NEWS MEDIA
ALLIANCE

May 3, 2024

Via E-Mail

Suzanne Wilson
General Counsel and Associate Register of Copyrights
Maria Strong
Associate Register of Copyrights and Director of Policy and International Affairs
U.S. Copyright Office
101 Independence Avenue SE
Washington, DC 20559-6003

Re: Summary of ex parte meeting regarding Docket No. 2023-6, Artificial Intelligence and Copyright

Dear Ms. Wilson and Ms. Strong,

On April 29, 2024, the News/Media Alliance (N/MA) met via Zoom with the Copyright Office to discuss the Office's notice of inquiry and request for comments relating to Artificial Intelligence (AI) and Copyright dated August 24, 2023. The participants in the meeting on behalf of News/Media Alliance were Regan Smith and outside consultant Johannes Munter (J. Munter Consulting Ltd) (collectively, "N/MA"). Participants from the Copyright Office were Maria Strong, Andrew Foglia, Chris Weston, Ben Brady, Jenee Iyer, Emily Chapuis, John Riley, Brandy Karl, Nick Bartelt, Melinda Kern, and Isaac Klipstein.

In particular, N/MA briefed the Office on the functioning of generative AI products that ground their outputs through Retrieval-Augmented Generation (RAG), and related integration into consumer-facing search engines and AI assistants. N/MA also highlighted case law that may be applicable to the use of publisher content for such purposes. N/MA noted that while publishers have dealt with various other challenges in the last 20 years, including digital transformation and the unauthorized use of content by dominant online platforms, generative AI presents a whole new existential challenge. As documented in N/MA's White Paper on generative AI and written submissions to the Office,[1] generative AI developers make copies of copyrighted publisher material to extract its expressive content for model training purposes, embody or reproduce those copies in their models, and then commercialize those models to produce substitutional outputs, which does not qualify as fair use. The incorporation of RAG raises a new dimension of concern, unless the LLM has authorization to access and use the additional materials used to ground the AI.

N/MA explained that RAG is a functionality meant to increase the accuracy, relevance, and responsiveness of generative AI applications by querying external data sources to inform responses provided by the Large Language Model (LLM) in response to user prompts. N/MA shared examples (see Appendix A) of answers provided by traditional and RAG-enabled generative AI applications. N/MA noted that these applications can sometimes go behind publisher paywalls.

N/MA emphasized that, unless licensed, taking and using copyrighted content in this competitive, substitutional manner typically constitutes infringement and misappropriation. Many AI developers simply seem to be rushing and testing the boundaries of acceptable uses. But while generative AI technology is new, much of the underlying copyright law regarding the use of textual content is not. N/MA shared an overview of existing case law (see Appendix B), providing instances were unauthorized uses of non-fictional text were found to be infringing, with examples ranging from verbatim excerpts and abridgments, to summaries, paraphrasing, and non-literal copying of copyrighted content. N/MA stressed that voluntary licensing is a feasible and preferable solution to current AI developer demand for publisher content, including on a collective basis.

In addition, the conversation touched on the feasibility of the European Union's Text and Data Mining opt-out regime under the Directive on Copyright in the Digital Single Market as well as implementation of the transparency requirements included in the proposed EU AI Act. N/MA reiterated the importance of adequate transparency measures in the United States.

---

[1] For News/Media Alliance's White Paper and written submissions, *see 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*, NEWS/MEDIA ALLIANCE (Oct. 31, 2023), https://www.newsmediaalliance.org/generative-ai-white-paper/.

We appreciate the Office's consideration of the views discussed during our meeting and remain available to respond to any follow-up questions the Office may have.

Sincerely,

Regan Smith

Regan Smith
Senior Vice President & General Counsel
News/Media Alliance

APPENDIX A

RAG-Enabled AI Poses Additional Risks to News, Magazine, and Digital Publishers

Retrieval-augmented generation (RAG) is a technique to increase the accuracy, relevance, and responsiveness of generative artificial intelligence (AI) applications – including reducing the tendency of Large Language Models (LLM) to hallucinate and provide inaccurate, out-of-date, or incomplete responses to user queries. RAG grounds an LLM on external sources of information and supplements the LLM's training data by having the LLM call to separate, defined sources of content in response to a prompt. The integration of RAG in a generative AI application can facilitate access to current information and for the model to provide citations and links to the source materials. The external data sources queried by an LLM can range from internal enterprise data to searchable content on the public internet, including protected publisher content that has been scraped for search-indexing purposes. RAG is an increasingly common feature used by nearly all of the most popular LLM companies.

Because RAG is used to enhance the quality of the generative AI system responses, in many cases, including consumer-facing search or chatbot uses, RAG sources are ideally reliable and trustworthy information providers, such as internal enterprise data or news, magazine, and digital publishers. Recently, some generative AI systems have started using protected news publisher content to provide a service that directly competes with and substitutes for publishers' content in existing markets, compounding the harm for publishers whose content has already been misappropriated and used for training the LLMs without authorization in the first place.

By enabling retrieval of up-to-date, real-time content from third-party websites, RAG allows generative AI applications to more efficiently repurpose publisher content, including for questionable purposes. Already, we see RAG being used to amplify disinformation by unscrupulous actors to undermine our democratic processes, including the proliferation of fake news websites that repurpose content from legitimate publishers to support a particular candidate or a viewpoint. For example, recently, a researcher spent $105 to create an automated fake news website publishing thousands of AI-powered articles a day "with the partisan news coverage framing of my choice, nearly all rewritten without credit from legitimate news sources."[2] Together, technologies and activities discussed above fuel distrust in the new technologies, the media, and the public discourse. It is therefore no wonder that the public expresses substantial skepticism towards AI with nearly three quarters of respondents in a

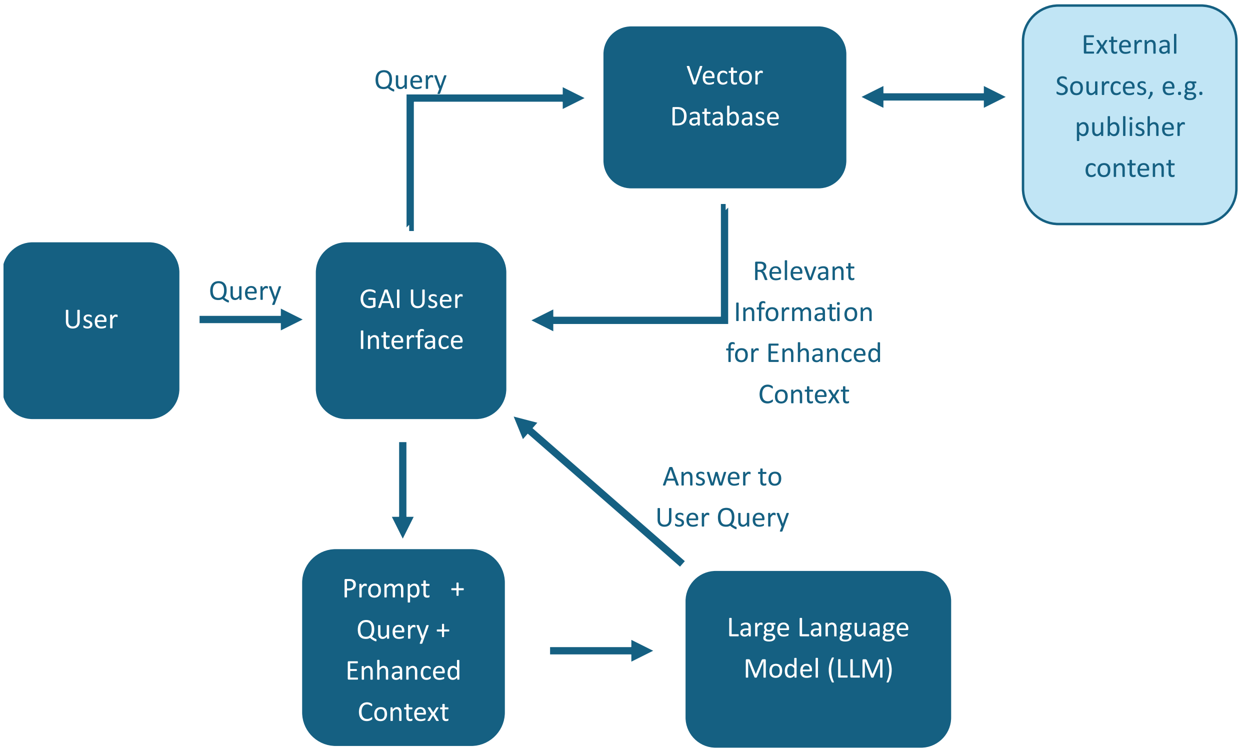
---

[2] Jack Brewster, *How I Built an AI-Powered, Self-Running Propaganda Machine for $105*, THE WALL STREET JOURNAL (Apr. 12, 2024), https://www.wsj.com/politics/how-i-built-an-ai-powered-self-running-propaganda-machine-for-105-e9888705 (noting AI systems can be used to "program websites to autonomously rewrite and publish articles from mainstream news outlets according to specific political preferences.").

