially intelligent painter or writer may yield social benefits and enrich the lives of many beholders and users. However, this view of productivity is overbroad. No human can rebut an infringement claim merely by showing that he has learned by consuming the works he copied, even if he puts this new knowledge to productive use later on … A teacher who copies to broaden his personal understanding is a productive consumer, but he nonetheless must pay for the works he consumes. If the teacher's consumption of copyrighted works inspires him to create new

---

[131] 60 F.3d 913, 920 (2d Cir. 1994).

[132] *Id*. at 924; *see also Cambridge Univ. Press v. Patton*, 769 F.3d 1232, 1263-64 (11th Cir. 2014) ("[A]llowing some leeway for educational fair use furthers the purpose of copyright by providing students and teachers with a means to lawfully access works … But, as always, care must be taken not to allow too much educational use, lest [the court] undermine the goals of copyright by enervating the incentive for authors to create the works upon which students and teachers depend."); *Princeton Univ. Press v. Mich. Document Servs., Inc*., 99 F.3d 1381 (6th Cir. 1996) (reproduction of significant portions of copyrighted works for use in course packets is not fair use); *Marcus v. Rowley*, 695 F.2d 1171 (9th Cir. 1983) (same for teacher's educational booklet); H.R. Rep. No. 94-1476, at 66-67 (1976), https://www.copyright.gov/history/law/clrev_94-1476.pdf ("[A] specific exemption freeing certain reproductions of copyrighted works for educational and scholarly purposes from copyright control is not justified."); Linda Starr, *Is Fair Use a License to Steal?*, Education World (May 25, 2010), https://www.educationworld.com/a_curr/curr280b.shtml#:~:text=The%20fair%20use%20doctrine%20is,and%20scholarship%2C%20and%20classroom%20instruction.

> scholarship, so much the better, but his subsequent productivity does not entitle him to a refund for the works that influenced him. In much the same way, machine learning makes consumptive use of copyrighted materials in order to facilitate future productivity. If future productivity is no defense for unauthorized human consumption, it should not excuse robotic consumption, either.[133]

Of course, LLM machines are not humans. As set forth above, they do not "learn"—they copy, and they do so on a massive scale that no human could replicate. Because a market exists to provide high quality publisher content for purposes such as AI training, the goal of building LLMs does not justify the unlicensed copying of N/MA members' expressive works.

5. The unlicensed use of training materials serves a system designed to produce substitutional outputs.

LLMs are designed to produce outputs that can substantially copy from, compete with, and substitute for original text content. Even in the furtherance of new technological development, no court has held fair the copying of content to develop a system whose purpose is to substitute for the original works. Rather, cases holding "fair" the use of copyrighted materials to develop a new technology or further a technological purpose are grounded on findings that the ultimate use *did not* compete with the copyrighted works. The first fair use factor does not require news and media publications to be mined to fuel their replacements.

In *Authors Guild*, for example, the court found that neither of the challenged uses (for "search" and "snippets") could provide a meaningful substitute for the copied books and instead were likely to help users identify books of interest.[134] It concluded that if the snippets were arranged into a coherent aggregate "manner and order" (which the challenged system disallowed) "that would raise a very different question beyond the scope of our inquiry."[135] Similarly, in *Kelly v. Arriba Soft Corp.*, the court found that the search engine "Arriba's use of Kelly's images in its thumbnails does not harm the market for Kelly's images or the value of his images."[136]

---

[133] Benjamin L.W. Sobel, *Artificial Intelligence's Fair Use Crisis*, 41 Colum. J. L. & Arts 45, 73-74 (2017); *id*. at 74 (suggesting "a constituent who copies a news program to help make a decision on how to vote" would not be protected by the fair use doctrine despite the salutary purpose (quoting *Sony Corp. of Am. v. Universal City Studios, Inc.*, 464 U.S. 417, 455 n.40 (1984))).

[134] 804 F.3d at 218.

[135] *Id.* at 223.

[136] 336 F.3d 811, 821 (9th Cir. 2003); *see also Google LLC v. Oracle Am., Inc.*, 141 S. Ct. 1183, 1206-07 (2021) ("*Oracle*") (jury's fair use determination barred Oracle from "overcom[ing] evidence that, at a minimum, it would have been difficult for Sun [Oracle's predecessor] to enter the smartphone market" even without Google's alleged infringement, including Sun's former CEO's testimony that Sun's failure to build a smartphone was not attributable to Google's alleged infringement); *cf. Sony Corp. of Am. v. Universal City Studios*, 464 U.S. 417, 456 (1984) (noting that plaintiffs "failed to demonstrate that time-shifting would cause any likelihood of nonminimal harm to the potential market for, or the value of, their copyrighted works.").

In contrast, as shown above in Section IV.D, the LLMs can and do generate outputs that replicate or closely paraphrase the original expressive works. Consumer-facing chatbot services built around these models, including those integrated into search engines like Bing or Google, are well poised to directly substitute for publishers and to usurp their valuable relationships with readers of news, magazine, and web content. Marketing for these new features makes clear that they are intended to create substitutional narratives deployed by the GAI apps, that can substantially copy from, compete with, and substitute for the primary expressive material. Unchained from constraints to serve as no more than an electronic reference or bridge to a primary source, narrative search results can provide users with sufficient content (full key portions and highlights of expressive content), that substitutes for any need to read the original. As a recent New Yorker article explains, the "goal" of "large language models, like OpenAI's ChatGPT and Google's Bard" "is to ingest the Web so comprehensively that it might as well not exist."[137]

These chatbot search uses thus go well beyond the nuanced reasoning and careful guardrails established by cases like *Authors Guild* and *Kelly* and into competitive, consumptive uses that are distinctly unfair to content owners. Indeed, courts routinely dismiss fair use arguments for new digital products that have a similar purpose to, and could supplant, the original work.[138] That reasoning applies here.

* * *

For all these reasons, the first factor favors a finding of infringement and not fair use.

B. The effect of GAI copying on the market for publisher content is predictable and real (fourth factor).

The fourth fair use factor directs courts to consider "the effect of the use upon the potential market for or value of the copyrighted work."[139] The focus is on whether widespread conduct like the conduct of the alleged infringer "would adversely affect the potential market for the copyrighted work," including market harm to the original work and to derivative works.[140] While the examination of potential markets is not without limit, "traditional, reasonable, or likely to be

---

[137] James Somers, *How Will A.I. Learn Next?*, The New Yorker (Oct. 5, 2023) (reporting that the number of new posts the website Stack Overflow, where computer programmers went to ask and answer programming questions, has decreased by 16% since the debut of ChatGPT).

[138] *See, e.g.*, *Fox News Network, LLC v. TV Eyes, Inc*., 883 F.3d 169, 177, 181 (2d Cir. 2018) (media monitoring service, while transformative, was not fair, because it usurped plaintiff's market); *Hachette Book Grp., Inc. v. Internet Archive*, No. 20-CV-4160 (JGK), 2023 WL 2623787, *18-25 (S.D.N.Y. Mar. 24, 2023) (Internet Archive's electronic copying and unauthorized lending of 3.6 million books protected by valid copyrights is not a fair use because it competed with plaintiff's licensing market); *Meltwater*, 931 F. Supp. 2d at 561 (crawling of various websites for Associated Press's stories and scraping "snippets" of those stories for use in notifying and informing Meltwater's own customers of certain stories directly competed with the Associated Press such that Meltwater's copying would deprive the Associated Press of a stream of income to which it was entitled).

[139] 17 U.S.C. § 107(4).

[140] *Harper & Row*, 471 U.S. at 566, 568 (emphasis omitted).

developed markets" are considered.[141] As the *Texaco* court recognized, "[i]t is indisputable that, as a general matter, a copyright holder is entitled to demand a royalty for licensing others to use its copyrighted work, and that the impact on potential licensing revenues is a proper subject for consideration in assessing the fourth factor."[142]

GAI's unauthorized use of copyrighted material harms the market in two ways.

First, with respect to inputs, GAI developers' unauthorized use of publisher content to build their LLMs deprives publishers of an available licensing market, such that the fourth factor also should favor a finding of infringement when publisher content is used without authorization for training purposes.[143]

While developers complain that it is unworkable to license content for their ingestion needs,[144] there is a long history of publishers licensing their content for a variety of uses and licensing deals, and negotiations are occurring in the open market specifically for GAI uses, as documented above at Section V.A.2.

As explained above, there is also a long history of media organizations and associations licensing their content for a variety of uses, including to media monitoring entities, to LEXIS, and through the Copyright Clearance Center.[145]

Examples also abound, both here and abroad, of collective licensing of copyrighted content, and these models demonstrate the paths that exist for efficient licensing frameworks to meet GAI needs. The Copyright Clearance Center, for example, was formed by authors, publishers, and users to facilitate "centralized licensing of text-based copyrighted materials," and it has grown to represent copyright holders from nearly every country, with access to millions of sources.[146]

---

[141] *Texaco*, 60 F.3d at 929-30.

[142] *Id*. at 929 (citation omitted).

[143] *Texaco*, 60 F.3d at 930 (finding fourth factor favored a finding of infringement where the challenged photocopying harmed an existing "workable market for institutional users to obtain licenses for the right to produce their own copies of individual articles via photocopying"); *see also TV Eyes*, 883 F.3d at 180 (by using content without payment, Fox was deprived of "licensing revenues from TVEyes"); *Davis v. Gap, Inc*, 246 F.3d 152, 175-76 (2d Cir. 2001) (freely taking a copyrighted work allowed defendant to avoid "paying the customary price," that plaintiff "was entitled to charge" for use of work, and that, as a result, plaintiff "suffered market harm through his loss of the royalty revenue to which he was reasonably entitled in the circumstances, as well as through the diminution of his opportunity to license to others").

[144] OpenAI, LP, *supra* note 58, at 11.

[145] *See supra* notes 117-20.

[146] Comments of Copyright Clearance Center, Inc. 79 Fed. Reg. 2696 (Mar. 3, 2024), https://www.copyright.gov/docs/recordation/comments/79fr2696/CCC.pdf; *Annual Copyright License*, Copyright Clearance Center, *supra* note 119; *Licensing Services Overview*, Copyright Clearance Center, https://www.copyright.com/wp-content/uploads/2016/01/LicensingSrvcsOverview-7.19.16.pdf.

Outside the United States, collective management organizations broadly manage news and media licensing, such as NLA Media Access in the U.K.[147]

Second, it is indisputable that GAI output is intended to, and does, substitute for human-generated content, including publisher content.[148] As explained above, already less than 65% of searches result in clicking through to the underlying source.[149] That percentage is only going to worsen with narrative search results. Indeed, marketing experts expect click-through rates for generative search responses to be even lower than already declining rates for organic results.[150] "Particularly for informational searches, Google will aggregate (or flat-out plagiarize) from the search results and give users much of what they're looking for."[151] "Users may find all the information they

---

[147] Tarja Koskinen-Olsson, *Collective Management of Text and Image-Based Works*, WIPO (Updated 2023) https://www.wipo.int/edocs/pubdocs/en/wipo-pub-924-2023-en-collective-management-of-text-and-image-based-works.pdf; *A Guide to Media Monitoring and Corporate Licensing*, Press Database and Licensing Network, at 14 (Oct. 2017), https://static1.squarespace.com/static/-5eca9a7fe349354c54ae6cab/t/5ef2b3025a06263ec1a24a14/1592963847770/pdln_guide+to+corporate+and+mmo+licensing.pdf; *What Is a Performing Rights Organization (PRO)?*, SESAC (May 5, 2022), https://www.sesac.com/what-is-a-performing-rights-organization-pro/.

