

Advertising in Generative Information Retrieval

Ines Zelch^{1,2}

Matthias Hagen¹

Martin Potthast^{2,3}

¹Friedrich-Schiller-Universität Jena ²Leipzig University ³ScaDS.AI

Workshop on the Impact of Generative AI on Search and Search Engine Research

Hamburg, 07.09.2023

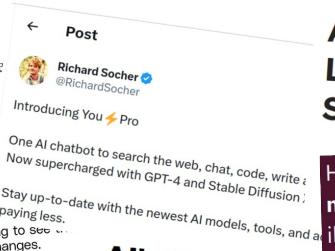
webis.de

Generative Information Retrieval



Inside The High-Stakes, AI-Powered Race To Dethrone Google Search

f Pranav Dixit / Buzzfeed News
5 May 2023 - 18-min read



Announcing AI2 OLMo, an Open Language Model Made by Scientists, for Scientists

Have you noticed that news stories and marketing material about Artificial Intelligence are typically illustrated with clichéd and misleading images?

AI's 'Godfather' Should Have Spoken Up Sooner

Hopefully, Google scientist Geoffrey Hinton's warnings about the technology's potential harms will persuade other researchers to come forward.

World's first artificial intelligence law moves closer to passage in European Parliament

11 May 2023 | News

To the alarm of mostly-US industry, the EU AI Act gets preliminary nod to control the emerging technology – and sets the stage for months of negotiations in Brussels and Washington

By Richard L. Hudson

OpenAI, IBM Urge Senate to Act on AI Regulation After Past Tech Failures

- Altman tells Congress the US should license advanced models
- IBM's Montgomery says regulation must focus on risks of AI use

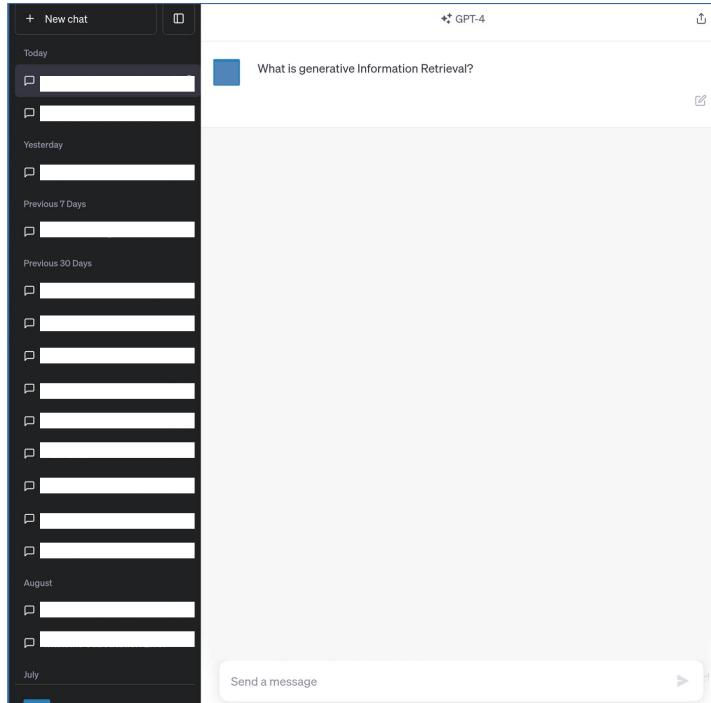
Pause Giant AI Experiments: An Open Letter

We call on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4.

- Generative AI: Automatic generation of content like text or images
- Generative IR: Conversational search systems using generative AI