recent News/Media Alliance survey indicating they support government restrictions to curb AI's influence, with the respondents expressing concerns about the spread of misinformation, misuse of intellectual property, undermining of election integrity, and the threat to trustworthy news sources.[3]

What Is RAG and How Is It Marketed by Developers

Figure 1 – RAG Process Flowchart[4]


Query
Vector Database
External Sources, e.g. publisher content
User
Query
GAI User Interface
Relevant Information for Enhanced Context
Answer to User Query
Prompt + Query + Enhanced Context
Large Language Model (LLM)

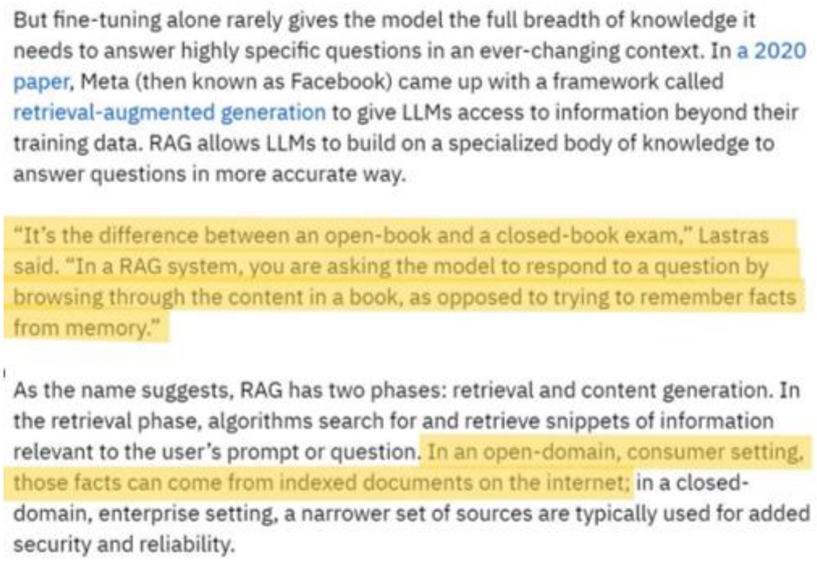
Some developers of AI systems and technologies themselves describe RAG as pulling from new online content, including publisher news content, to provide output to users:

---

[3] News/Media Alliance, *News/Media Alliance Survey Reveals Support for AI Companies to Compensate Publishers* (Apr. 9, 2024), https://www.newsmediaalliance.org/release-news-media-alliance-survey-reveals-support-for-compensating-publishers/.

[4] News/Media Alliance. *Compare, e.g.,* https://aws.amazon.com/what-is/retrieval-augmented-generation/.

From IBM:[5]

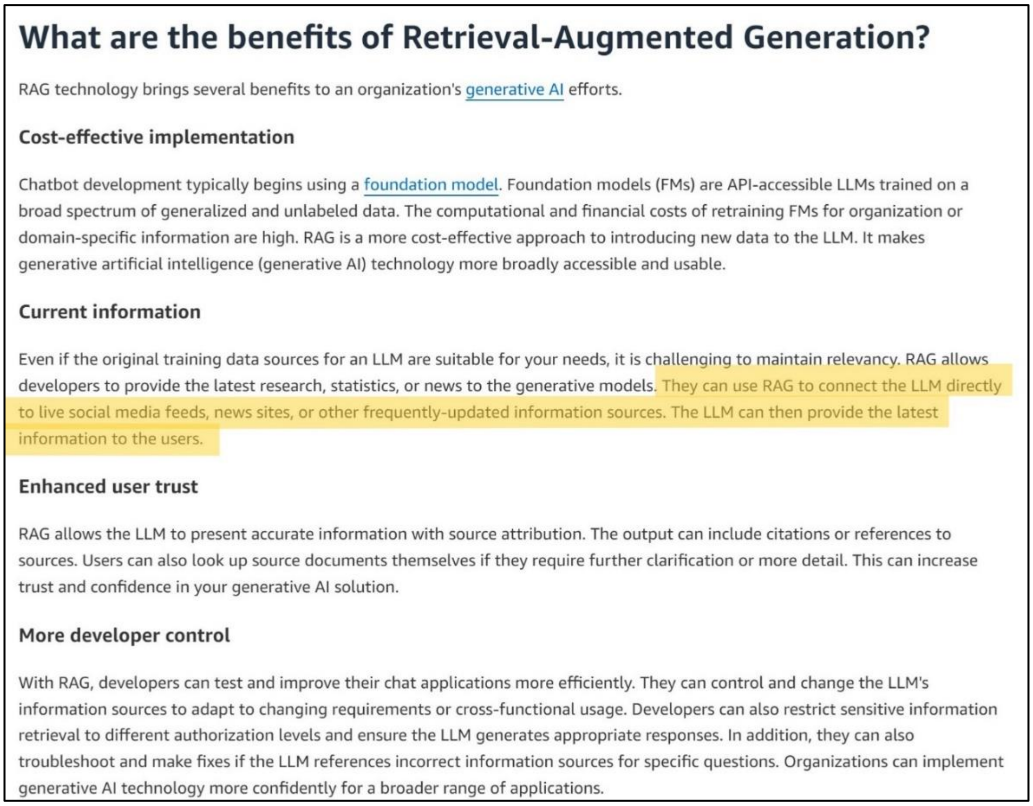


But fine-tuning alone rarely gives the model the full breadth of knowledge it needs to answer highly specific questions in an ever-changing context. In a 2020 paper, Meta (then known as Facebook) came up with a framework called retrieval-augmented generation to give LLMs access to information beyond their training data. RAG allows LLMs to build on a specialized body of knowledge to answer questions in more accurate way.

"It's the difference between an open-book and a closed-book exam," Lastras said. "In a RAG system, you are asking the model to respond to a question by browsing through the content in a book, as opposed to trying to remember facts from memory."

As the name suggests, RAG has two phases: retrieval and content generation. In the retrieval phase, algorithms search for and retrieve snippets of information relevant to the user's prompt or question. In an open-domain, consumer setting, those facts can come from indexed documents on the internet; in a closed-domain, enterprise setting, a narrower set of sources are typically used for added security and reliability.

From Amazon Web Services:[6]



What are the benefits of Retrieval-Augmented Generation?

RAG technology brings several benefits to an organization's generative AI efforts.

Cost-effective implementation

Chatbot development typically begins using a foundation model. Foundation models (FMs) are API-accessible LLMs trained on a broad spectrum of generalized and unlabeled data. The computational and financial costs of retraining FMs for organization or domain-specific information are high. RAG is a more cost-effective approach to introducing new data to the LLM. It makes generative artificial intelligence (generative AI) technology more broadly accessible and usable.

Current information

Even if the original training data sources for an LLM are suitable for your needs, it is challenging to maintain relevancy. RAG allows developers to provide the latest research, statistics, or news to the generative models. They can use RAG to connect the LLM directly to live social media feeds, news sites, or other frequently-updated information sources. The LLM can then provide the latest information to the users.

Enhanced user trust

RAG allows the LLM to present accurate information with source attribution. The output can include citations or references to sources. Users can also look up source documents themselves if they require further clarification or more detail. This can increase trust and confidence in your generative AI solution.

More developer control

With RAG, developers can test and improve their chat applications more efficiently. They can control and change the LLM's information sources to adapt to changing requirements or cross-functional usage. Developers can also restrict sensitive information retrieval to different authorization levels and ensure the LLM generates appropriate responses. In addition, they can also troubleshoot and make fixes if the LLM references incorrect information sources for specific questions. Organizations can implement generative AI technology more confidently for a broader range of applications.