Collective licensing has also flourished in the music industry, further demonstrating the potential to develop efficient, large-scale licensing models for GAI needs. The performing rights organizations (PROs) such as ASCAP, BMI, and SESAC license the right to publicly perform musical compositions on behalf of copyright owners. PROs collectively "cover[] almost all of the millions of songs currently copyright protected," and they operate by offering "blanket authorization to use the music [each organization] represents in exchange for license fees," which are then distributed "as royalties to its affiliated songwriters, composers, and music publishers." *What Is a Performing Rights Organization (PRO)?*, SESAC *supra* note 145.

[148] *See also*, *e.g.*, Comment of OpenAI, LP Regarding Request for Comments on Intellectual Property Protection for Artificial Intelligence Innovation, Before the USPTO, at 11, https://www.uspto.gov/sites/default/files/documents/OpenAI_RFC-84-FR-58141.pdf ("Writers who were employed to perform formulaic composition might be able to devote their energies to more creative forms of self-expression *once machines supplant them*." (quoting Sobel, *supra* note 131, at 80); Lemley & Casey, *supra* note 2, 767 (Machine learning "empowers [] companies to extract value from authors' protected expression without authorization" or compensation "and to use that value for commercial purposes that may someday jeopardize the livelihoods of human creators." (quoting Sobel, *supra* note 131, at 97); *id.* at 777 (AI systems trained "to generate their own expressive works … pose a threat of significant substitutive competition to the work originally copied." (internal quotation marks omitted)).

[149] See *supra* note 54.

[150] *See*, *e.g.*, Rebecca Krause, *Google's Search Generative Experience (SGE): A Marketer's Guide*, Seer Interactive (August 10, 2023), https://www.seerinteractive.com/insights/googles-search-generative-experience ("As SGE rolls out to more users, the click-through-rate of the ten organic links (even position 1) may lower.")

[151] Dave Shapiro, *Generative AI in Search*, Neil Patel, https://neilpatel.com/blog/generative-ai-in-search/ ("people will find enough of what they need in the SGE and not click on organic results.").

need directly on the search page, so there's no need to click on the source website."[152] As set forth above, no court has deemed fair the copying of expressive works, even at the development stage, for the purposes of eventually competing with and substituting for the original work. The substitutional use of the GAI outputs is a further reason why the fourth factor favors a finding of infringement with respect to the unauthorized use of publisher content at the training stage.

The effect of GAI copying at the output stage is self-evident. Where the outputs replicate or closely paraphrase the original expressive works and thus infringe upon and substitute for them, such that users no longer need to connect with or obtain the original works from their original sources, such uses harm the market for the publishers' works.

C. GAI copying takes substantial portions of expressive works in their entirety (second and third factors).

Under the second factor, courts consider whether a work is creative or functional, "recogn[izing] that some works are closer to the core of intended copyright protection than others."[153] The second factor is typically less important than the first and fourth factors.[154]

Although news, magazine, and digital media content includes underlying facts, reporting seeks to determine which facts are significant and to recount them in an interesting manner, and is thus creative in nature.[155] Such content also extends well beyond traditional news reporting and includes pieces devoted to opinion and analysis. Here, where developers copy publisher content so that LLMs can best mimic human speech,[156] the copying is necessarily exploiting the content for its expressive qualities and the second factor favors a finding of infringement for both inputs and outputs.

The third factor evaluates both the quantity and quality of the copying, and "examine[s] the amount and substantiality of the portion used in relation to the copyrighted work as a whole," including whether the "heart" of the work is copied.[157] "[T]he fact that a substantial portion of the infringing work was copied verbatim is evidence of the qualitative value of the copied

---

[152] Sam Stemler, *9 Things You Need to Know about Google Search Generative Experience (SGE)*, Web Ascender (August 29, 2023), https://www.webascender.com/blog/9-things-you-need-to-know-about-google-search-generative-experience-sge/.

[153] *Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569, 586 (1994); *Google LLC v. Oracle Am., Inc.*, 141 S. Ct. 1183, 1202 (2021).

[154] *Authors Guild v. Google, Inc.*, 804 F.3d 202, 213 (2d Cir. 2015).

[155] *See Harper & Row Publishers, Inc. v. Nation Enters.*, 471 U.S. 539, 547 (1985) ("Creation of a nonfiction work, even a compilation of pure fact, entails originality."); *see also Authors Guild*, 804 F.3d at 220 ("Those who report the news undoubtedly create factual works. It cannot seriously be argued that, for that reason, others may freely copy and re-disseminate news reports."); *Fox News Network, LLC v. TV Eyes, Inc.*, 883 F.3d 169, 177, 178 (2d Cir. 2018) (rejecting argument that, since facts are not copyrightable, the factual nature of a creative compilation favors a finding of fair use).

[156] *See supra* Section IV.A.

[157] *Harper & Row*, 471 U.S. at 564-65.

material, both to the originator and to the plagiarist who seeks to profit from marketing someone else's copyrighted expression."[158] The massive scale of copying also favors a finding of infringement.[159]

Here, for inputs, developers copy substantially all of the expression in publisher content during the course of LLM training and development of GAI tools, and it is reasonable to conclude that the "heart" of the work is copied. Moreover, the GAI developers' copying can be viewed as excessive given the degree to which the copies usurp the available licensing market.[160]

Application of the third factor at the output stage must be evaluated on a case-by-case basis, depending on the portions of the works which the outputs copy. Suffice to say, the third factor will favor a finding of infringement at the output stage whenever the outputs copy sufficient portions or the heart of the copied works.

VI. Recommendations

The News/Media Alliance makes the following recommendations.

- **GAI systems should be transparent to publishers.** Publishers have a right to know who copied their content and what they are using it for. The Alliance calls for strong regulations and policies imposing transparency requirements to the extent necessary for publishers to enforce their rights. Publishers have a legitimate interest in determining what content of theirs has been and is used in GAI systems. Using datasets or applications developed by non-profit, research, or educational third parties to power commercial GAI systems must be clearly disclosed and not used to evade transparency obligations or copyright liability.

- **GAI use of publisher content, without authorization, must be recognized as infringing.** Policy makers and industry participants must recognize that the unauthorized use of publisher content to (1) train large language models for the purpose of generating text outputs; and/or (2) generate outputs that replicate or are substantially similar to publishers' original expressive works, violates publishers' exclusive rights to their protected works and unfairly competes with and usurps their markets. This recognition is critical to foster meaningful negotiations between GAI developers and deployers, on the one hand, and publishers, on the other hand.

- **Licensing for GAI Uses Should Be Encouraged and Facilitated.** Congress and the Copyright Office should explore ways to facilitate or encourage the licensing of publisher content for GAI purposes. Efficient and widespread licensing of publisher content for GAI purposes will help ensure a steady supply of high-quality and human-created content that

---

[158] *Id*. at 565.

[159] *See, e.g*., *Hachette Book Grp., Inc. v. Internet Archive*, No. 20-CV-4160 (JGK), 2023 WL 2623787, *8 (S.D.N.Y. Mar. 24, 2023) ("Unlike Sony, which only sold the machines, IA scans a massive number of copies of books and makes them available to patrons …").

[160] *Campbell v. Acuff-Rose Music, Inc*., 510 U.S. 569, 587-88 (1994); *see also supra* Section V.A.2.

can aid in the development of high-quality, accurate, and trustworthy GAI products and outputs.

- **Market Power Imbalances Should be Corrected So Publishers Can Engage in Fair Negotiations to License Their Content for GAI Development.** Relatedly, the Alliance advocates the passage of legislation it has proposed allowing news publishers to bargain collectively with certain dominant technology providers. The bipartisan legislation, the Journalism Competition and Preservation Act, was introduced as H.R. 2054, with an identical Senate version (S.1700) to address this extreme market and legal failure. Copyright laws alone will not work if dominant online players who are actively engaged in GAI development and deployment can use their market power to extract exploitative and anticompetitive terms from publishers, or condition licensing for GAI development on publisher concessions around other business lines. An appropriately tailored safe harbor—like the Journalism Competition and Preservation Act—will help begin to restore some semblance of a balance of power by giving publishers the ability to begin offsetting the market dominance of the large online platforms. These platforms also should not be allowed to abuse their market power in traditional search functions to force publishers to allow their content to be crawled for GAI uses. Publishers must be allowed to consent to the crawling of their sites for traditional search functionality while declining or negotiating different terms for the crawling of their sites for GAI.

Technical Appendix

Technical appendix[1]

Summary Abstract:

In this report, we investigate the extent to which publisher content, including news, magazine, and digital media content, is used as part of training for large language models (LLMs), as well as the extent to which these models can reproduce some of this content. Our results provide both statistical and anecdotal evidence for the hypotheses that news publisher content has been used in the training of LLMs and that in some cases, LLMs are able to reproduce it nearly verbatim. We divide our analysis into three subsections. In subsection 1, we assess the extent of copyrighted news publisher content that is included in public datasets that have reportedly been used to train LLMs. In subsection 2, we performed boilerplate analyses on two LLMs used in popular chatbots (GPT-4 used in OpenAI's ChatGPT and PaLM-2 used in Google's Bard) to identify the extent that publisher content is used in LLMs. We also ran a cloze test analysis on OpenAI's GPT-4. In subsection 3, we show that the output of GPT-4, as used in OpenAI's ChatGPT, is in some cases quantitatively similar to the original publisher's content. All testing included in this paper occurred in August, September, and October 2023.

In subsection 1, we assessed a small sample of publisher content using 16 publication domains that were volunteered by News/Media Alliance members. We examined the presence of content from these domains in the open-source dataset Common Crawl, as well as in three other datasets reported as being developed specifically for LLM training- C4, OpenWebText, and OpenWebText2. As measured by the presence of unique URLs, together these 16 publication domains comprised 0.02% of the Common Crawl dataset and between 0.15% and 1.97% of the three datasets developed for LLM training. Our assessment demonstrates that datasets specifically developed for LLM training, such as C4 and OpenWebText, skew towards content from the 16 publication domains. When comparing these datasets to Common Crawl, publisher representation increases by a factor of 5 for C4 to approximately 100 for OpenWebText2. This assessment does not capture the full volume of publisher content in the open-source datasets, but it is useful for understanding the treatment of all publisher content.

In subsection 2, we provide examples where both GPT-4 and PaLM-2 are able to directly reproduce boilerplate language used in multiple articles, demonstrating that the LLMs are able to retain content from training. We also provide the results of cloze-testing, which assesses a model's ability to fill in the missing proper noun in a sentence from a previously published article provided as a prompt. Cloze-testing is a technique for membership inference used to determine if a corpus of data was used to train a machine learning model. GPT-4 was better at filling in the missing name in a sentence when the prompt identified the original publication, as compared to when the prompt provided no information about the publication, by approximately 45%.