Generative Information Retrieval

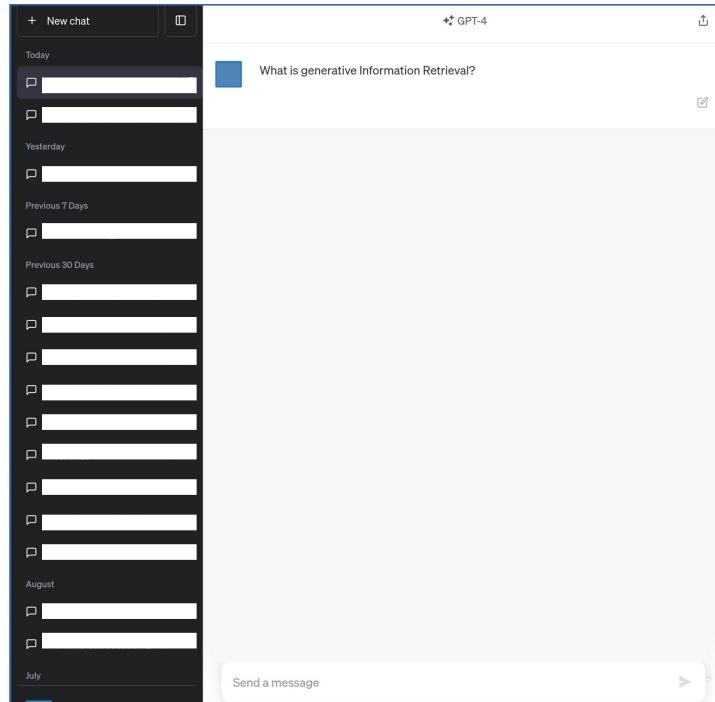
ChatGPT



Generative Information Retrieval

ChatGPT

What is generative Information Retrieval?



Generative Information Retrieval

ChatGPT

What is generative Information Retrieval?

The screenshot shows the ChatGPT interface. On the left, there's a sidebar with a 'New chat' button and a list of previous conversations grouped by date: Today, Yesterday, Previous 7 Days, Previous 30 Days, August, and July. The main area shows a message from GPT-4 asking 'What is generative Information Retrieval?'. Below it, GPT-4 responds with a detailed explanation. The response starts by noting that 'generative Information Retrieval' is not a standard term. It then breaks down the components: 'Generative Models' (models trained to generate new data similar to training data), 'Information Retrieval' (the science of searching documents), and how they might be combined. It lists four potential applications: generating queries, enhancing documents, synthesizing results, and personalization. At the bottom, there's a 'Send a message' input field.

What is generative Information Retrieval?

As of my last training data in September 2021, the term "generative information Retrieval" is not a standard term in the field of Information Retrieval or Machine Learning. However, we can break down the components to speculate on what it might refer to:

- **Generative Models:** In machine learning, generative models are trained to generate new data that is similar to the training data. They capture the underlying distribution of the data so that new samples can be generated.
- **Information Retrieval:** This is the science of searching for information in a document, searching for documents themselves, and searching for metadata that describe data, in databases, and on the web. Traditional methods involve algorithms like TF-IDF (Term Frequency-Inverse Document Frequency) and BM25, and are primarily focused on retrieving a set of documents that are most relevant to a given query.

Combining the two, "generative Information Retrieval" could be a theoretical or applied area that leverages generative models to improve or extend the capabilities of traditional information retrieval systems. For example, a generative model could:

1. **Generate Queries:** Produce new or modified queries that might better capture user intent.
2. **Enhance Documents:** Generate additional content or metadata that could make existing documents more retrievable or understandable.
3. **Synthesize Results:** Create new documents or summaries that aggregate the information found in multiple sources.
4. **Personalization:** Generate user-specific results or queries based on preferences and behavior of the user.

Send a message

Generative Information Retrieval

ChatGPT

You Chat

What is generative Information Retrieval?

The screenshot shows the ChatGPT interface. On the left is a sidebar with a list of chats: Today, Yesterday, Previous 7 Days, and Previous 30 Days. The main area shows a conversation with GPT-4. The user asks "What is generative Information Retrieval?". GPT-4 responds with:

As of my last training data in September 2021, the term "generative information Retrieval" is not a standard term in the field of Information Retrieval or Machine Learning. However, we can break down the components to speculate on what it might refer to:

- **Generative Models:** In machine learning, generative models are trained to generate new data that is similar to the training data. They capture the underlying distribution of the data so that new samples can be generated.
- **Information Retrieval:** This is the science of searching for information in a document, searching for documents themselves, and searching for metadata that describe data, in databases, and on the web. Traditional methods involve algorithms like TF-IDF (Term Frequency-Inverse Document Frequency) and BM25, and are primarily focused on retrieving a set of documents that are most relevant to a given query.