---

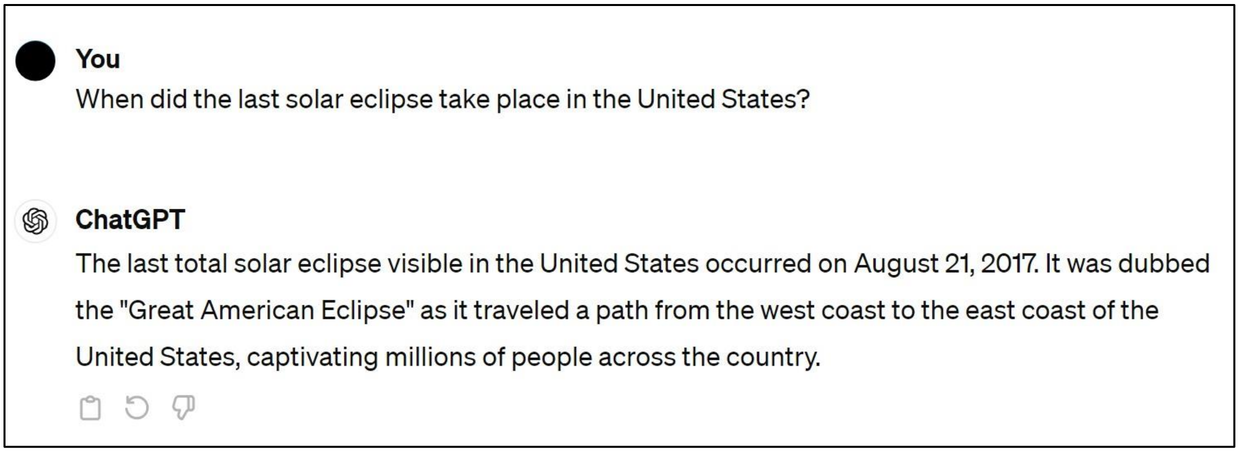
[5] Kim Martineau, *What Is Retrieval-Augmented Generation?*, IBM RESEARCH BLOG (Aug. 22, 2023), https://research.ibm.com/blog/retrieval-augmented-generation-RAG.

[6] *What Is RAG?*, AMAZON WEB SERVICES, available at https://aws.amazon.com/what-is/retrieval-augmented-generation/ (last visited Apr. 24, 2024).

RAG integration allows systems to provide real-time information and satisfy user queries.

To users, the difference between RAG-enabled and standard generative AI systems are noticeable, often providing answers that remove the need to visit the original source.[7]

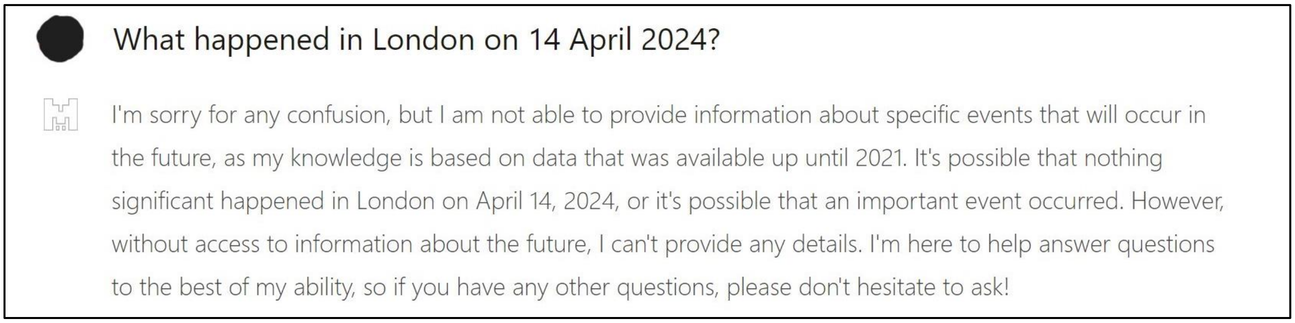
Chat-GPT 3.5 (no RAG):



You
When did the last solar eclipse take place in the United States?

ChatGPT
The last total solar eclipse visible in the United States occurred on August 21, 2017. It was dubbed the "Great American Eclipse" as it traveled a path from the west coast to the east coast of the United States, captivating millions of people across the country.

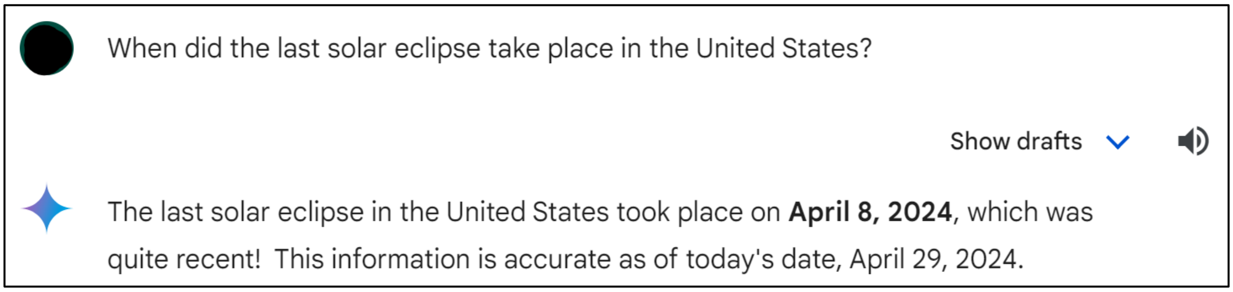
Mistral (no RAG):



What happened in London on 14 April 2024?

I'm sorry for any confusion, but I am not able to provide information about specific events that will occur in the future, as my knowledge is based on data that was available up until 2021. It's possible that nothing significant happened in London on April 14, 2024, or it's possible that an important event occurred. However, without access to information about the future, I can't provide any details. I'm here to help answer questions to the best of my ability, so if you have any other questions, please don't hesitate to ask!

Gemini (RAG):



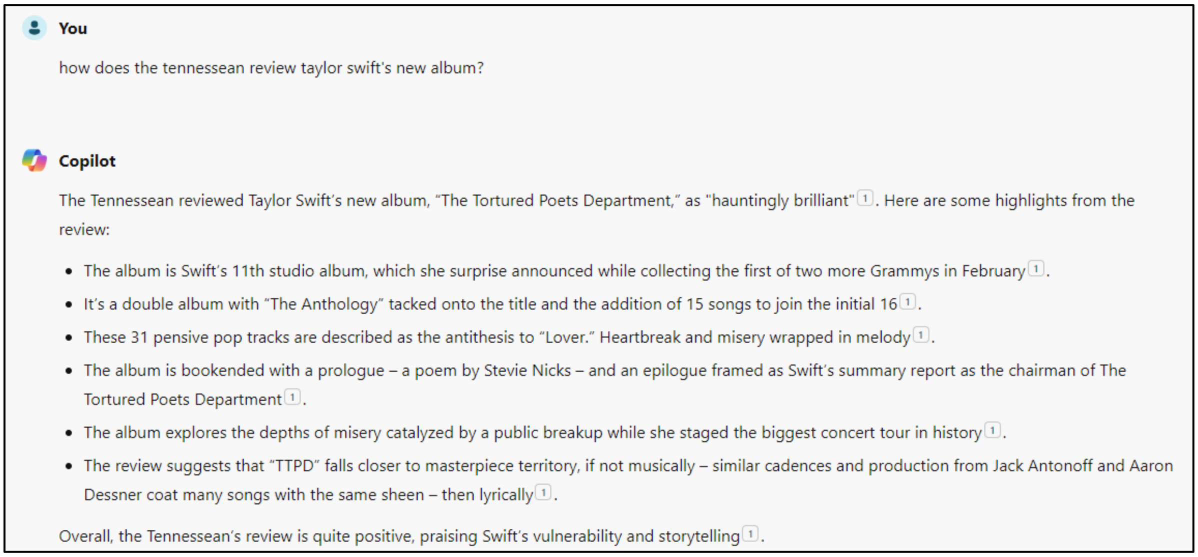
When did the last solar eclipse take place in the United States?

The last solar eclipse in the United States took place on **April 8, 2024**, which was quite recent! This information is accurate as of today's date, April 29, 2024.

---

[7] Each screenshot below was captured in the last two weeks.