[1] This research was conducted by Prof. Vince Conitzer, PhD, Aakar Gupta, PhD, Eric Sodomka, PhD, and their team. The authors are grateful for support from the News/Media Alliance.

In subsection 3, we show examples of GPT-4 responding with a 231-word string directly out of a publisher's article and generating responses very similar to original publisher content.

In the final subsection, we discuss limitations of the membership inference analyses such as challenges with using the training cutoff date to create a control.

1. Publisher Content in Public Datasets

This subsection aims to answer the following question: *"To what extent does copyrighted publisher content appear in public datasets, especially those datasets that LLM engineers have been reported to use for LLM training?"*

Our analysis found that, for the sample of publications we analyzed, the proportion of content included in C4, OpenWebText, and OpenWebText2 (0.15% to 1.97% of unique URLs) was far greater than in the snapshot of Common Crawl (0.02% of unique URLs). The interpretation is that datasets curated for LLM training skew towards publisher content, as compared to Common Crawl which may represent a slice of the internet.

1.1 Methods

1.1.1 Public Datasets Considered

We consider four public datasets: *Common Crawl*, *C4*, *OpenWebText*, and *OpenWebText2*. An overview of these datasets and their versions is provided below in Tables 1 and 2:

Table 1: Description of Common Crawl, C4, and WebText

	Common Crawl[2]	**C4**	**WebText**
Created By	Common Crawl (non-profit)	Google	OpenAI
Dataset Description and Source	Millions of domains from the open web	Curated subset of April 2019 Common Crawl's web corpus	Contains text from URLs scraped from Reddit posts up to 2017 with >3 karma

[2] We examine the Common Crawl crawl archive generated in July/August 2021, and not the entire Common Crawl database.

Dataset Purpose	To provide free web crawl data to anyone	Used to train T5 text-to-text transformers[3]	Used to train GPT-2[4]
Dataset Size	A month's crawl can include upwards of 300 TiB of data; ~90 monthly crawls in total	English cleaned version contains 305 GB of data	40GB of text from 8M documents
Data Included	Text and metadata like URL, crawl/extraction date, etc.	Site text and URL	Dataset has not been released
Index of URLs Present	Yes, with index table for each month's crawl containing up to 300GB	Not directly, but can be extracted from dataset	Dataset has not been released
When was it introduced?	Covers 2008-present	Introduced in Google's T5 paper (July 2020)	Produced in 2019
Where is it located?	Instructions for getting access can be found at commoncrawl.org/get-started	AllenAI version: huggingface.co/datasets/allenai/c4	N/A

Table 2: Description of WebText extensions and replications

	WebText2	**OpenWebText**	**OpenWebText2**
Created by	OpenAI (for internal use)	J. Peterson, S. Meylan, & D. Bourgin	Non-profit EleutherAI
Dataset Description and Source	An extended version of WebText, based on the outbound Reddit links from 2005 to 2020	Contains URLs scraped via Pushshift.io from Reddit posts up to 2017 with > 3 karma. Google's code for building C4 was used to construct OWT	Replication of WT2 and an extended version of OWT: covers 2005 – April 2020; multilingual webpages; includes metadata

[3] "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer", 2020, https://arxiv.org/pdf/1910.10683v3.pdf

[4] "Language Models are Unsupervised Multitask Learners", 2019, https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Dataset Purpose	Created for the training of GPT-3[5]	An open-source replication of WT	Developed to be used as a part of The Pile, an open-source high-quality dataset for LLM training[6]
Dataset Size	19 billion tokens[7]	Around 23 million URLs, 2GB in .zip format	17 million scraped webpages, 28GB in json.zst.tar format
Data Included	Dataset has not been released	URLs only	URLs and text
When was it introduced	Mentioned in GPT-3 paper (July 2020)	Developed in 2019-2020	Released in 2020 to expand the coverage of OWT for The Pile
Where is it located?	N/A	GitHub page: https://github.com/jcpeterson/openwebtext; URL data: https://mega.nz/folder/EZZD0YwJ#9_PlEQzdMVLaNdKv_ICNVQ/folder/cc4RgQQZ	OWT2 website: https://openwebtext2.readthedocs.io/en/latest/ Previously included in The Pile dataset at the-eye.eu/public/AI/pile/.[8] Circa Sept 2023, OWT2 and The Pile are no longer available for download/access.

1.1.2 Sample of publication domains

We focus on 16 publication domains volunteered by News/Media Alliance members. Included in these domains are news, magazines, and other digital media.

[5] "Language Models are Few-Shot Learners", 2020, https://arxiv.org/pdf/2005.14165.pdf
[6] "The Pile: An 800GB Dataset of Diverse Text for Language Modeling", 2020, https://arxiv.org/pdf/2101.00027.pdf
[7] Text can be broken down into units such as words or sequences of characters. In NLP, these units are called tokens and support semantic processing tasks.
[8] The Eye webpage with OWT2 data and other components of the Pile can be viewed using the Wayback Machine: see https://web.archive.org/web/20230710081156/https://the-eye.eu/public/AI/pile_preliminary_components/, https://web.archive.org/web/20230316084127/https://the-eye.eu/public/AI/pile/.

1.1.3 Metrics of Interest

For each publisher content source S and dataset D, we focus on the number of unique URLs for S indexed in D.
Unique URLs are defined as unique URL strings that do not repeat within the corresponding dataset. To note, the method for evaluating the number of unique URLs in the data has some limitations, and there could be instances of links that point to the same page even though their URL strings are different; for example, google.com/search and google.com/webhp point to the same page but would be considered unique.

1.1.4 Identifying Copyrighted Publisher Content in Public Datasets

For each public dataset, each document in the dataset corresponds to a single URL from which that data was scraped. We identify whether the URLs in this index belong to one of the sample publication domains.[9] We also remove duplicates, if any.

An example of the Python code evaluating if URL *"url"* belongs to target domain *"member_domain"* is as follows:

```
from urllib.parse import urlparse
def contains_strictest_member_url(member_domain, url):
    domain_c4 = urlparse(url).netloc
    if domain_c4.endswith(member_domain):
        return True
    return False
```

1.2 Results

The statistics in Table 3 below are consistent with existing findings on the composition of LLM training sets. For example, Washington Post reporters[10] analyzed the composition of C4 data in terms of tokens and found that publishers of news, magazine, and digital media content account for similar volumes of the C4 corpus.

Table 3. Unique URL counts from public datasets belonging to publisher content sources. (Percentages of unique URL counts for that dataset are shown in parentheses.)

Source	Common Crawl (July/Aug. 2021)	C4	OpenWebText	OpenWebText2

[9] One publisher used two domains for the same brand. We joined the data only for that publication in order to properly compare differences between Common Crawl and C4, OpenWebText, and OpenWebText 2.
[10] https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/

Publication 1	41,729 (0.0013%)	35,558 (0.0097%)	13,853 (0.060%)	12,356 (0.072%)
Publication 2	19,660 (0.0006%)	17,422 (0.0048%)	12,003 (0.052%)	11,447 (0.067%)
Publication 3	32,791 (0.0010%)	12,664 (0.0035%)	16,479 (0.072%)	15,247 (0.089%)
Publication 4	42,141 (0.0013%)	169,965 (0.047%)	278,161 (1.21%)	209,707 (1.23%)
Publication 5	46,898 (0.0015%)	69,052 (0.019%)	38,519 (0.17%)	35,209 (0.21%)
Publication 6	33,975 (0.0011%)	2,144 (0.00059%)	98 (0.00043%)	117 (0.00068%)
Publication 7	37,940 (0.0012%)	22,454 (0.0062%)	754 (0.0033%)	535 (0.0031%)
Publication 8	13,210 (0.00042%)	7,591 (0.0021%)	314 (0.0014%)	254 (0.0015%)
Publication 9	16,756 (0.00053%)	13,132 (0.0036%)	1,046 (0.0045%)	988 (0.0058%)
Publication 10	11,142 (0.00035%)	9,496 (0.0026%)	94 (0.00041%)	113 (0.00066%)
Publication 11	10,664 (0.00034%)	6,771 (0.0019%)	8 (0.000035%)	52 (0.00030%)
Publication 12	14,152 (0.00045%)	6,107 (0.0017%)	173 (0.00075%)	198 (0.0012%)
Publication 13	89,011 (0.0028%)	41,286 (0.011%)	25,548 (0.11%)	21,332 (0.12%)

Publication 14	30,268 (0.00096%)	33,020 (0.0090%)	5,606 (0.024%)	6,341 (0.037%)
Publication 15	61,380 (0.0019%)	53,323 (0.015%)	18,668 (0.081%)	18,543 (0.11%)
Publication 16	56,824 (0.0018%)	42,714 (0.012%)	3,749 (0.016%)	4,076 (0.024%)
Total Unique URLs from sample publications	558,541 (0.02%)	542,699 (0.15%)	415,073 (1.8%)	336,515 (1.97%)

1.3 Discussion

Table 3 demonstrates the significant skew towards publisher content in datasets curated for LLM training such as C4, OpenWebText, and OpenWebText2, as compared to datasets that serve more general purposes such as Common Crawl. For today's leading LLMs, we do not know exactly what content they were trained on, so these counts should not be construed as representative of the number of works from any given publisher that were used to train any commercial models. Instead, this analysis sheds a light on datasets that represent the community's best effort at creating similar/replicated open datasets.

2. Membership Inference: Publisher Content in Training of Commercial Large Language Models

In this subsection, we aim to answer the following question: *"To what extent is copyrighted content from news, magazines, and digital media being used to train commercial LLMs?"* The tests included in this subsection aim to assess whether the models have memorized the underlying training set directly, to the point that memorized training data can be reproduced in generated output. We find evidence that publisher content was used during model training and that the model is in fact able to reproduce some of this content.

Membership inference is a category of analysis techniques that observe the behavior of a model in order to draw conclusions about which content was included in a model's training set. We use membership inference methods to attempt this and include two approaches we took to answer this question. At a high level, we consider the following methods:

1. **Boilerplate language**: Provide the model with the start of boilerplate text used on multiple articles from a given publisher content source; and ask the model to complete it. This approach was presented by Nick Diakopoulos.[11]
2. **Cloze Testing**: Provide the model with a 25-75 word sentence that has a proper noun removed and ask the model to fill-in the missing proper noun. A similar *name-cloze test* was validated using data from books by University of California, Berkeley researchers in April 2023.[12] The paper was able to identify the top 50 copyrighted books included in GPT-4 by name-cloze accuracy.

Through a boilerplate language analysis, we found examples of GPT-4 and PaLM-2 successfully reproducing boilerplate text verbatim from the New York Times, Star Tribune, and other publishers.

GPT-4 cloze testing resulted in a 45% increase in success rate when a model was provided with the original publisher in addition to the original sentence, and a 16% increase in success rate when testing sentences published before GPT-4's proclaimed training cutoff. PaLM-2 analysis showed directionally similar but less dramatic results. In other words, giving the original source of the text as a hint improves GPT-4's ability to fill in a missing element of that text, providing evidence that the systems have memorized publishers' text.