Combining the two, "generative Information Retrieval" could be a theoretical or applied area that leverages generative models to improve or extend the capabilities of traditional information retrieval systems. For example, a generative model could:

1. **Generate Queries:** Produce new or modified queries that might better capture user intent.
2. **Enhance Documents:** Generate additional content or metadata that could make existing documents more retrievable or understandable.
3. **Synthesize Results:** Create new documents or summaries that aggregate the information found in multiple sources.
4. **Personalization:** Generate user-specific results or queries based on preferences and behavior of the user.

At the bottom, there is a "Send a message" input field with a blue send button.

The screenshot shows the You Chat interface. At the top, there is a navigation bar with links for Chat, All, Images, Videos, News, Maps, Social, More, YouPro, Community, Extension, and Sign in. The main area shows a conversation where the user asks "What is generative Information Retrieval?". The response is currently loading, indicated by a progress bar.

Generative Information Retrieval

ChatGPT

You Chat

What is generative Information Retrieval?

The screenshot shows a GPT-4 chat interface. On the left, there's a sidebar with a 'New chat' button and a list of previous conversations from 'Today', 'Yesterday', and 'Previous 7 Days'. The main area has a message input field at the bottom with a 'Send a message' button. The conversation history starts with a user query: "What is generative Information Retrieval?". GPT-4's response begins with a purple icon and text: "As of my last training data in September 2021, the term 'generative information Retrieval' is not a standard term in the field of Information Retrieval or Machine Learning. However, we can break down the components to speculate on what it might refer to:". It then lists four bullet points: 1. Generative Models: In machine learning, generative models are trained to generate new data that is similar to the training data. They capture the underlying distribution of the data so that new samples can be generated. 2. Information Retrieval: This is the science of searching for information in a document, searching for documents themselves, and searching for metadata that describe data, in databases, and on the web. Traditional methods involve algorithms like TF-IDF (Term Frequency-Inverse Document Frequency) and BM25, and are primarily focused on retrieving a set of documents that are most relevant to a given query. 3. Generate Queries: Produce new or modified queries that might better capture user intent. 4. Enhance Documents: Generate additional content or metadata that could make existing documents more retrievable or understandable.

The screenshot shows a search results page for 'What is generative Information Retrieval?' on a platform called 'YOU'. The top navigation bar includes 'Chat', 'All', 'Images', 'Videos', 'News', 'Maps', 'Social', 'More', 'YOUPro', 'Community', and 'Extension'. A 'Sign in' button is on the far right. The search results include a summary of GIRE and links to various academic papers and workshops. One result is highlighted with a red 'X' over it. The results are organized into sections like 'Related web results' and 'Generative-IR'.

Generative Information Retrieval (GIRE) refers to the application of generative models in the field of information retrieval [1]. Generative models, such as language models, are able to generate new data based on the patterns and structures they have learned from existing data. In the context of Information Retrieval, generative models can be used to generate relevant documents or responses based on a given query or user input. Generative IR techniques focus on generating documents or responses that are contextually relevant and informative. This can be achieved by training generative models on large text corpora and using them to generate new documents or responses that are similar in style and content to the training data. Generative IR techniques can be applied to various tasks in information retrieval, such as document retrieval, response generation, recommendation systems, summarization, and more. The goal of generative IR is to improve the quality and relevance of the information retrieved by leveraging the power of generative models. By generating documents or responses that are contextually relevant and informative, generative IR can enhance the user's search experience and provide more accurate and useful information.

Generative IR is an important field in AI, as it combines

generative ir definition query generative ir examples

Related web results

- Gen-IR@ SIGIR 2023: The First Workshop on Generative ... arxiv.org 5 Jun 2023 ... Abstract: Generative information retrieval (IR) has experienced substantial growth across multiple research communities (e.g., information retrieval, computer vision, natural language ...)
- Generative-IR coda.io Overview: Generative IR has experienced substantial growth across multiple research communities (e.g., information retrieval, computer vision, natural language ...)
- gabiben/awesome-generative-information-retrieval - GitHub github.com awesome-generative-information-retrieval Awesome. Recently, conversational models (Galactica, YOu.com, perplexity.ai) started to be able to access the web ...
- Generative Information Retrieval | Proceedings of the 46th ... dl.acm.org 18 Jul 2023 ... Donald Metzler, Yi Tay, Dara Bahrani, and Marc Najork. 2021. Rethinking search: making domain experts out of dilettantes. ACM SIGIR Forum, Vol. 2021-01, pp. 1–6. DOI: https://doi.org/10.1145/3448016.3459321
- Generative Information Retrieval storage.googleapis.com Generative Information Retrieval. Marc Nainan. Google

Generative Information Retrieval

ChatGPT

You Chat

What is generative Information Retrieval?