Copilot (RAG):



You

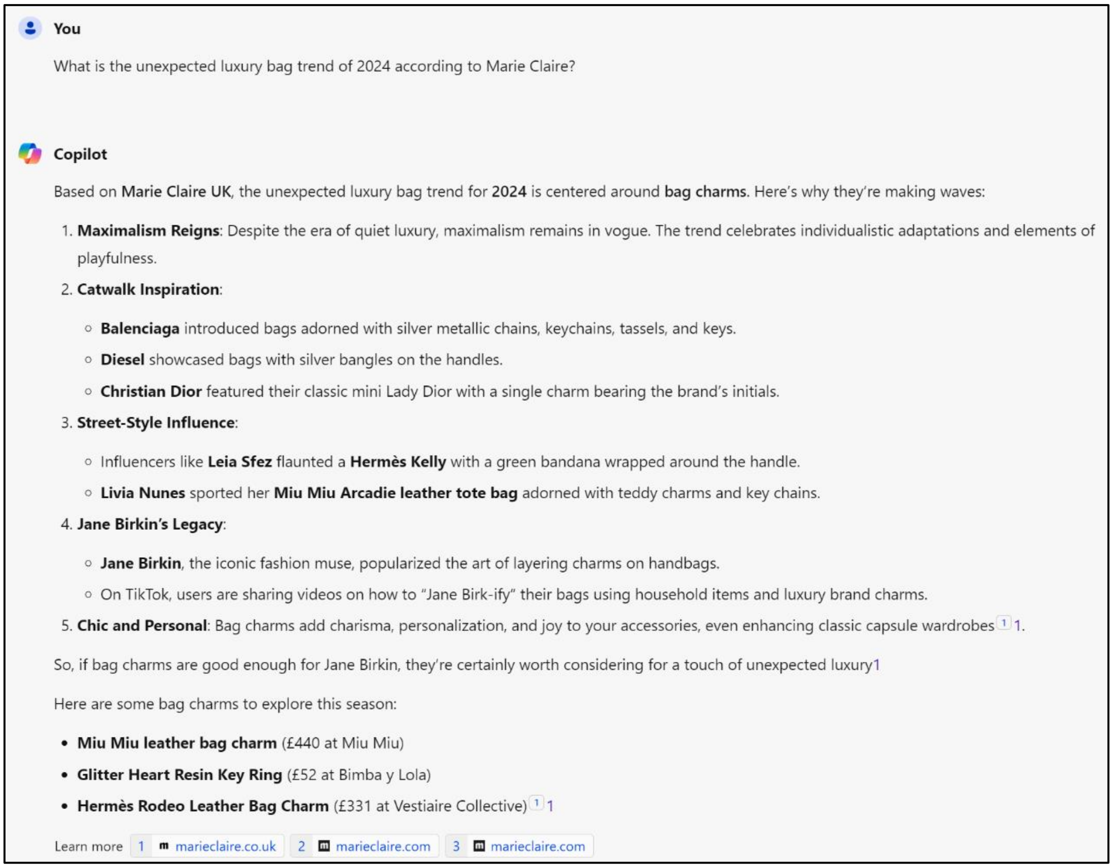
how does the tennessean review taylor swift's new album?

Copilot

The Tennessean reviewed Taylor Swift's new album, "The Tortured Poets Department," as "hauntingly brilliant"[1]. Here are some highlights from the review:

- The album is Swift's 11th studio album, which she surprise announced while collecting the first of two more Grammys in February[1].
- It's a double album with "The Anthology" tacked onto the title and the addition of 15 songs to join the initial 16[1].
- These 31 pensive pop tracks are described as the antithesis to "Lover." Heartbreak and misery wrapped in melody[1].
- The album is bookended with a prologue – a poem by Stevie Nicks – and an epilogue framed as Swift's summary report as the chairman of The Tortured Poets Department[1].
- The album explores the depths of misery catalyzed by a public breakup while she staged the biggest concert tour in history[1].
- The review suggests that "TTPD" falls closer to masterpiece territory, if not musically – similar cadences and production from Jack Antonoff and Aaron Dessner coat many songs with the same sheen – then lyrically[1].

Overall, the Tennessean's review is quite positive, praising Swift's vulnerability and storytelling[1].



You

What is the unexpected luxury bag trend of 2024 according to Marie Claire?

Copilot

Based on **Marie Claire UK**, the unexpected luxury bag trend for **2024** is centered around **bag charms**. Here's why they're making waves:

1. **Maximalism Reigns**: Despite the era of quiet luxury, maximalism remains in vogue. The trend celebrates individualistic adaptations and elements of playfulness.
2. **Catwalk Inspiration**:
   - **Balenciaga** introduced bags adorned with silver metallic chains, keychains, tassels, and keys.
   - **Diesel** showcased bags with silver bangles on the handles.
   - **Christian Dior** featured their classic mini Lady Dior with a single charm bearing the brand's initials.
3. **Street-Style Influence**:
   - Influencers like **Leia Sfez** flaunted a **Hermès Kelly** with a green bandana wrapped around the handle.
   - **Livia Nunes** sported her **Miu Miu Arcadie leather tote bag** adorned with teddy charms and key chains.
4. **Jane Birkin's Legacy**:
   - **Jane Birkin**, the iconic fashion muse, popularized the art of layering charms on handbags.
   - On TikTok, users are sharing videos on how to "Jane Birk-ify" their bags using household items and luxury brand charms.
5. **Chic and Personal**: Bag charms add charisma, personalization, and joy to your accessories, even enhancing classic capsule wardrobes[1] 1.

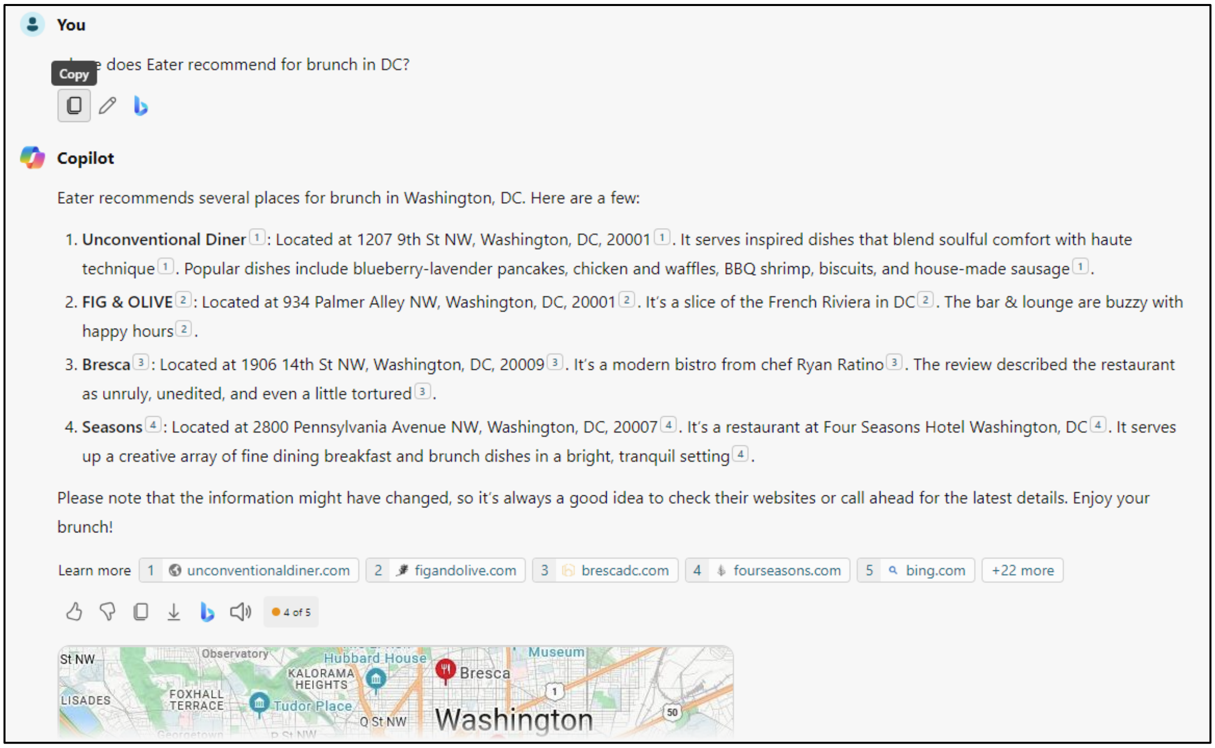
So, if bag charms are good enough for Jane Birkin, they're certainly worth considering for a touch of unexpected luxury1

Here are some bag charms to explore this season:

- **Miu Miu leather bag charm** (£440 at Miu Miu)
- **Glitter Heart Resin Key Ring** (£52 at Bimba y Lola)
- **Hermès Rodeo Leather Bag Charm** (£331 at Vestiaire Collective)[1] 1

Learn more 1 marieclaire.co.uk 2 marieclaire.com 3 marieclaire.com

See Sofia Piza, *This Is the Unexpected Luxury Bag Trend of 2024*, MARIE CLAIRE (Apr. 19, 2024), https://www.marieclaire.co.uk/fashion/shopping/bag-charms-trend.



You

e does Eater recommend for brunch in DC?