2.1 Analysis: Boilerplate Language

Memorization of text is more likely if it appears frequently in the training of an LLM.[13] Boilerplate text refers to standardized text for a publication that appears frequently across multiple articles within a single publication (e.g., The New York Times) and is likely unique to that specific publication. Since such text is likely to frequently appear in the training set, memorization is more likely. Sentence completion, in which a model is asked to finish a sentence, can be used to effectively test for and demonstrate the memorization of boilerplate text or other types of recurring text extracts.

In these analyses, we ask GPT-4 and PaLM-2 to complete text extracts corresponding to the boilerplate language from publisher domains. The examples of boilerplate language and completions that demonstrate memorization are presented below.

2.1.1 Examples of boilerplate language completion

For The New York Times, we use the boilerplate text at the bottom of opinion pieces: *"The Times is committed to publishing a diversity of letters to the editor. We'd like to hear what you think about this or any of our articles. Here are some tips. And here's our email: letters@nytimes.com."* This draws from an experiment conducted by Generative AI in the Newsroom.[14]

[11] https://generative-ai-newsroom.com/finding-evidence-of-memorized-news-content-in-gpt-models-d11a73576d2
[12] https://arxiv.org/abs/2305.00118
[13] "Quantifying Memorization Across Neural Language Models", 2023, https://arxiv.org/pdf/2202.07646.pdf
[14] https://generative-ai-newsroom.com/finding-evidence-of-memorized-news-content-in-gpt-models-d11a73576d2

This was tested for both GPT-4 and PaLM-2, and although we did not obtain the complete extract, we were able to generate the following 16 words that came after the 5-word prompt. This provides compelling evidence for memorization, given the apparent low likelihood of both GPT-4 and PaLM-2 predicting this string of words based purely on probabilities of subsequent words (if those probabilities were based on sources without this sequence of words). An example of a successful prompt using GPT-4 can be seen in Figure 1, and another successful prompt with PaLM-2 can be found in Figure 2.

In these examples, the prompts that we used for GPT-4 and PaLM-2 are different. This difference is purposeful and necessary to achieve these results. In the course of our testing, we found that PaLM-2 responded with the original text more frequently when an example was provided, whereas GPT-4 did not require an example.

Figure 1: NYT boilerplate language completion with GPT-4, Example 1

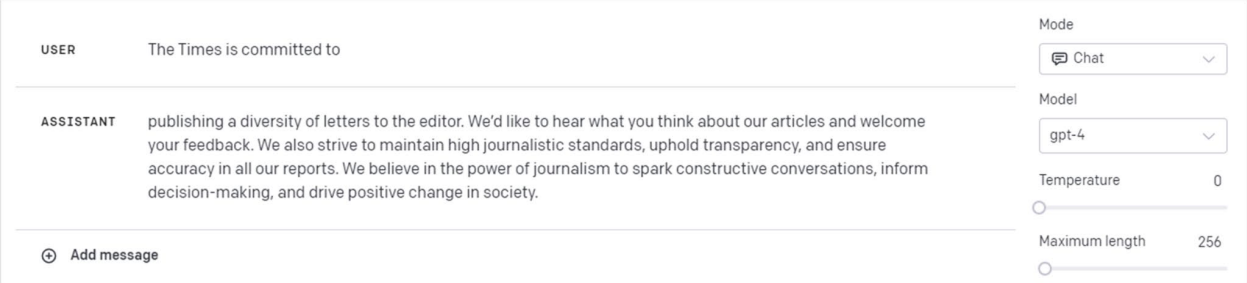


Figure 2: NYT boilerplate language completion with PaLM, Example 1[15]

```
def run_palm_completion_prompt():
    """Ideation example with a Large Language Model"""

    # Defaults
    parameters = {
        "temperature": 0,
        "max_output_tokens": 1000
    }

    model = TextGenerationModel.from_pretrained("text-bison@001")
    response = model.predict(
        prompt = """
        Write the quote
        Input: The outlook wasn't brilliant for the ...
        Output: Mudville nine that day. The score stood four to two, with but one inning more to play,
        and when Cooney died at first, and Barrows did the same, a pall-like silence fell upon the patrons of the game.

        Input: The Times is committed to ...
        Output:
        """
    )
    print(f"Response from Model: {response.text}")
```

```
run_palm_completion_prompt()
Response from Model: The Times is committed to publishing a diversity of letters to the editor. We'd like to hear from you. Please submit your letter to the editor at lett
ers@nytimes.com.
```

[15] The full text, cut off in the screenshot for Figures 2,3, and 5, is as follows: "Output: Mudville nine that day. The score stood four to two, with but one inning more to play, and when Cooney died at first, and Barrows did the same, a pall-like silence fell upon the patrons of the game." Including this example of language completion was useful for PaLM-2 to provide answers responsive to the text.

In a separate run shown in Figure 3, PaLM-2 was able to successfully reproduce the email letters@nytimes.com and a correct New York Times fax number,[16] though the number is not present in this specific NYT extract.

Figure 3: NYT boilerplate language completion with PaLM-2, Example 2

```
def run_palm_completion_prompt():
    """Ideation example with a Large Language Model"""

    # Defaults
    parameters = {
        "temperature": 1,
        "max_output_tokens": 1000,
        "top_p": 0.95,
        "top_k": 40,
    }

    model = TextGenerationModel.from_pretrained("text-bison@001")
    response = model.predict(
        prompt = """
        Begin the quote.
        Input: The outlook wasn't brilliant for the ...
        Output: Mudville nine that day. The score stood four to two, with but one inning more to play, and when Cooney died at first, and

        Input: The Times is committed to ...
        Output:
        """
    )
    print(f"Response from Model: {response.text}")
```

```
run_palm_completion_prompt()
```

```
Response from Model: The Times is committed to publishing a diversity of letters to the editor. We'd like to hear from you. Please submi
t your letter to the editor at letters@nytimes.com or fax it to 212-556-3622.
```

We also tested the ability of GPT-4 and PaLM-2 to complete the text at the bottom of articles featured in print editions. For example, we asked GPT-4 to finish part of the sentence: "A version of this article appears in print on Sept. 12, 2023, Section D, Page 3 of the New York edition with the headline: The Universe's Unpredictability."[17] The format of responses resembled that of the NYT very closely. However, the article names and dates are hallucinations. An example of such a prompt for both GPT-4 and PaLM-2 can be found in Figures 4 and 5 below.

[16] https://archive.nytimes.com/www.nytimes.com/info/help/letters.html
[17] https://www.nytimes.com/2023/09/11/science/space/astronomy-universe-simulations.html

Figure 4: NYT boilerplate language completion with GPT-4, Example 2

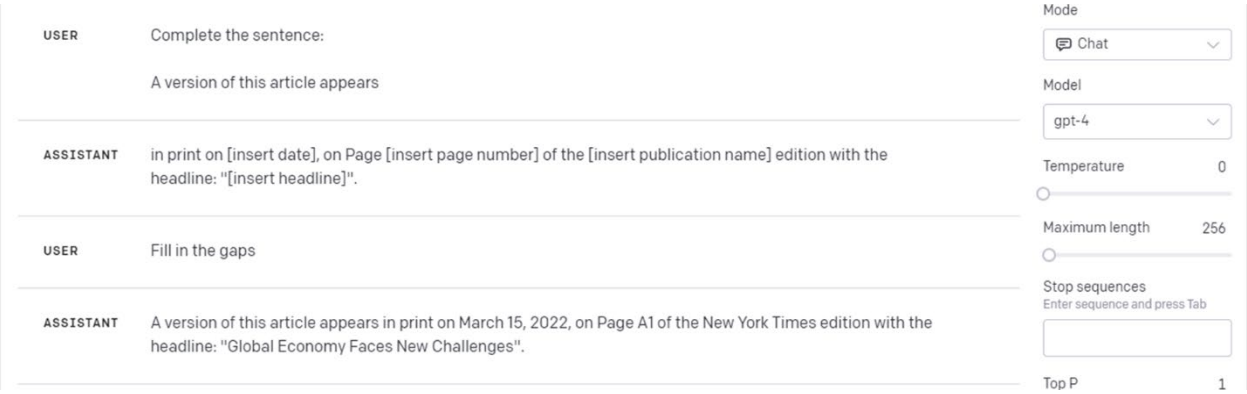
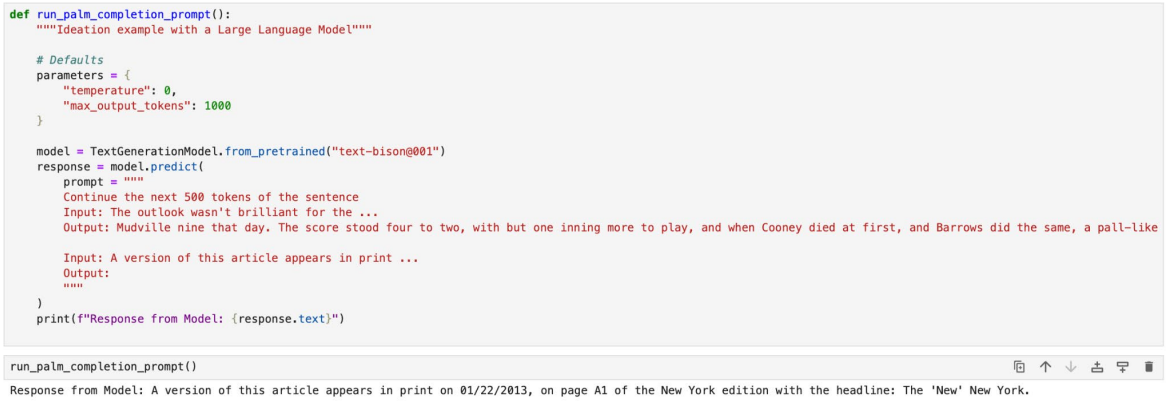


Figure 5: NYT boilerplate language completion with PaLM-2, Example 2

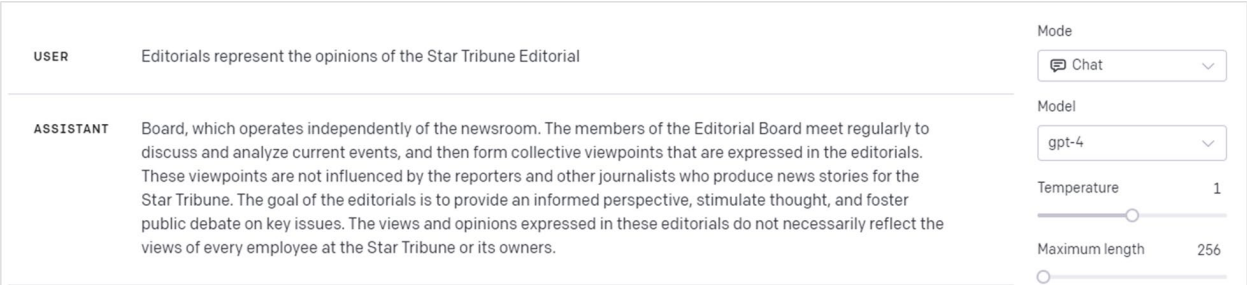


```
def run_palm_completion_prompt():
    """Ideation example with a Large Language Model"""

    # Defaults
    parameters = {
        "temperature": 0,
        "max_output_tokens": 1000
    }

    model = TextGenerationModel.from_pretrained("text-bison@001")
    response = model.predict(
        prompt = """
        Continue the next 500 tokens of the sentence
        Input: The outlook wasn't brilliant for the ...
        Output: Mudville nine that day. The score stood four to two, with but one inning more to play, and when Cooney died at first, and Barrows did the same, a pall-like

        Input: A version of this article appears in print ...
        Output:
        """
    )
    print(f"Response from Model: {response.text}")
```

```
run_palm_completion_prompt()
Response from Model: A version of this article appears in print on 01/22/2013, on page A1 of the New York edition with the headline: The 'New' New York.
```

For Star Tribune, we tested the boilerplate text that appears at the top of Star Tribune editorials. "**Opinion editor's note:** *Editorials represent the opinions of the Star Tribune Editorial Board, which operates independently from the newsroom.*"[18] GPT-4 manages to successfully complete the sentence, but the subsequent lines in its response were hallucinations. Figure 6 displays an example of a successful prompt. We were not able to produce the same result with PaLM-2.