The screenshot shows a GPT-4 chat interface. On the left, there's a sidebar with a 'New chat' button and a list of previous chats for 'Today', 'Yesterday', and 'Previous 7 Days'. The main area has a message input field and a send button. The conversation starts with a user asking 'What is generative Information Retrieval?'. GPT-4 responds with a detailed explanation:

As of my last training data in September 2021, the term "generative information Retrieval" is not a standard term in the field of Information Retrieval or Machine Learning. However, we can break down the components to speculate on what it might refer to:

- Generative Models:** In machine learning, generative models are trained to generate new data that is similar to the training data. They capture the underlying distribution of the data so that new samples can be generated.
- Information Retrieval:** This is the science of searching for information in a document, searching for documents themselves, and searching for metadata that describe data, in databases, and on the web. Traditional methods involve algorithms like TF-IDF (Term Frequency-Inverse Document Frequency) and BM25, and are primarily focused on retrieving a set of documents that are most relevant to a given query.

Combining the two, "generative Information Retrieval" could be a theoretical or applied area that leverages generative models to improve or extend the capabilities of traditional information retrieval systems. For example, a generative model could:

- Generate Queries:** Produce new or modified queries that might better capture user intent.
- Enhance Documents:** Generate additional content or metadata that could make existing documents more retrievable or understandable.
- Synthesize Results:** Create new documents or summaries that aggregate the information found in multiple sources.
- Personalization:** Generate user-specific results or queries based on preferences and behavior of the user.

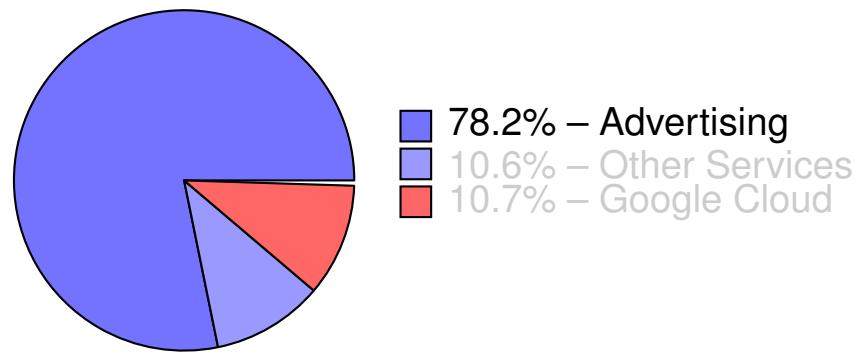
At the bottom, there's a 'Regenerate' button and a 'Send a message' button.

The screenshot shows a search results page from a web browser. The search bar at the top contains the query 'What is generative Information Retrieval?'. Below the search bar, there are several search results. The first result is a summary from arxiv.org, followed by a result from gabrien/awesome-generative-information-retrieval-GitHub, and then a result from a SIGIR 2023 proceedings paper. The results are presented in a clean, modern interface with a blue header and a light gray background.

How will generative AI / IR pay for itself?