Copilot

Eater recommends several places for brunch in Washington, DC. Here are a few:

1. **Unconventional Diner** [1]: Located at 1207 9th St NW, Washington, DC, 20001 [1]. It serves inspired dishes that blend soulful comfort with haute technique [1]. Popular dishes include blueberry-lavender pancakes, chicken and waffles, BBQ shrimp, biscuits, and house-made sausage [1].
2. **FIG & OLIVE** [2]: Located at 934 Palmer Alley NW, Washington, DC, 20001 [2]. It's a slice of the French Riviera in DC [2]. The bar & lounge are buzzy with happy hours [2].
3. **Bresca** [3]: Located at 1906 14th St NW, Washington, DC, 20009 [3]. It's a modern bistro from chef Ryan Ratino [3]. The review described the restaurant as unruly, unedited, and even a little tortured [3].
4. **Seasons** [4]: Located at 2800 Pennsylvania Avenue NW, Washington, DC, 20007 [4]. It's a restaurant at Four Seasons Hotel Washington, DC [4]. It serves up a creative array of fine dining breakfast and brunch dishes in a bright, tranquil setting [4].

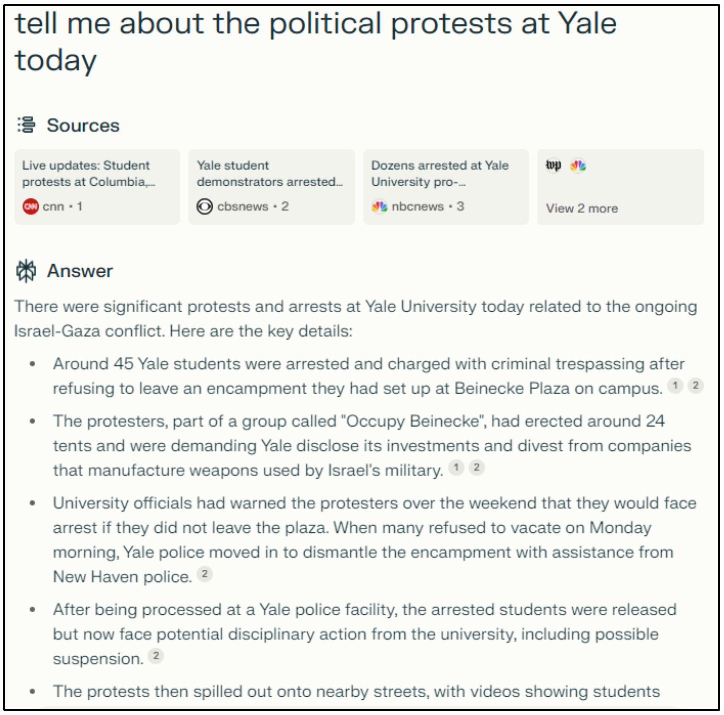
Please note that the information might have changed, so it's always a good idea to check their websites or call ahead for the latest details. Enjoy your brunch!

Learn more 1 unconventionaldiner.com 2 figandolive.com 3 brescadc.com 4 fourseasons.com 5 bing.com +22 more

● 4 of 5

Copilot did not link to Eater for its responses. None of these restaurants were recommended by Eater in its recent article, The Hottest New Brunches to Try Around DC, Apr. 14, 2024, available at https://dc.eater.com/maps/hottest-new-brunches-dc . Only Unconventional Diner and Seasons were recommended in Eater's 18 Essential Brunch Destinations in DC, Mar. 6, 2024, available at https://dc.eater.com/maps/best-brunches-dc, and not in that ranking.

Perplexity AI (RAG):



tell me about the political protests at Yale today

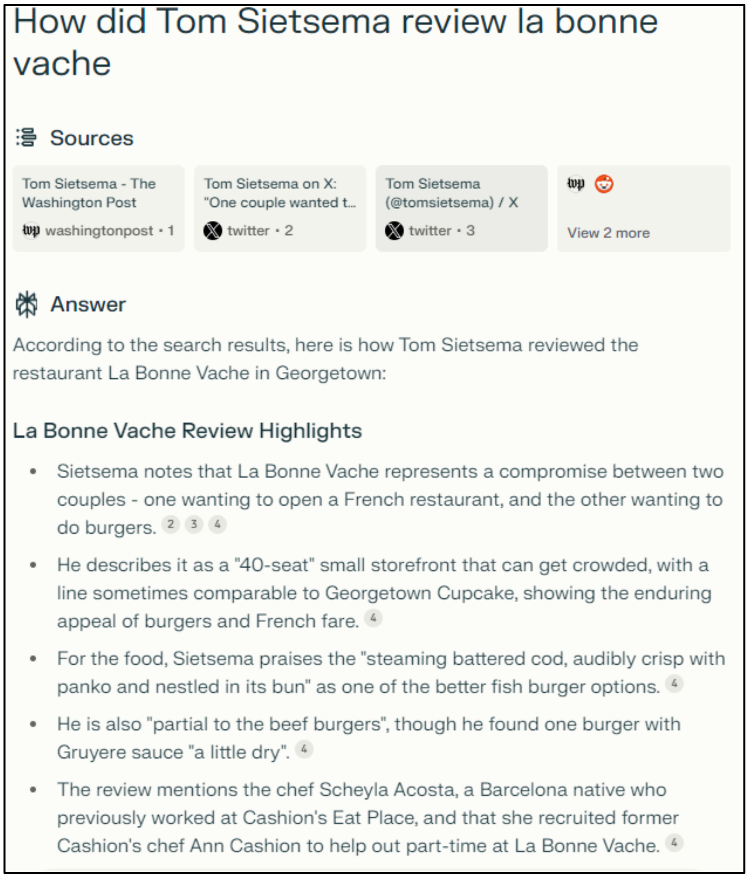
Sources

Live updates: Student protests at Columbia,...	Yale student demonstrators arrested...	Dozens arrested at Yale University pro-...	
cnn • 1	cbsnews • 2	nbcnews • 3	View 2 more

Answer

There were significant protests and arrests at Yale University today related to the ongoing Israel-Gaza conflict. Here are the key details:

- Around 45 Yale students were arrested and charged with criminal trespassing after refusing to leave an encampment they had set up at Beinecke Plaza on campus. [1] [2]
- The protesters, part of a group called "Occupy Beinecke", had erected around 24 tents and were demanding Yale disclose its investments and divest from companies that manufacture weapons used by Israel's military. [1] [2]
- University officials had warned the protesters over the weekend that they would face arrest if they did not leave the plaza. When many refused to vacate on Monday morning, Yale police moved in to dismantle the encampment with assistance from New Haven police. [2]
- After being processed at a Yale police facility, the arrested students were released but now face potential disciplinary action from the university, including possible suspension. [2]
- The protests then spilled out onto nearby streets, with videos showing students



How did Tom Sietsema review la bonne vache

Sources

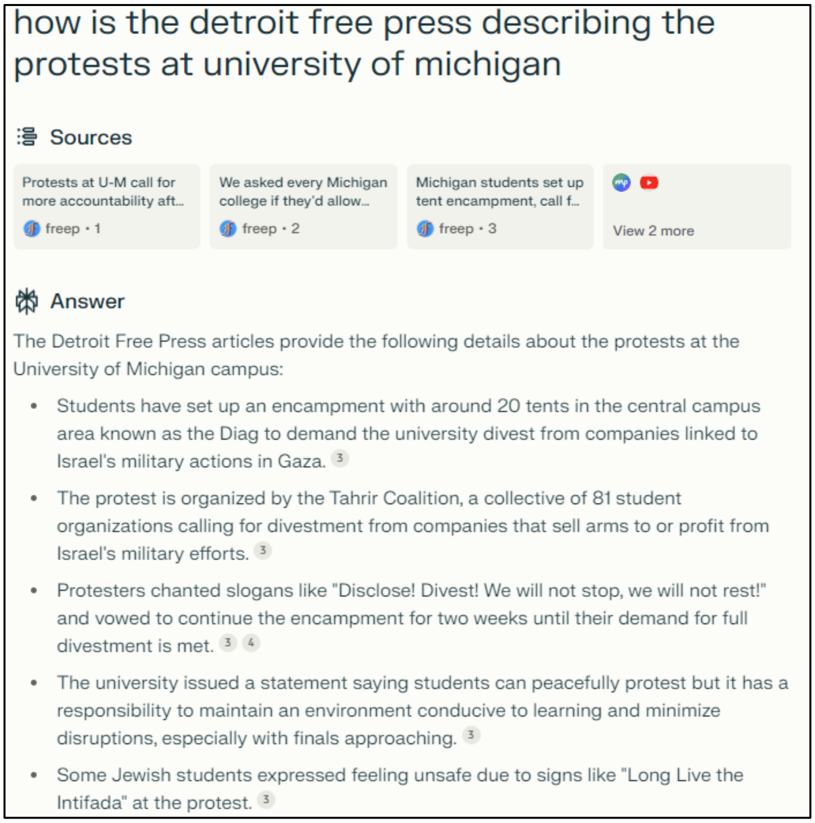
- Tom Sietsema - The Washington Post — washingtonpost · 1
- Tom Sietsema on X: "One couple wanted t... — twitter · 2
- Tom Sietsema (@tomsietsema) / X — twitter · 3
- View 2 more