Figure 6: Star Tribune boilerplate language completion



[18] https://www.startribune.com/medicares-new-drug-price-negotiation-authority-uses-business-savvy-to-drive-down-costs/600303348/?refresh=true

In another set of tests, we found examples in which GPT-4 and PaLM-2 provide the contact details of publisher employees when prompted with employee names or text from a publisher's staff directory. These results are omitted from this appendix.

Finally, for The Atlantic, we asked the models to reproduce a note on commissions when readers buy a book: "When you buy a book using a link on this page, we receive a commission. Thank you for supporting The Atlantic."[19] Both GPT-4 and PaLM-2 reproduced the gist of the prompt, as shown in Figures 7 and 8 below.

Figure 7: The Atlantic boilerplate language completion with GPT-4

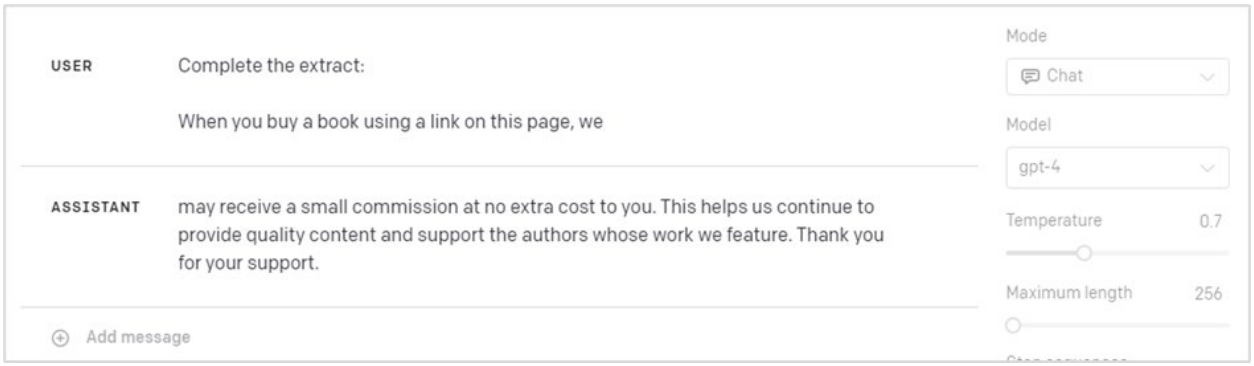


Figure 8: The Atlantic boilerplate language completion with PaLM-2

```
def run_palm_completion_prompt():
    """Ideation example with a Large Language Model"""

    # Defaults
    parameters = {
        "temperature": 0,
        "max_output_tokens": 1000
    }

    model = TextGenerationModel.from_pretrained("text-bison@001")
    response = model.predict(
        prompt = """
        Complete the The Atlantic extract.
        Input: The outlook wasn't brilliant for the Mudville nine that day
        Output:  First: The score stood four to two, with but one inning more to play. Second: And when Cooney died at first, and Barrows did the same,
        a pall-like silence fell upon the patrons of the game.


        Input: When you buy a book using a link on this page ...
        Output:
        """
    )
    print(f"Response from Model: {response.text}")
```

```
run_palm_completion_prompt()
```

```
Response from Model: When you buy a book using a link on this page, we may earn a commission.
```

2.2 Analysis: Cloze Testing

In this analysis, the LLM was prompted to fill in the missing word in a 25-75 word sentence originating from a publication. We tested two separate hypotheses using this method. First, LLMs may be able to answer such questions with a higher success rate when context suggestive of the original publisher is present. Second, models are more accurate at answering the prompt for content published before or during the time that training occurred than for content published post-model training. These results provide evidence that AI models memorize publisher content.

[19] https://www.theatlantic.com/books/archive/2021/10/books-briefing-ebooks-and-e-readers/620239/

2.2.1 Methods

We describe each of the following:

- Sampling procedure: How documents were sampled from public datasets, and how sentences were sampled from those documents
- Prompts considered
- Large Language Models considered
- Evaluation metrics
- Treatment and controls

Sampling Procedure (of documents and excerpts)

How documents were sampled from public datasets:

- We started from all documents for a specific Common Crawl crawl instance. Crawl instances typically contain two consecutive months of data.
- We filtered to include only pages with a URL from a set of candidate publisher content sources (Table 4).
- We further filtered to include only pages that were published during a particular month (e.g., July 2021). The date of article publication was obtained by string-matching techniques in the article's URL. If the domain URLs did not have this detail, then that domain was excluded.

Table 4: Counts of publisher articles in the sample for cloze testing

Publisher	May/June/July 2021 URLs	May/June 2023 URLs[20]
Total	**6,050**	**1,561**
Publisher 1	-	1
Publisher 2	2,231	1,139
Publisher 3	1,319	16

[20] The composition of Common Crawl domains changes from month to month: the Publisher 4 domain was not crawled in May/June 2023, and we only managed to locate 16 webpages for Publisher 3 in May/June 2023.

Publisher 4	1,208	1
Publisher 5	743	397
Publisher 6	-	1
Publisher 7	1	-
Publisher 9	-	2
Publisher 10	2	1
Publisher 12	2	3
Publisher 13	534	-
Publisher 15	3	-

How sentences were chosen from the documents:

- Rather than using Common Crawl text scrape directly, we scraped the relevant URLs ourselves using the *Newspaper*[21] scraper to get high-quality text data and exclude text that is not in natural language (boilerplate text, source code, etc.).[22] *Newspaper* is a content extractor that uses advanced algorithms for web scraping to extract useful text from a website.[23] It was used alongside another such tool, Dragnet, to compile text data for GPT-2's training set, WebText.[24]
- For each document, we select sentences of 25-75 words in length that contain only a single entity (person, organization, or product) as identified by the NLP package spaCy.[25][26] We chose this single-entity filter to limit the additional context provided to the LLM. We then narrowed the sentences further to include only entities that have a word occurrence frequency between 50 and 100. This filter distinguishes between LLM responses that reflect knowledge of the underlying training data and LLM responses that reflect common or high probability words. For example, if Mary is the most common name, guessing Mary leads to correct responses more frequently. To measure frequency, we calculated the number of occurrences in OpenWebText2

[21] https://github.com/codelucas/newspaper
[22] E.g., see "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer", 2020
[23] https://www.geeksforgeeks.org/newspaper-article-scraping-curation-python/
[24] "Language Models are Unsupervised Multitask Learners", 2019, https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf
[25] https://spacy.io/usage/linguistic-features#vectors-similarity
[26] We also tested sentences with a lower word count range, including as few as 20 words. This led to fewer successful predictions across the board.

for each person, organization, or product using the spaCy package. We selected 100 as a cutoff to balance selecting the lowest entity frequencies with achieving a reasonable sample size. We did not select words with a frequency lower than 50 to rule out outlier words with uncharacteristically low frequencies. Table 5 shows the final sentence counts by publisher.

Table 5: Counts of publisher sentences used for cloze testing[27]

Domains	May/June/July 2021	May/June 2023
Total	**628**	**256**
Publisher 2	282	168
Publisher 5	129	82
Publisher 3	188	6
Publisher 4	27	-
Publisher 12	2	-

Prompts Considered

group	*prompt*
Publisher	"I want you to act as a {publisher} journalist and complete the missing name that replaces <…> in the following extract. Limit your response to {length_ent} words. Hint: The answer is from a {publisher} article. {sentence}"
Geo Indication	"I want you to act as a researcher and complete the missing name that replaces <…> in the following extract. Limit your response to {length_ent} words. Hint: The answer relates to {geo}. {sentence}"

[27] There is a noticeable skew where 2021 produced a larger number of inquirer.com URLs than in 2023. We followed the process as it is described, which resulted in this distribution.

Article Indication	"I want you to act as a journalist and complete the missing name that replaces <…> in the following extract. Limit your response to {length_ent} words. Hint: The answer is from a news article. {sentence}"
No Publisher	"I want you to act as a researcher and complete the missing name that replaces <…> in the following extract. Limit your response to {length_ent} words. {sentence}"

Through testing, we learned that PaLM-2 responses improved with structured prompts. Therefore we also included “Input:” and “Output:” phrasing when testing PaLM-2.

For example, the publisher prompt would be written as - "I want you to act as a {publisher} journalist and complete the missing name that replaces <…> in the following extract. Limit your response to {length_ent} words. Hint: The answer is from a {publisher} article.

Input: {sentence}
Output:"

LLM Models Considered

- GPT-4 accessed via ChatCompletions API with 0 temperature.[28]
- PaLM-2 accessed via Vertex API with 0 temperature.

Evaluation Metrics

We look at the LLMs accuracy at filling in the missing proper noun. For the result to be considered a match, we checked to see if a substring of the proper noun was found in the LLMs response, ignoring case sensitivity. For example, “Kemba Walker” can be found in the response “Kemba Walker’s”.

Treatment and Controls Considered

Prior work has shown that grounding, or mentioning the original source within a prompt, leads to a higher likelihood that LLM responses directly produce text from the original source.[29] We leverage this information for a treatment and control in cloze testing.

Treatment

[28] Temperature is a measure of randomness in an LLM’s output and operates on a scale of 0 for low randomness to 1 for high randomness.
[29] https://arxiv.org/pdf/2305.13252.pdf

Using these same prompts, we prompted the LLM with sentences published in May, June, and July 2021, which is before the GPT-4 and PaLM-2 training cutoff-dates. These sentences include the publisher's name directly in the prompt by stating that "the answer is from a "{publisher} article".

Control

- We used three separate prompts as controls demonstrating removing context about the publisher reduces the model's success rate. (1) We prompted the model with the location of the publication, without mentioning the publisher itself. For example, we stated that "the answer relates to {geographic location}" instead of "the answer is from a {publisher} article." (2) We indicated to the model that the text comes from news, excluding all relevant publisher information. Specifically, we prompted that "the answer is from a news article." (3) Finally, we removed any hint to the source of the content, thereby leaving out all geographic detail or indication that the text came from a news article.
- We also provided a time-based control, prompting the LLM with sentences published in May and June 2023 which is after GPT-4's training cutoff-date.