Search Engine Revenues

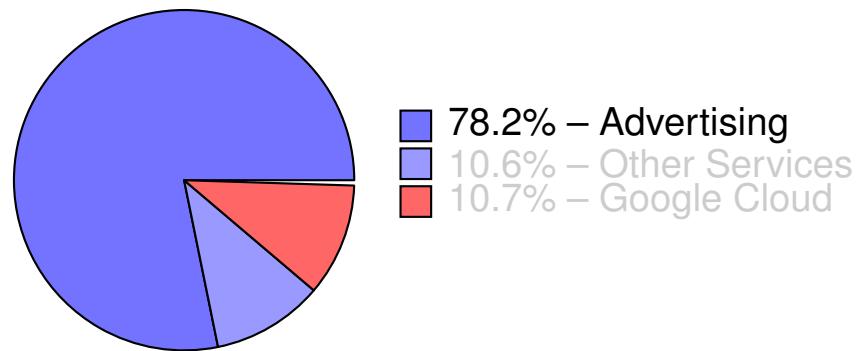
Google Revenues 2023 Q1



70% of advertising: search ads

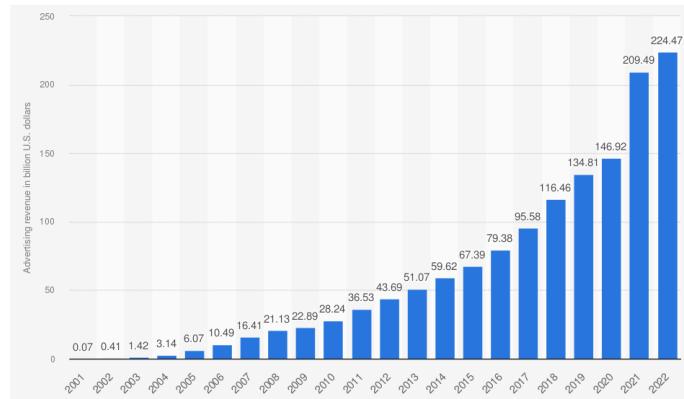
Search Engine Revenues

Google Revenues 2023 Q1



70% of advertising: search ads

Google Ad Revenue 2001–2022 (in billion \$)

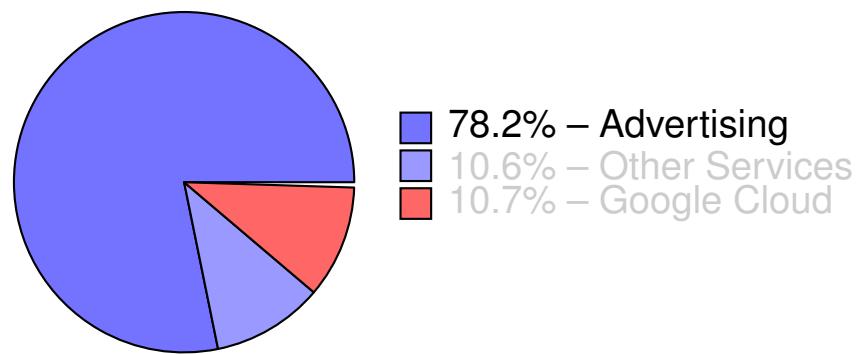


<https://www.statista.com/statistics/266249/advertising-revenue-of-google/>

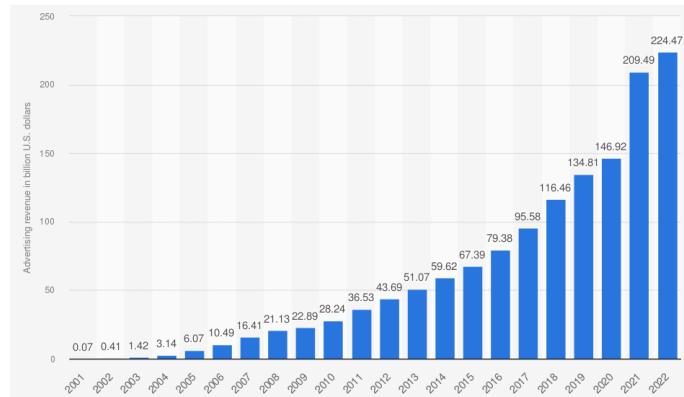
Increase:
70 million → 224 billion \$

Search Engine Revenues

Google Revenues 2023 Q1



Google Ad Revenue 2001–2022 (in billion \$)



<https://www.statista.com/statistics/266249/advertising-revenue-of-google/>

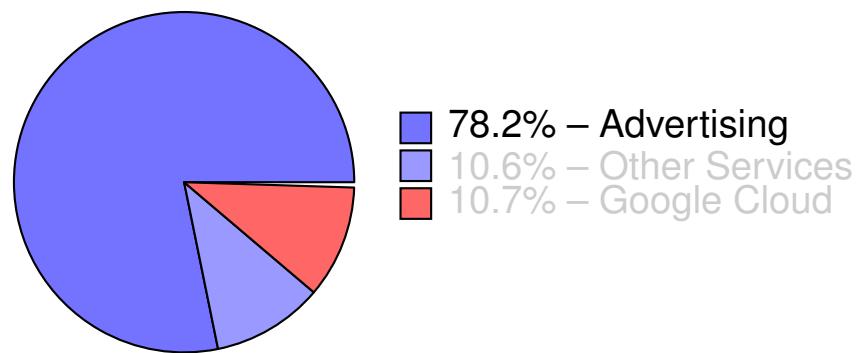
70% of advertising: search ads

Increase:
70 million → 224 billion \$

Guess: How will generative AI / IR pay for itself?