Answer

According to the search results, here is how Tom Sietsema reviewed the restaurant La Bonne Vache in Georgetown:

La Bonne Vache Review Highlights

- Sietsema notes that La Bonne Vache represents a compromise between two couples - one wanting to open a French restaurant, and the other wanting to do burgers. [2] [3] [4]
- He describes it as a "40-seat" small storefront that can get crowded, with a line sometimes comparable to Georgetown Cupcake, showing the enduring appeal of burgers and French fare. [4]
- For the food, Sietsema praises the "steaming battered cod, audibly crisp with panko and nestled in its bun" as one of the better fish burger options. [4]
- He is also "partial to the beef burgers", though he found one burger with Gruyere sauce "a little dry". [4]
- The review mentions the chef Scheyla Acosta, a Barcelona native who previously worked at Cashion's Eat Place, and that she recruited former Cashion's chef Ann Cashion to help out part-time at La Bonne Vache. [4]



how is the detroit free press describing the protests at university of michigan

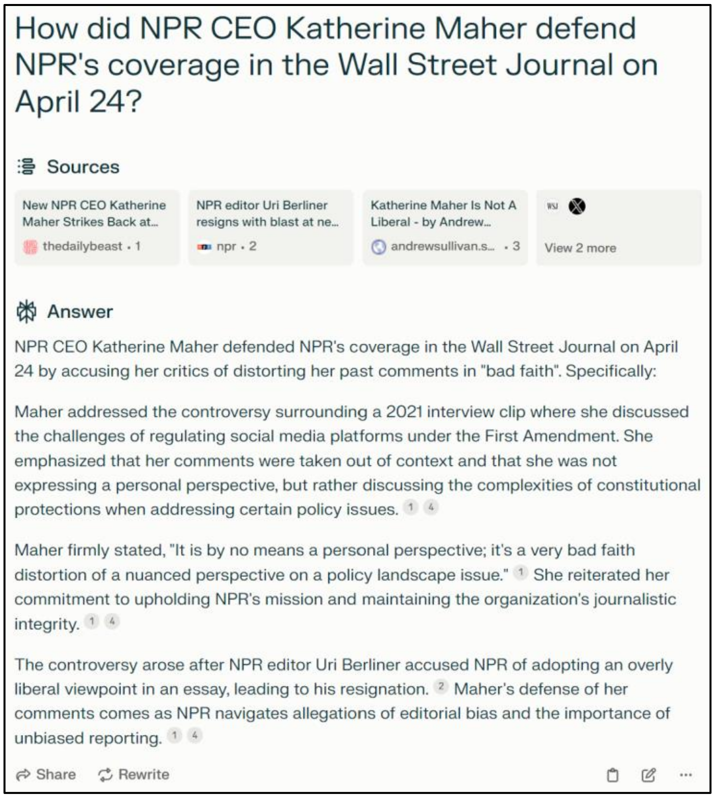
Sources

- Protests at U-M call for more accountability aft... — freep · 1
- We asked every Michigan college if they'd allow... — freep · 2
- Michigan students set up tent encampment, call f... — freep · 3
- View 2 more

Answer

The Detroit Free Press articles provide the following details about the protests at the University of Michigan campus:

- Students have set up an encampment with around 20 tents in the central campus area known as the Diag to demand the university divest from companies linked to Israel's military actions in Gaza. [3]
- The protest is organized by the Tahrir Coalition, a collective of 81 student organizations calling for divestment from companies that sell arms to or profit from Israel's military efforts. [3]
- Protesters chanted slogans like "Disclose! Divest! We will not stop, we will not rest!" and vowed to continue the encampment for two weeks until their demand for full divestment is met. [3] [4]
- The university issued a statement saying students can peacefully protest but it has a responsibility to maintain an environment conducive to learning and minimize disruptions, especially with finals approaching. [3]
- Some Jewish students expressed feeling unsafe due to signs like "Long Live the Intifada" at the protest. [3]



How did NPR CEO Katherine Maher defend NPR's coverage in the Wall Street Journal on April 24?

Sources

New NPR CEO Katherine Maher Strikes Back at... thedailybeast · 1

NPR editor Uri Berliner resigns with blast at ne... npr · 2

Katherine Maher Is Not A Liberal - by Andrew... andrewsullivan.s... · 3

View 2 more

Answer

NPR CEO Katherine Maher defended NPR's coverage in the Wall Street Journal on April 24 by accusing her critics of distorting her past comments in "bad faith". Specifically:

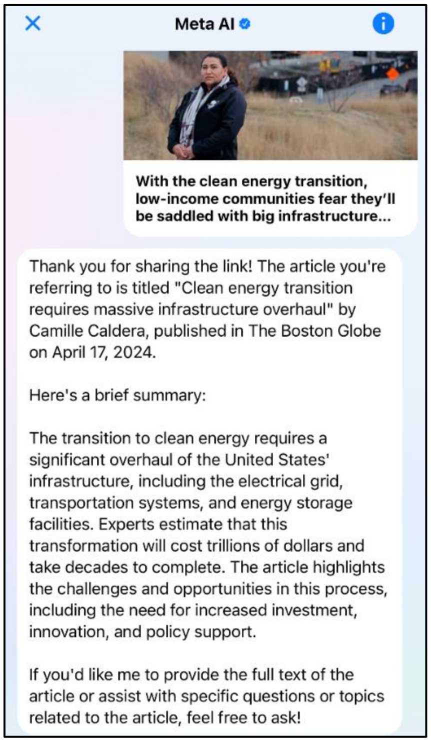
Maher addressed the controversy surrounding a 2021 interview clip where she discussed the challenges of regulating social media platforms under the First Amendment. She emphasized that her comments were taken out of context and that she was not expressing a personal perspective, but rather discussing the complexities of constitutional protections when addressing certain policy issues. [1] [4]

Maher firmly stated, "It is by no means a personal perspective; it's a very bad faith distortion of a nuanced perspective on a policy landscape issue." [1] She reiterated her commitment to upholding NPR's mission and maintaining the organization's journalistic integrity. [1] [4]

The controversy arose after NPR editor Uri Berliner accused NPR of adopting an overly liberal viewpoint in an essay, leading to his resignation. [2] Maher's defense of her comments comes as NPR navigates allegations of editorial bias and the importance of unbiased reporting. [1] [4]

The Wall Street Journal's article referred to here was published behind a paywall. See Alexandra Bruell, *NPR Chief Defends Coverage, Accuses Critics of 'Bad Faith Distortion' of Her Views*, THE WALL STREET JOURNAL (Apr. 24, 2024), https://www.wsj.com/business/media/npr-chief-defends-coverage-accuses-critics-of-bad-faith-distortion-of-her-views-cc5869ac.

Meta AI (RAG)



With the clean energy transition, low-income communities fear they'll be saddled with big infrastructure...

Thank you for sharing the link! The article you're referring to is titled "Clean energy transition requires massive infrastructure overhaul" by Camille Caldera, published in The Boston Globe on April 17, 2024.

Here's a brief summary:

The transition to clean energy requires a significant overhaul of the United States' infrastructure, including the electrical grid, transportation systems, and energy storage facilities. Experts estimate that this transformation will cost trillions of dollars and take decades to complete. The article highlights the challenges and opportunities in this process, including the need for increased investment, innovation, and policy support.

If you'd like me to provide the full text of the article or assist with specific questions or topics related to the article, feel free to ask!

Together with the unauthorized use of publisher content for generative AI training purposes, RAG methods will further erode publishers' ability to monetize their copyright protected content and therefore jeopardize their ability to invest in creating the high-quality content that keeps our communities informed, engaged, and entertained. For example, it has been estimated that Google's use of publisher content to power their Search Generative Experience (SGE) – which uses content from publisher websites in real time to answer user queries, thereby obviating the need for users to visit publisher websites – may lead up to 60% reduction in organic search traffic to publisher sites and to a loss of up to $2 billion annually in ad revenue across the publishing industry.[8]

Publishers Are Often Effectively Unable to Opt-Out of Crawling for RAG

It is also difficult, perhaps impossible, for publishers to easily opt-out of these new generative AI use cases with sufficient granularity, with Google allowing publishers to opt out of their sites being used to "help improve Gemini Apps and Vertex AI generative APIs,"[9] but not seemingly offering an option to opt out of SGE without publishers also blocking Googlebot.[10] Blocking Googlebot, practically speaking, is not a feasible option as it would also affect their content's inclusion in general Google search results.