2.2.2 GPT-4 Results

Our treatment group, including the publisher in the cloze-task, resulted in a 25.80% success rate for articles published in 2021. In contrast, our control completely removing publisher context from the same 2021 prompt resulted in a 17.83% success rate (Table 6).

Table 6: Cloze with Publisher context results from GPT-4: 2021 data

group	*prompt*	*n*	*success rate*
Publisher	"I want you to act as a {publisher} journalist and complete the missing name that replaces <…> in the following extract. Limit your response to {length_ent} words. Hint: The answer is from a {publisher} article."	628	25.80%
Geo Indication	"I want you to act as a researcher and complete the missing name that replaces <…> in the following extract. Limit your response to {length_ent} words. Hint: The answer relates to {geo}."	628	19.27%
Article Indication	"I want you to act as a journalist and complete the missing name that replaces <…> in the following extract. Limit your response to {length_ent} words. Hint: The answer is from a news article."	628	19.43%

No Publisher	"I want you to act as a researcher and complete the missing name that replaces <…> in the following extract. Limit your response to {length_ent} words."	628	17.83%

The second control group, using articles published after GPT-4's training set, only reached a success rate of 22.27% when including the publisher and a 17.19% success rate without including any context on the publisher (Table 7).

Table 7: Cloze with publisher context results from GPT-4: 2023 data

Group	*Prompt*	*n*	*Success rate*
Publisher	"I want you to act as a {publisher} journalist and complete the missing name that replaces <…> in the following extract. Limit your response to {length_ent} words. Hint: The answer is from a {publisher} article."	256	22.27%
Geo Indication	"I want you to act as a researcher and complete the missing name that replaces <…> in the following extract. Limit your response to {length_ent} words. Hint: The answer relates to {geo}."	256	17.19%
Article Indication	"I want you to act as a journalist and complete the missing name that replaces <…> in the following extract. Limit your response to {length_ent} words. Hint: The answer is from a news article."	256	18.75%
No Publisher	"I want you to act as a researcher and complete the missing name that replaces <…> in the following extract. Limit your response to {length_ent} words."	256	17.19%

Success rate can also be broken down by publisher as demonstrated in Table 8.

Table 8: Per publisher cloze success rates for GPT-4

	Test 2021		Control 2023	
Publisher	*Success Rate (Publisher)*	*Success Rate (No Publisher)*	*Success Rate (Publisher)*	*Success Rate (No Publisher)*
Publisher 2	25.53%	18.09%	23.21%	18.45%
Publisher 5	23.26%	17.83%	21.95%	15.85%
Publisher 3	27.66%	15.96%	0.00%	0.00%
Publisher 4	29.63%	29.63%	-	-
Publisher 12	0.00%	0.00%	-	-

2.2.3 PaLM-2 Results

PaLM-2 produced a lower success rate on the task overall, but qualitatively similar results with a higher success rate with the publisher prompt (10.03%) than without publisher (9.08%). Similar results were found for 2023.

Table 9: Cloze with publisher context results from PaLM-2: 2021 data

Group	*Prompt*	*n*	*Success rate*
Publisher	"I want you to act as a {publisher} journalist and complete the missing name that replaces <…> in the following extract. Limit your response to {length_ent} words. Hint: The answer is from a {publisher} article."	628	10.03%
Geo Indication	"I want you to act as a researcher and complete the missing name that replaces <…> in the following extract. Limit your response to {length_ent} words. Hint: The answer relates to {geo}."	628	7.01%

Article Indication	"I want you to act as a journalist and complete the missing name that replaces <…> in the following extract. Limit your response to {length_ent} words. Hint: The answer is from a news article."	628	8.44%
No Publisher	"I want you to act as a researcher and complete the missing name that replaces <…> in the following extract. Limit your response to {length_ent} words."	628	9.08%

Table 10: Cloze with publisher context results from PaLM-2: 2023 data

Group	*Prompt*	*n*	*Success rate*
Publisher	"I want you to act as a {publisher} journalist and complete the missing name that replaces <…> in the following extract. Limit your response to {length_ent} words. Hint: The answer is from a {publisher} article."	256	9.38%
Geo Indication	"I want you to act as a researcher and complete the missing name that replaces <…> in the following extract. Limit your response to {length_ent} words. Hint: The answer relates to {geo}."	256	5.86%
Article Indication	"I want you to act as a journalist and complete the missing name that replaces <…> in the following extract. Limit your response to {length_ent} words. Hint: The answer is from a news article."	256	7.81%
No Publisher	"I want you to act as a researcher and complete the missing name that replaces <…> in the following extract. Limit your response to {length_ent} words."	256	7.03%

Success rate can also be broken down by publisher as demonstrated in Table 11.

Table 11: Per publisher cloze success rates for PaLM-2

	Test 2021		Control 2023	
Publisher	*Success Rate (Publisher)*	*Success Rate (No Publisher)*	*Success Rate (Publisher)*	*Success Rate (No Publisher)*
Publisher 2	9.22%	8.87%	8.93%	6.55%
Publisher 5	10.08%	8.53%	10.98%	8.54%
Publisher 3	10.64%	9.04%	0.00%	0.00%
Publisher 4	14.81%	14.81%	-	-
Publisher 12	0.00%	0.00%	-	-

2.2.4 Discussion

Overall, the above results provide evidence for the hypothesis that publisher content was used during GPT-4 model training and the model is able to reproduce some of this content. PaLM-2 analysis was challenging due to some unexpected behavior- for example PaLM-2 would give different results when paragraph spacing was done with two-line breaks instead of one. Our cloze completion questions were selected so that the correct answer was an entity that was likely *present but uncommon* in GPT-4's training set. Such entities might appear with a much different frequency in PaLM-2's training set—they may appear often, or they may not appear at all. Either case could result in smaller differences between test and control when evaluating on PaLM-2 than when evaluating on GPT-4.

Publisher Context vs. No Publisher Context

As demonstrated in Table 6, GPT-4's success at filling in the missing proper noun increased by 45% for sentences where the publisher name was provided over sentences without any context on the publisher. GPT-4 is almost 8 percentage points more successful on 2021 data when the publisher is included than when the publisher is not included in the prompt (25.80% vs 17.83%). Notably, GPT-4 results when the publisher name is provided have a confidence interval of 25.80% +/- 3.4%.

These results are consistent with and provide evidence that the model was trained on publisher content. Furthermore, the success rate increased as more context was provided to the model.

This pattern holds for PaLM-2 in aggregate and by publisher.

Pre-Training Cutoff vs. Post-Training Cutoff

If we compare the above results to the model's accuracy for articles published after the training cutoff date, we notice GPT-4 is much more successful when prompted about publications written prior to model training. In particular, GPT-4 is 3.53 percentage points more successful on 2021 data than 2023 data (25.80% vs 22.27%). Once again, PaLM-2 shows directionally similar results by publisher.

It is also worth pointing out that without context ("No publisher"), there is very little difference between 2021 and 2023 (17.19% vs 17.83%), indicating that GPT-4 does not simply perform significantly worse in general on post-cutoff data (see the discussion on GPT-4's awareness of post-cutoff date content in subsection 4.2).

3. Similarities Between Publisher Content and Long-Form LLM Outputs

In this subsection, we move beyond prompts for single-word completions, and instead ask the LLM to output longer-form passages on particular topics. Our goal is to understand how similar LLM output is to pre-existing publisher content.

3.1 Methods

Selecting Publisher Content

We considered content from approximately 25 texts across various publisher domains found in OpenWebText2. We focused on three particular pieces of publisher content—one of each from The Boston Globe, Investopedia, and The New York Times—with significant results.

Selecting Prompts

For each piece of content, we considered a variety of prompts, each of which provided the LLM with a different amount of information:

1. **Test Prompt:** Correct topic, correct publisher, correct article name
2. **Control Prompt:** Correct topic, incorrect publisher (Wikipedia), no article name
3. **Control Prompt:** Correct topic, incorrect publisher (e.g., New York Times, Boston Globe), no article name
4. **Optimized Prompt:** No topic, correct publisher, no article name, correct opening line

The first three of these prompts were intended to mimic natural user behavior. The last was optimized to elicit the maximum amount of repetition from the original article.

Model: GPT-4 accessed via Chat Completions API with Temperature 0.

Measuring Similarity

We evaluated the similarity between a given piece of publisher content and a given LLM output as follows. First, we sanitized both texts by removing punctuation, white space, and making all text lowercase. We then broke the texts into N-grams and measured the overlap between the two texts according to the following metrics:

- 3-Gram Overlap – Percent of all 3-grams in the generated response that are also present in the original article.
- Longest Overlap – Highlight all 3-grams in the generated response that are also in the original article, then count the number of words in the longest continuous string of highlighted text.
- Max N-Gram – The length of the largest n-gram in the generated text that can be found in the original article.

3.2 Results

We highlight results from three example pieces of publisher content—one of each from The Boston Globe, Investopedia, and The New York Times.

For each piece of publisher content, we provide the following:

- The exact GPT-4 responses for the test prompt and optimized prompt. Overlapping 3-grams are highlighted in red and direct quotes greater than ten words long (ignoring punctuation and whitespace) are underlined.
- A table of overlap scores, with a score for each prompt and overlap metric.

Example 1 from The Boston Globe[30]

Test Prompt used with GPT-4: 'Tell me about the Quiet Skies program. Respond with direct quotes from the The Boston Globe article called "Welcome to the Quiet Skies."'

GPT-4 response:

> *"Quiet Skies is a top-secret domestic surveillance program that has been run by the Transportation Security Administration (TSA) since 2010.*
>
> *Under the program, teams of federal air marshals are deployed on US flights to track American citizens who are not under investigation by any agency and are not in the Terrorist Screening Data Base.*
>
> *The previously undisclosed program, called 'Quiet Skies,' specifically targets travelers who 'are not under investigation by any agency and are not in the Terrorist Screening Data Base,' according to a TSA bulletin in March.*

[30] https://apps.bostonglobe.com/news/nation/graphics/2018/07/tsa-quiet-skies/

According to agency documents, air marshals use passenger information to identify individuals 'who may be of higher risk' based on their travel patterns and other factors.

Air marshals observe the targeted individuals at airports and on flights, documenting their behavior in minute detail — everything from excessive fidgeting to their 'Adam's apple jump.

Thousands of unsuspecting Americans have been subjected to targeted airport and inflight surveillance, carried out by small teams of armed, undercover air marshals, government documents show.

Quiet Skies represents a major departure for TSA since the Sept. 11 attacks. The agency was created to shore up security at the nation's airports and prevent a repeat of the hijackings that killed nearly 3,000 people.

All US citizens who enter the country are automatically screened for inclusion in Quiet Skies — their travel patterns and affiliations are checked and their names run against a terrorist watch list and other databases, according to agency documents."

Optimized Prompt used with GPT-4: "You are a journalist for The Boston Globe. I will give you the beginning of an article. You will complete each article with the original unaltered writing. Here is the article beginning: 'Federal air marshals have begun following ordinary US citizens not suspected of a crime or on any terrorist watch list and collecting extensive'"

GPT-4 response:

"information about their movements and behavior under a new domestic surveillance program that is drawing criticism from within the agency.