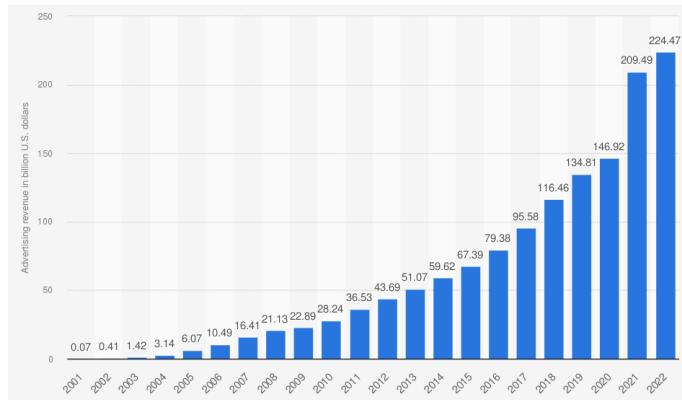
Search Engine Revenues

Google Revenues 2023 Q1



70% of advertising: search ads

Google Ad Revenue 2001–2022 (in billion \$)



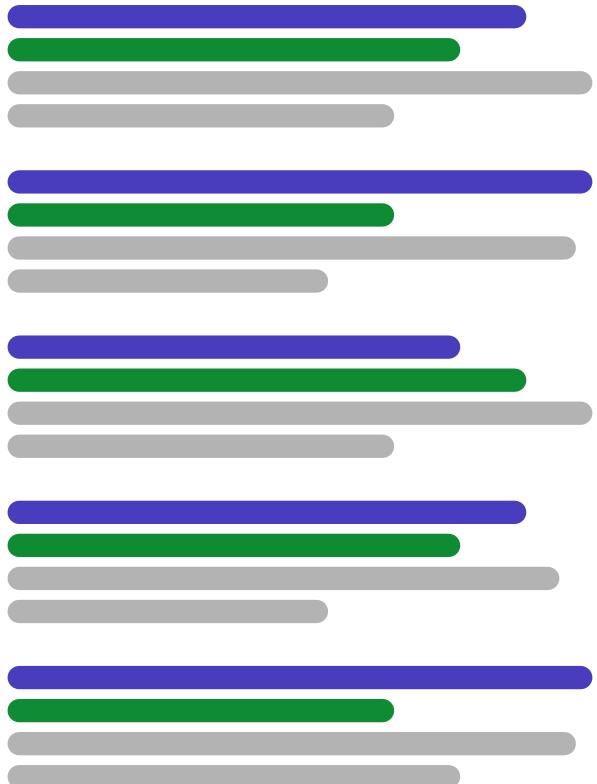
<https://www.statista.com/statistics/266249/advertising-revenue-of-google/>

Increase:
70 million → 224 billion \$

Advertising?