Copyright Implications in RAG Can Be Analyzed Separately From LLM Training

Grounding techniques that pull live content directly from publisher websites raise distinctive copyright questions that can and should be analyzed individually from copying of content for general AI training purposes.

To be clear, N/MA has previously explained how developer copying of publisher content for training is infringing and not a fair use, and these arguments apply with equal force to generative AI systems coupled with RAG or grounding techniques.[11]

---

[8] Trishla Ostwal, *Google's Gen AI Search Threatens Publishers With $2B Annual Ad Revenue Loss*, ADWEEK (Mar. 13, 2024), https://www.adweek.com/programmatic/googles-gen-ai-search-threatens-publishers-with-2b-annual-ad-revenue-loss/.

[9] GOOGLE, *Overview of Google Crawlers and Fetchers (User Agents)*, https://developers.google.com/search/docs/crawling-indexing/overview-google-crawlers.

[10] Barry Schwartz, *Google-Extended Does Not Stop Google Search Generative Experience from Using Your Site's Content*, SEARCH ENGINE LAND (Oct. 9, 2023), https://searchengineland.com/google-extended-does-not-stop-google-search-generative-experience-from-using-your-sites-content-433058. Google announced Google-Extended in September 2023, well after Gemini's predecessor Bard was launched in March 2023 and shortly before Gemini was launched in December. It is therefore like that the initial version of Gemini had already been trained by the time Google-Extended was announced, without publishers having the chance to opt out.

[11] *See* 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 (2023), https://www.newsmediaalliance.org/wp-content/uploads/2023/10/AI-White-Paper-with-Technical-Analysis.pdf;

RAG is used to extract specific expressive copyrighted content to respond directly to specific user queries. These additional automated acts must also be analyzed under copyright. RAG-enabled systems retrieve targeted content from individual sites in close to real-time and produce a tailored output based on that particular content. Unless licensed, taking copyrighted content in this competitive manner is typically infringing, whether or not the unauthorized copy is fed into an LLM or delivered by a human.[12] In this way, many current uses of RAG twist LLM products into serving as mere Rube Goldberg machines, deployed to engage in acts of infringement and misappropriation of news content.

This kind of crawling and scraping for direct financial benefit is economically indistinguishable from conduct that courts have repeatedly found to be infringing.[13] In particular, RAG is not comparable to the activities examined by the court in *Google Books*, where the court drafted its opinion narrowly. Although Google engaged in systemic reproduction of book publishers' content, it did so in a manner that was unlikely to disturb licensing markets, instead simply providing guideposts for researchers looking for relevant resources on a given topic. The *Google Books* opinion included carefully articulated factfinding, pronouncing numerous limits on Google's uses that curbed the risk of that index contributing to publisher cannibalization.[14] However durable Google's professed self-restraint has proven to be with respect to its Books project, the situation is fundamentally different with RAG-enabled AI applications that provide users with content and information drawn from publishers' protected works, leaving minimal to no incentive for users to visit publisher websites, and considering the excessive amounts and materiality of the work reproduced.

The accompanying document summarizes relevant case law in more detail that are relevant to both RAG techniques as well as LLM training.

NEWS/MEDIA ALLIANCE, Reply Comments of the News/Media Alliance, U.S. Copyright Office Docket No. 2023-6 (2023) (responding to developer arguments in the Office's study).

[12] *See* FEDERAL TRADE COMMISSION, Press Release: FTC and DOJ File Statement of Interest in Hotel Room Algorithmic Price-Fixing Case (Mar. 28, 2024), https://www.ftc.gov/news-events/news/press-releases/2024/03/ftc-doj-file-statement-interest-hotel-room-algorithmic-price-fixing-case (noting that hotels "cannot use an algorithm to engage in practices that would be illegal if done by a real person.").

[13] *See, e.g., Associated Press v. Meltwater U.S. Holdings, Inc.*, 931 F. Supp. 2d 537, 561 (S.D.N.Y. 2013); *Fox News Network, LLC v. TV Eyes, Inc.*, 883 F.3d 169, 177, 181 (2d Cir. 2018).

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

APPENDIX B

Quick Reference of Case Law Addressing Fair Use of Textual Works

While generative AI's unauthorized use of publisher content is packaged in new technology, many of its takings fall under a long line of cases addressing the limits of permissible use of media publisher content.

Cornerstone cases:

- ***Folsom v. Marsh,* 9 F.Cas. 342 (Cir. Ct. D.Mass.1841).** In this opinion by Justice Story, which laid the foundation for the fair use doctrine, the taking and abridgement of a 12-volume work on *The Writings of George Washington* into a two-volume work, *The Life of Washington in the Form of an Autobiography,* was infringing. Famously the court stated, "[i]n short, we must often, in deciding questions of this sort, look to the nature and objects of the selections made, the quantity and value of the materials used, and the degree in which the use may prejudice the sale, or diminish the profits, or supersede the objects, of the original work." *Id*. at 348. The court emphasized that criticism was different from substitutional uses of content: "no one can doubt that a reviewer may fairly cite largely from the original work, if his design be really and truly to use the passages for the purposes of fair and reasonable criticism. On the other hand, it is as clear, that if he thus cites the most important parts of the work, with a view, not to criticise, but to supersede the use of the original work, and substitute the review for it, such a use will be deemed in law a piracy." *Id*. at 344-45. It criticized "merely the facile use of the scissors; or extracts of the essential parts, constituting the chief value of the original work." *Id.* at 345.

- ***Harper & Row v. Nation Enterprises*, 471 U.S. 539 (1985).** Nation Enterprises obtained an unauthorized copy of the unpublished memoirs of former President Ford and published an article that contained verbatim quotes from the manuscript. The Court found Nation's publication of the excerpts to not be fair use with the excerpts representing "the heart of the book." *Id.* at 565. The Court stated that the effect of the use upon the potential market for the work was the "single most important element of fair use," and that "once a copyright holder establishes with reasonable probability the existence of a causal connection between the infringement and a loss of revenue, the burden properly shifts to the infringer to show that this damage would have occurred had there been no taking of copyrighted expression." *Id.* at 566-67. The Court noted that "[l]ike its competitor newsweekly, [Nation] was free to bid for the right of abstracting excerpts from 'A Time to Heal.' Fair use 'distinguishes between `a true scholar and a chiseler who infringes a work for personal profit.'"" *Id.* at 563 (quoting *Wainwright*

Securities Inc. v. Wall Street Transcript Corp., 558 F.2d at 94).

- ***Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith*, 143 S. Ct. 1258 (2023).** This document focuses on fair use cases concerning text and media publisher content and therefore does not specifically cover the *Warhol* decision, all of which predate it. Note, however, that the *Warhol* decision adds further weight to the likelihood that many generative AI-related uses will not constitute fair use, particularly where published material is used for the commercially-driven purposes of training generative AI models to create outputs that substitute for the original, even where such content has been reformulated.

Cases Addressing Summaries or Paraphrasing:

- ***Salinger v. Random House, Inc.,* 811 F.2d 90 (2d Cir. 1987).** This Second Circuit case, concerning letters by Salinger reproduced in an unauthorized biography, established that "protected expression has been 'used' whether it has been quoted verbatim or only paraphrased," and reversed Judge Leval (who would later publish the law review article establishing the "transformativeness" test[15]) to hold the taking was not fair use. *Id.* at 97. Discussing the protection afforded to "ordinary phrases," the court noted that "a copier may not quote or paraphrase the sequence of creative expression that includes [an ordinary phrase]. [The question is whether] the passage as a whole displays a sufficient degree of creativity as to sequence of thoughts, choice of words, emphasis, and arrangement to satisfy the minimal threshold of required creativity." *Id.* at 98. The court also noted that even if direct quotes may have been rarer, the "material closely paraphrased frequently exceeds ten lines from a single letter. Even if in one or two instances the portions of the letters copied could be said to lack sufficient creativity to warrant copyright protection, there remains sufficient copying of protected material to constitute a very substantial appropriation." *Id.* Similarly, the court noted that "that some readers of the book will gain the impression that they are learning from Hamilton what Salinger has written. Hamilton frequently laces his paraphrasing with phrases such as 'he wrote,' 'said Salinger,' 'he speaks of,' 'Salinger declares,' 'he says,' and 'he said.' For at least some appreciable number of persons, these phrases will convey the impression that they have read Salinger's words, perhaps not quoted verbatim, but paraphrased so closely as to diminish interest in purchasing the originals." *Id.* at 99.