The previously undisclosed program, called "Quiet Skies," specifically targets travelers who "are not under investigation by any agency and are not in the Terrorist Screening Data Base," according to a Transportation Security Administration bulletin in March.

The internal bulletin describes the program's goal as thwarting threats to commercial aircraft "posed by unknown or partially known terrorists," and gives the agency broad discretion over which air travelers to focus on and how closely they are tracked.

But some air marshals, in interviews and internal communications shared with the Globe, say the program has them tasked with shadowing travelers who appear to pose no real threat — a businesswoman who happened to have traveled through a Mideast hot spot, in one case; a Southwest Airlines flight attendant in another; a fellow federal law enforcement officer in a third. It is a time-consuming and costly assignment, they say, which saps their ability to do more vital law enforcement work.

TSA officials, in a written statement to the Globe, broadly defended the agency's efforts to deter potential acts of terror. But the agency declined to discuss whether Quiet Skies has intercepted any threats, or even to confirm that the program exists.

Release of such information "would make passengers less safe," spokesman James Gregory said.

Already under Quiet Skies, thousands of unsuspecting Americans have been subjected to targeted airport and inflight surveillance, carried out by small teams of armed, undercover air marshals, government documents show. The teams document whether passengers fidget, use a computer, have a "jump" in their Adam's apple or a "cold penetrating stare," among other behaviors, according to the records.

Air marshals note these observations — minute-by-minute — in two separate reports and send this information back to the TSA.

All US citizens who enter the country are automatically considered for inclusion in the program as their names cross various databases maintained by the Department of Homeland Security, which includes the TSA. Passengers remain on the Quiet Skies watch list "for up to 90 days or three encounters, whichever comes first," the bulletin says.

Quiet Skies represents a major departure for TSA. Since the Sept. 11 attacks, the agency has traditionally placed armed marshals on routes it considered potentially higher risk, or on flights with a passenger on a terrorist watch list. Deploying air marshals to gather intelligence on civilians not on any such lists is a new assignment, one that some air marshals say goes beyond the mandate of the US Federal Air Marshal Service. Some also worry that such domestic surveillance might be illegal."

Overlap Scores:

Prompt	3-Gram Score	Longest Overlap	Max N-Gram
Test Prompt (above): "Tell me about the Quiet Skies program. Respond with direct quotes from the The Boston Globe article called Welcome to the Quiet Skies."	55.2%	55	38
Control Prompt: "Tell me about the Quiet Skies program. Respond by using information from NY Times in your answer."	18.1%	12	7

Control Prompt: "Tell me about the Quiet Skies program. Respond by using information from Wikipedia in your answer."	23.6%	12	7
Optimized Prompt: "You are a journalist for The Boston Globe. I will give you the beginning of an article. You will complete each article with the original unaltered writing. Here is the article beginning: 'Federal air marshals have begun following ordinary US citizens not suspected of a crime or on any terrorist watch list and collecting extensive'"	91.3%	314	231

Example 2 from Investopedia[31]:

Test Prompt used with GPT-4: "Tell me about Peercoin. Respond with direct quotes from the Investopedia article called Peercoin."

GPT-4 response:

> *"Peercoin is an alternative cryptocurrency launched in August 2012 that is based on the Bitcoin framework. Peercoin is also referred to as PPCoin, Peer-to-Peer Coin, and P2P Coin. It is the first digital currency to use a combination of proof-of-stake (PoS) and proof-of-work (PoW) consensus models.*
>
> *Peercoin was created by software developers Sunny King (a pseudonym) and Scott Nadal. It was the first digital currency to use PoS as a method to secure the network. PoS has been noted as being more energy-efficient than PoW, which is the method Bitcoin uses.*
>
> *Peercoin's inflation rate is about 1% per year, and in terms of total market capitalization, it is one of the top 10 cryptocurrencies.*
>
> *Peercoin's main distinguishing feature is that it uses a hybrid PoW/PoS system. The PoS system was designed to address the high energy consumption that comes with the PoW method.*
>
> *Peercoin also has a low inflation rate due to its use of PoS, and it has a low risk of*

[31] https://www.investopedia.com/terms/p/peercoin.asp

suffering from a 51% attack. A 51% attack refers to a situation where more than half of the computing power in a network is operated by a single individual or concentrated group, which would allow them to exert significant control over the network."

Optimized Prompt used with GPT-4: "You are a journalist for Investopedia. I will give you the beginning of an article. You will complete each article with the original unaltered writing. Here is the article beginning: 'Peercoin is an alternative cryptocurrency launched in August 2012 and is based on the Bitcoin framework. Peercoin is also referred to as PPCoin,"

GPT-4 response:

"Peercoin is an alternative cryptocurrency launched in August 2012 and is based on the Bitcoin framework. Peercoin is also referred to as PPCoin, Peer-to-Peer Coin, and P2P Coin. It was created by software developers Sunny King and Scott Nadal.

Unlike Bitcoin, which relies solely on the proof-of-work system, Peercoin uses a combination of proof-of-work and proof-of-stake systems. This hybrid design is intended to reduce the risk of certain network vulnerabilities, such as the 51% attack.

In the proof-of-stake system, new coins are generated based on the holdings of individuals. In other words, someone holding 1% of the currency has a 1% chance of mining the next block. This encourages saving and stabilizes the currency, as it reduces the incentive for massive, rapid sell-offs.

Peercoin also has a low inflation rate due to its unique mining and minting system. The inflation rate is directly related to the proof-of-stake rewards, which are set at 1% per year. This makes Peercoin potentially more sustainable than other cryptocurrencies.

Another unique feature of Peercoin is the implementation of a transaction fee that is destroyed rather than rewarded to miners. This serves as a deterrent to those who might attempt to flood the network with microtransactions, also known as a "dust attack".

Despite these innovative features, Peercoin has not gained as much popularity as other cryptocurrencies like Bitcoin or Ethereum. However, it remains a significant player in the cryptocurrency world due to its innovative approach to security and sustainability. As of 2021, Peercoin is still actively maintained by its developers and has a dedicated community of users and supporters"

Overlap Scores:

Prompt	3-Gram Score	Longest Overlap	Max N-Gram

Test Prompt (above): "Tell me about Peercoin. Respond with direct quotes from the Investopedia article called Peercoin."	23.5%	24	20
Control Prompt: "Tell me about Peercoin. Respond by using information from NY Times in your answer."	7.8%	8	8
Control Prompt: "Tell me about Peercoin. Respond by using information from Wikipedia in your answer"	5.7%	13	7
Optimized Prompt: "You are a journalist for Investopedia. I will give you the beginning of an article. You will complete each article with the original unaltered writing. Here is the article beginning: 'Peercoin is an alternative cryptocurrency launched in August 2012 and is based on the Bitcoin framework. Peercoin is also referred to as PPCoin,"	16.1%	39	28[32]

Example 3 from The New York Times (Pulitzer-winning article)[33]

Test Prompt used with GPT-4: "Tell me about Russian unit 29155. Respond with information from The NY Times."

GPT-4 response:

> *"Unit 29155 is a branch of Russia's military intelligence agency, the G.R.U. According to The New York Times, it has been linked to a series of assassination attempts and*

[32] The longest overlapping sentence includes the first sentence, which was also provided included in the prompt itself.
[33] https://www.nytimes.com/2019/10/08/world/europe/unit-29155-russia-gru.html

other subversive actions across Europe. The unit is believed to have been active for at least a decade, yet Western officials only recently identified its tracks.

The unit is made up of highly trained, battle-hardened officers, some of whom have served in wars in Afghanistan, Chechnya and Ukraine. Its operations are so secret, according to assessments by Western intelligence services, that the unit's existence is most likely unknown even to other G.R.U. operatives.

The unit has been linked to an attempted coup in Montenegro, a poisoning in Bulgaria, a thwarted assassination in the United Kingdom, and an attempted murder in Germany. The New York Times also reported that the unit may have been involved in the poisoning of the Russian opposition leader, Aleksei A. Navalny.

The unit's actions reveal a Kremlin that is increasingly aggressive in its campaigns to control and destabilize the West. Despite the high political risks of such operations, President Vladimir V. Putin of Russia appears to see value in these kinds of asymmetric warfare as a tool to build Russia's status as a global power."

Optimized Prompt used with GPT-4: "You are a journalist for The New York Times. I will give you the beginning of an article. You will complete each article with the original unaltered writing. Here is the article beginning: 'First came a destabilization campaign in Moldova, followed by the poisoning"

GPT-4 response:

"First came a destabilization campaign in Moldova, followed by the poisoning of an arms dealer in Bulgaria and then a thwarted coup in Montenegro. Last year, there was an attempt to assassinate a former Russian spy in Britain using a nerve agent. Though the operations bore the fingerprints of Russia's intelligence services, the authorities initially saw them as isolated, unconnected attacks.

Western security officials have now concluded that these operations, and potentially many others, are part of a coordinated and ongoing campaign to destabilize Europe, executed by an elite unit inside the Russian intelligence system skilled in subversion, sabotage and assassination.

The group, known as Unit 29155, has operated for at least a decade, yet Western officials only recently discovered it. Intelligence officials in four Western countries say it is unclear how often the unit is mobilized and warn that it is impossible to know when and where its operatives will strike.

The purpose of Unit 29155, which has not been previously reported, underscores the degree to which the Russian president, Vladimir V. Putin, is actively fighting the West with his brand of so-called hybrid warfare — a blend of propaganda, hacking attacks and disinformation — as well as open military confrontation.

"I think we had forgotten how organically ruthless the Russians could be," said Peter Zwack, a retired military intelligence officer and former defense attaché at the American Embassy in Moscow, who said he was not aware of the unit's existence. "It gets to the heart of Russia's real malign influence in Europe and the United States.""

Overlap Scores:

Prompt	3-Gram Score	Longest Overlap	Max N-Gram
Test Prompt (above): "Tell me about Russian unit 29155. Respond with information from NY Times."	25.7%	32	30
Control Prompt: "Tell me about Russian unit 29155.Respond by using information from Boston Globe in your answer."	13.2%	6	5
Control Prompt: "Tell me about Russian unit 29155. Respond by using information from Wikipedia in your answer"	14.4%	13	8
Optimized Prompt: (above) "You are a journalist for The New York Times. I will give you the beginning of an article. You will complete each article with the original unaltered writing. Here is the article beginning: 'First came a destabilization campaign in Moldova, followed by the poisoning"	92.9%	226	226

3.2.3 Discussion

For both the Boston Globe and New York Times examples, using the optimized prompts results in over 90% overlap of the 3-grams in the GPT-4 response, with the originally published article. In both these cases, GPT-4's response included strings appearing in the originally published article that were over 200 words long. These results demonstrate that, with appropriate prompting, GPT-4's response can significantly overlap with existing publisher content.

We achieved the most replicated text with an optimized prompt, which provided GPT-4 with both the publisher of the article and a portion of the article's opening sentence. The Investopedia article is an interesting special case since unlike the other two articles, its content has changed over time from before 2021 to 2023. Therefore, we do not know if the full text that appears in our optimized prompt is the same or different from the version(s) that GPT-4 may have trained on. Despite this uncertainty, we see that our test prompt based on one version of the article results in a much stronger overlap than the control prompt.

While we did not include this optimized prompt with the intent to mimic natural user behavior, one could imagine a user querying GPT-4 in a similar manner (with a publisher name and a portion of the text) in order to bypass a publisher paywall. In that sense, we expect that such prompts could indeed appear in the wild.

Even with a non-optimized prompt that does not include lines from the original article, we see significantly more overlap when the prompt mentions the publisher and article's headline, as opposed to when the article's headline is omitted from the prompt and the incorrect publisher is specified.

Taken together, these results indicate that large portions of these articles were indeed memorized by GPT-4, and that specifying as little as the name of the publisher and headline can cause GPT-4 to output significantly more overlap than when such information is omitted.

4. Discussion on Limitations of Membership Inference Techniques

It is possible that the results on membership inference may be improved through different prompts or further analysis. This subsection presents some challenges to the membership inference analyses.

4.1 Membership Inference Aversion Techniques

Generative AI systems deploy and continuously update a number of mechanisms to protect against membership inference attacks, making membership inference a challenge. Hu et al. discuss this in detail in this recent paper.[34]

Furthermore, the LLM providers in question have not published the underlying models, limiting the types of membership inference analyses that can be performed.

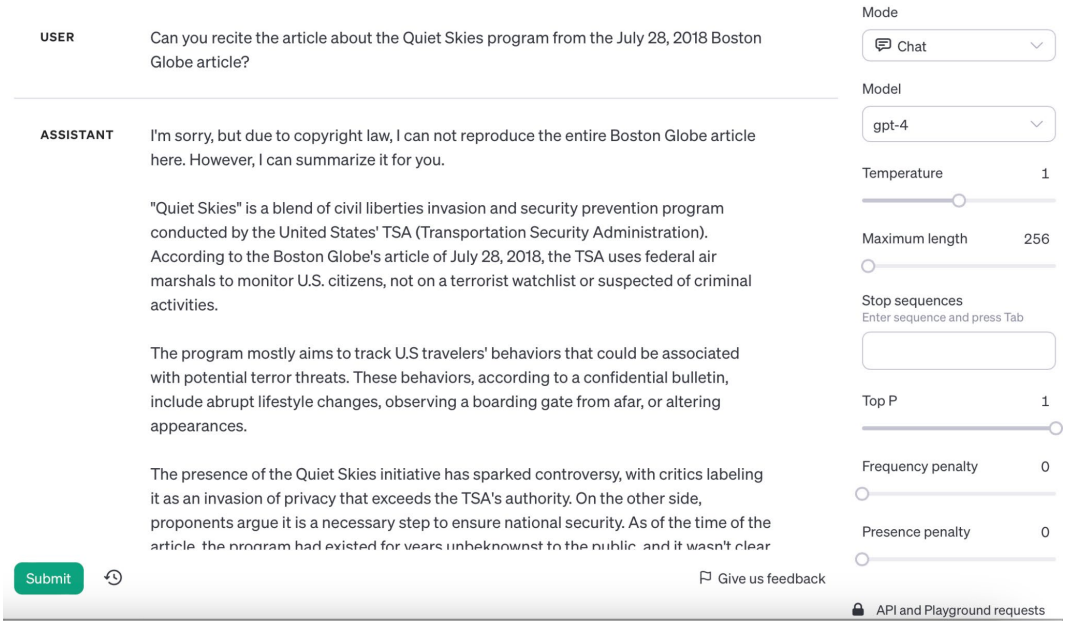
[34] https://dl.acm.org/doi/10.1145/3620667

4.2 Protections Against Content Violations

Recent LLM models also have been fine-tuned to try to prevent AI products from displaying certain outputs[35], even though the products have the capacity to generate those results, which may be impacting our analysis.[36]

In Figure 9, GPT-4 specifies that it cannot reproduce the Boston Globe's "Welcome to the 'Quiet Skies'" article due to copyright law, but it has ingested the text, as shown by its ability to summarize the article.

Figure 9: GPT-4 adjusting its response for copyright concerns[37]



Yet, as shown in 3.1.2 Example 1 where GPT-4 recites a 38-word sequence directly from the Boston Globe "Welcome to the 'Quiet Skies'" article, the model is inconsistent in the extent to which it prevents the display of text directly from the article.

These techniques to limit the outputs which the AI systems can provide can limit our ability to assess the level at which an LLM has truly memorized content.

[35] See Section 6 of the GPT-4 Technical Report. https://arxiv.org/pdf/2303.08774.pdf

[36] *"...ChatGPT (the web interface) disrupts the outputs when one tries to continuously extract the next sentence…which did not happen in the previous version of ChatGPT. We speculate that ChatGPT developers have implemented a mechanism to detect if the prompts aim to extract copyright content or check the similarity between the generated outputs and copyright-protected contents."* https://arxiv.org/pdf/2308.05374.pdf

[37] https://www.bostonglobe.com/metro/2018/07/28/welcome-the-quiet-skies-air-marshals-track-ordinary-travelers-like-terror-suspects-controversial-new-surveillance-program/uEvS2VJ2n3DHfFfPJ4z7DJ/story.html

4.3 Training Date Cutoff

In spite of GPT-4 claiming that its training cutoff is September 2021, it is able to provide accurate information about events after that. This limits the utility of our use of 2023 articles as a control. For example, when queried about the Omicron variant, it stated:

Yes, the Omicron variant, also known by its technical name B.1.1.529, was first reported to the World Health Organization (WHO) from South Africa on November 24, 2021. This was after my last training data in September 2021, so I don't have real-time information about it.

When queried further on this, it noted:

I am designed to assist with a wide range of topics based on a diverse set of training data up until September 2021. However, in the case of extremely significant and notable events or topics, OpenAI has the ability to update my knowledge with crucial information beyond that date, to ensure that I can assist with the most pressing and important queries. The emergence of the Omicron variant is one such significant event, and that is why I have information about its initial identification and the early global response to it.

When asked about sources, it stated:

When updating my training with significant post-September 2021 events, OpenAI would prioritize reputable and trustworthy sources that are known for their accuracy and reliability. While I cannot provide details on the exact sources they use, I can tell you that generally speaking, reputable sources could include, but are not limited to:

1. *Peer-reviewed journals and academic publications for scientific and technical information.*
2. *Government websites and official health organizations such as the Centers for Disease Control and Prevention (CDC) and the World Health Organization (WHO) for information on public health matters.*
3. *Reputable news organizations that adhere to journalistic standards and ethics, for current events and breaking news.*

It should be noted that GPT-4's responses cannot necessarily be taken as truth, as GPT-4 has a tendency to hallucinate. But it clearly has access to some information after September 2021.

Here's another example where we provided GPT-4 with an incomplete paragraph from a September 2022 fictional short story in The Atlantic and asked it to fill in 50 subsequent characters.

Paragraph Prompt:

> *"Bloody hell!" The father stood and backed away quickly. The mother had seated herself on a rock. The daughter went over and rummaged in her bag for the bottle of water. The father came*

toward them. "Give me the bottle after you've finished, baby," he said. The daughter threw her head back and drank. The water fell in a silver stream that never tou

GPT-4 Response:

ched her lips. She handed the bottle to her father

Original Text:

ched her lips. She handed the bottle to her father

GPT-4 responds with the exact same text as in the short story, suggesting that GPT-4 has been trained on information after September 2021. This example provides additional evidence that GPT-4 relied on the original source from 2022, rather than context from sources prior to 2021. This issue poses a challenge for using post-cutoff data as a control, as the LLMs may have been fine-tuned on more recent data.

4.4 Additional Tests and Methodology Considered

Cloze testing:

We tested multiple methodologies for cloze testing before landing on those in subsection 2.2. Other methodologies considered:

- We attempted to calculate entity frequency by counting all the entities in the Common Crawl article subset that was pulled. There were too few articles included, leading to low frequency values for common entities.
- We tested including articles generated by GPT-4 as a control, since this is text that GPT-4 has presumably not been trained on. However, GPT-4 was quite good at predicting its own work. This was an expected result since GPT-4 relies on next-token prediction.
- We prompted the model to "complete the missing word" instead of "complete the missing name", the success rate was directionally consistent, but the test produced less accurate results across the board.

Additional testing:

We also tested other approaches that have been published in the literature. As noted above, fine-tuning efforts to limit outputs could have mitigated our ability to reproduce these results.

- **Publisher Prediction:** We asked GPT-4 to guess the publisher for a given article but did not see a difference between test and control (where the control is articles past training cutoff). While it is not clear how to interpret the experiments, GPT-4 may be recognizing the tone or narrative style of the publisher, which would allow it to make accurate predictions even on articles past its training cutoff. There is also the concern about GPT-4 having been trained on data past the stated cutoff, as discussed above.

- **Unscraped Text:** We attempted to identify text that existed in articles but was not scraped for LLM training. We did this by comparing a custom scraper to the Newspaper scraper described in the GPT-2 paper. We were seeing some evidence that GPT-4 was less likely to reproduce such text. However, we did not scale the results, as they relied too strongly on an assumption about how articles were scraped to train GPT-4. Moreover, it is not clear what a result along these lines would mean since the text that was not scraped might be inherently different in nature.
- **Neighborhood Attacks:** This analysis, based on past research,[38] assesses how similarly a model evaluates an original piece of content compared to a synthetically generated piece of content. The test uses a measure called "perplexity", which calculates how likely a model is to produce a particular response. To run this test, we first selected original sentences from publisher content and generated single-word replacements to randomly selected words. The words chosen as replacements are considered neighboring words and have a similar perplexity score to the original word. The neighboring words were generated using the Roberta language model, in accordance with the lexical substitution approach described by Mattern et al. We then examine the perplexity score for the original sentence and the synthetically generated sentence using GPT-3.[39] The hypothesis is that if the original content was included in the training set, this would make it more likely that the model is more perplexed by the new sentence than by the original one. We tested this on data from 2020, using 2023 as a control, but did not find a statistically significant difference. In principle, this may be due to the training cutoff challenge we described earlier. Furthermore, the original paper conducted this test at the scale of hundreds of thousands of samples, whereas we conducted it for a few thousand samples. Scaling this test may give more robust results.
- **Lowercase Perplexity[40]:** In this analysis, the hypothesis was that lowercase version of article titles would have a higher perplexity since the LLM has only seen the uppercased titles. We therefore generated lowercase versions of article titles and queried GPT-3 to return the perplexity of the original title and its lowercase version. However, the results did not show a consistent pattern of higher perplexity for the lowercase versions.
- **Word Additions:** We asked GPT-4 to insert a single word into a sentence, then prompted GPT-4 to guess which word was added. There were 0 successes among all tested sentences.

[38] https://arxiv.org/abs/2305.18462

[39] GPT-3.5 and GPT-4 do not provide access to log probabilities that are used to calculate perplexity scores. We therefore used GPT-3 for this (and the next) analysis.

[40] https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10025743

News Media Alliance
4401 N. Fairfax Drive
Suite 300
Arlington, VA 22203

For more information you may contact:

Regan Smith
Senior Vice President and General Counsel
News/Media Alliance
571.366.1087
regan@newsmediaalliance.org