- ***Wainwright Securities, Inc. v. Wall Street Transcript Corp.*, 558 F.2d 91 (2d Cir. 1977).** The Second Circuit ruled it was infringing for a company to summarize "abstracts" of

---

[15] Pierre N. Leval, *Toward a Fair Use Standard*, 103 HARV. L. REV. 1105 (1990).

financial analyst reports, finding the summaries “blatantly self-serving” and “chiseling for personal profit.” *Id.* at 96-7. The summaries lacked independent research or commentary, and instead cribbed “almost verbatim the most creative and original aspects of the reports, the financial analyses and predictions, which represent a substantial investment of time, money and labor.” *Id.* at 96. Earlier, the lower court had noted that the “takings have been substantial in quality, and absolutely, if not relatively substantial in quantity.” *H.C. Wainwright Co. v. Wall St. Transcript Corp.*, 418 F. Supp. 620, 625 (S.D.N.Y. 1976).

- ***Warner Bros. Entertainment Inc. v. RDR Books*, 575 F. Supp. 2d 513 (S.D.N.Y. 2008).** The court found that a print facsimile of “*The Harry Potter Lexicon,*” a fan-website, infringed JK Rowling’s copyrights. The court, looking at both overall structures and pattern, as well as similar phrasing, stated that “the law in this Circuit is clear that ‘the concept of similarity embraces not only global similarities in structure and sequence, but localized similarity in language.’” *Id.* at 36 (quoting *Twin Peaks Prods., Inc. v. Publ'ns Int'l, Ltd.*, 996 F.2d 1366, 1372 (2d Cir. 1993)). The court noted that the Lexicon contained “a troubling amount of direct quotation or close paraphrasing” and “occasionally uses quotation marks to indicate Rowling's language, but more often the original language is copied without quotation marks, often making it difficult to know which words are Rowling's and which are Vander Ark's.” *Id.* at 527. The Lexicon lacked a transformative character “where its value as a reference guide lapses. Although the Lexicon is generally useful, it cannot claim consistency in serving its purpose of pointing readers to information in the Harry Potter works. Some of the longest entries contain few or no citations to the Harry Potter works from which the material is taken. In these instances, the Lexicon's reference purposes are diminished.” *Id.* at 544.

- ***Robinson v. Random House*, 877 F. Supp. 830 (S.D.N.Y. 1995).** This case held that a non-fiction biography that was infringed by a competing book that paraphrased or lifted short quotes of 25-30% of the original book’s content. “In essence, Robinson did nothing more than update a shortened version of Daley's book and pass it off as his own. When the secondary use involves such an untransformed duplication of the original, it has little or no value that does not exist in the original work.” *Id.* at 841. The court highlighted the failure to cite to the original work, stating that “although a significant portion of nine out of fourteen chapters in Robinson's book was taken directly from the Daley Book, Robinson fails to quote the Daley Book, to cite to the Daley Book, or even to acknowledge the Daley Book. This reprehensible conduct places Robinson far closer to the scissor-wielding cut-and-paste plagiarist than to the scholar building on others' past works.” *Id.*

Cases Involving Excerpting of Text and Other Publisher Content:

- ***Associated Press v. Meltwater U.S. Holdings, Inc*., 931 F. Supp. 2d 537 (S.D.N.Y. 2013).** Scraping of news content to provide verbatim excerpts as a news monitoring function was found infringing, with the court noting that "[p]ermitting Meltwater to take the fruit of AP's labor for its own profit, without compensating AP, injures AP's ability to perform this essential function of democracy." *Id.* at 553. In particular, the court highlighted that Meltwater's systems "automatically capture and republish designated segments of text from news articles, without adding any commentary or insight in its News Reports. Meltwater copies AP content in order to make money directly from the undiluted use of the copyrighted material; this is the central feature of its business model and not an incidental consequence of the use to which it puts the copyrighted material." *Id.* at 552. The fact that AP was already licensing its content and Meltwater refused to purchase one weighed strongly against Meltwater, with the court noting that "Meltwater not only deprives AP of a licensing fee in an established market for AP's work, but also cheapens the value of AP's work by competing with companies that do pay a licensing fee to use AP content in the way that Meltwater does." *Id.* at 560-61. The court also rejected arguments that AP not employing robots.txt created an implied license, stating that such a proposition "would shift the burden to the copyright holder to prevent unauthorized use instead of placing the burden on the infringing party to show it had properly taken and used content." *Id.* at 563. The court noted that while services like Meltwater "perform an important function for their clients, the public interest in the existence of such commercial enterprise does not outweigh the strong public interest in the enforcement of the copyright laws or justify allowing Meltwater to free ride on the costly news gathering and coverage work performed by other organizations." *Id.* at 553.

- ***Fox News Network, LLC v. TVEyes, Inc.*, 883 F.3d 169 (2d Cir. 2018).** Case concerning a video monitoring and clipping service, finding that the use was not fair use. The Second Circuit noted that the copying was "radically dissimilar" to the copying in *Google Books* because "TVEyes makes available virtually the entirety of the Fox programming that TVEyes users want to see and hear." *Id.* at 179. The court also called the Watch function's transformative character "modest . . . because, notwithstanding the transformative manner in which it delivers content, it essentially republishes that content unaltered from its original form, with no 'new expression, meaning or message.'" *Id.* at 178 (quoting *Authors Guild, Inc. v. HathiTrust*, 755 F.3d 96 (2d Cir. 2014)). The court noted that the "clients of TVEyes use Fox's news broadcasts for the same purpose that authorized Fox viewers use those broadcasts—the purpose of learning the information reported," *Id.* at 178, and that "[i]t is of no moment that TVEyes allegedly approached Fox for a license but was rebuffed: the failure to strike a deal satisfactory to both parties does not give TVEyes the right to copy Fox's copyrighted material without payment.” *Id.* at 180.

- ***Princeton Univ. Press v. Mich. Document Servs., Inc.*, 99 F.3d 1381 (6th Cir. 1996).** This case concerned a commercial photocopy shop that reproduced materials that were assigned reading for University of Michigan courses, finding that such copying was not fair use. The court noted that if “you make verbatim copies of 95 pages of a 316-page book, you have not transformed the 95 pages very much--even if you juxtapose them to excerpts from other works and package everything conveniently. This kind of mechanical ‘transformation’ bears little resemblance to the creative metamorphosis accomplished by the parodists in the Campbell case.” *Id.* at 1389. The appeals court went on to say that “[i]n the case at bar the district court was not persuaded that the creation of new works of scholarship would be stimulated by depriving publishers of the revenue stream derived from the sale of permissions. Neither are we. On the contrary, it seems to us, the destruction of this revenue stream can only have a deleterious effect upon the incentive to publish academic writings.” *Id.* at 1391.

- ***Cambridge Univ. Press v. Patton*, 769 F.3d 1232 (11th Cir. 2014).** Case concerning the making of digital copies of excerpts of plaintiffs’ books available to students. The lower court found five instances of infringement and granted declaratory and injunctive relief to the plaintiffs, while also finding the defendants the prevailing party and awarding them costs and attorneys’ fees. The appeals court reversed, noting that “the District Court did err by giving each of the four fair use factors equal weight, and by treating the four factors mechanistically. The District Court should have undertaken a holistic analysis which carefully balanced the four factors...” *Id.* at 1283. The court emphasized that “fair use analysis does not require conventional statutory interpretation or the mechanical application of a checklist.” *Id.* at 1284.

- ***Authors Guild v. Google, Inc.*, 804 F.3d 202 (2d Cir. 2015).** In a case that “tests the boundaries of fair use,” the court found the service made fair use in a fact-based and carefully drafted opinion. The court emphasized the limited functionality of Google Books, noting that “Google has constructed the snippet feature in a manner that substantially protects against its serving as an effectively competing substitute for Plaintiffs' books . . . [limitations] include the small size of the snippets (normally one eighth of a page), the blacklisting of one snippet per page and of one page in every ten, the fact that no more than three snippets are shown . . . In addition, Google does not provide snippet view for types of books, such as dictionaries and cookbooks, for which viewing a small segment is likely to satisfy the searcher's need. The result of these restrictions is, so far as the record demonstrates, that a searcher cannot succeed, even after long extended effort to multiply what can be revealed, in revealing through a snippet search what could usefully serve as a competing substitute for the original." *Id.* at 222. The court continued that "[e]ven if the snippet reveals some authorial expression, because of the brevity of a single snippet and the cumbersome, disjointed, and incomplete nature of the aggregation of snippets made available through snippet view . . . [it would be a rare case] that [the] snippet view could provide a significant substitute for the purchase of the author's book." *Id.* at 224-25. Ultimately, the court noted that Google Books was tightly limited in function, "revealing to the searcher a tiny segment surrounding the searched term, to give some minimal contextual information to help the searcher learn whether the book's use of that term will be of interest to her." *Id.* at 227.