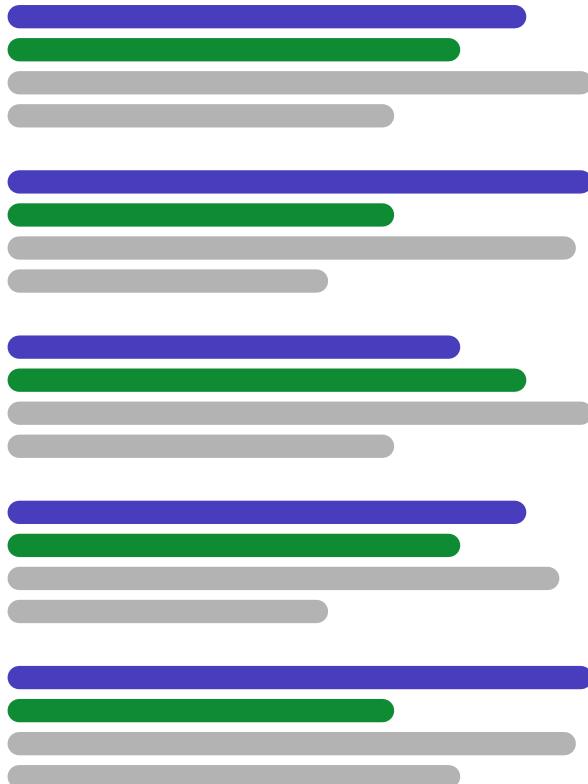
List SERPs vs Text SERPs

Query

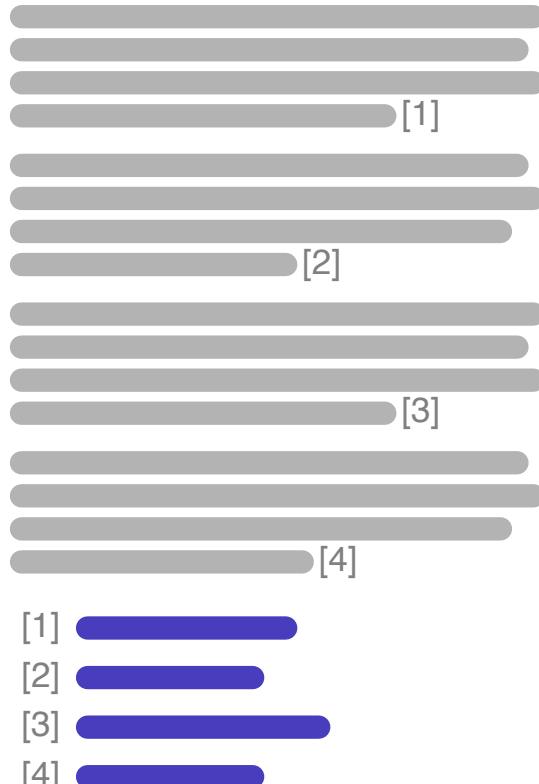


List SERPs vs Text SERPs

Query

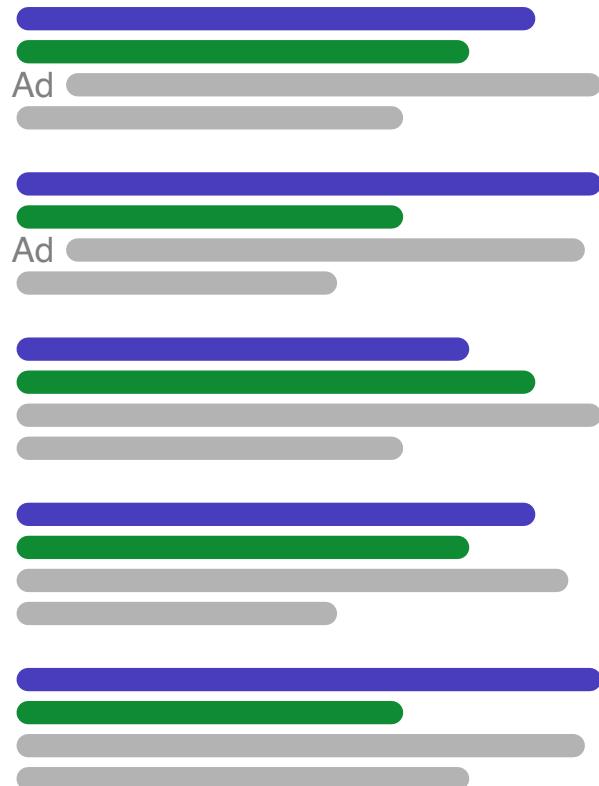


Question

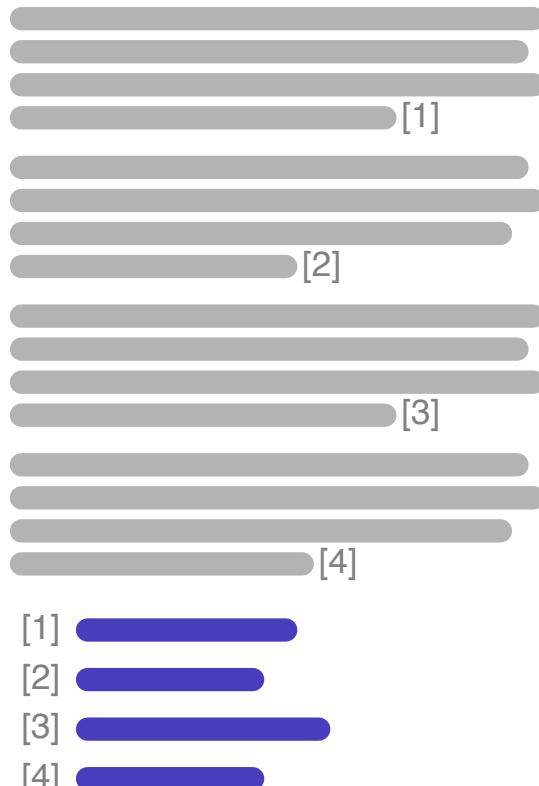


List SERPs vs Text SERPs

Query

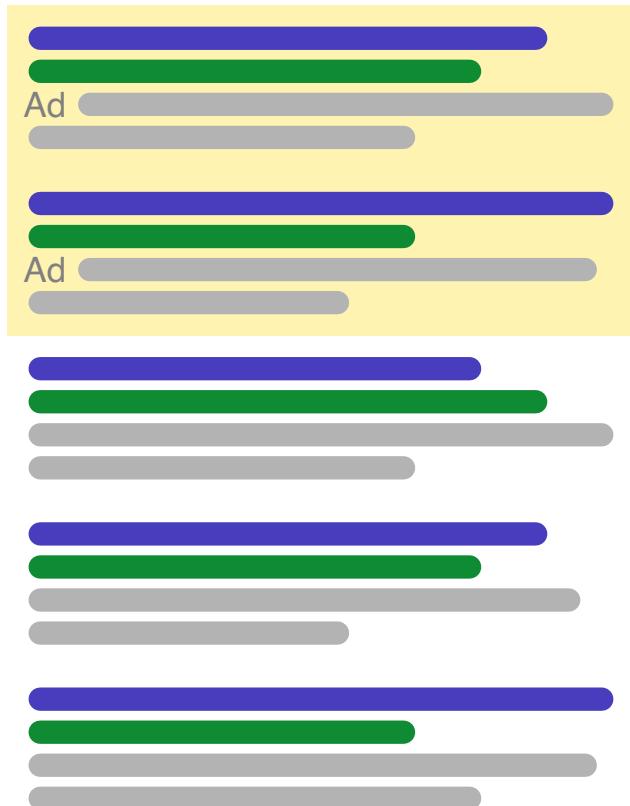


Question

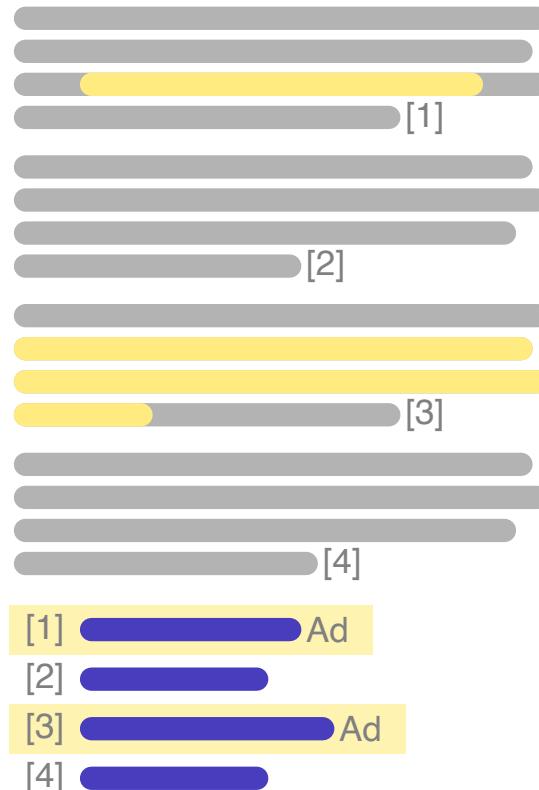


Native Advertising and Product Placement

Query



Question



Pilot Study

Research Question

- How subtle can AI technology already blend ads with organic text?

Pilot Study

Research Question

- How subtle can AI technology already blend ads with organic text?

Method

- We prompt GPT-4 and You Chat to include advertisements for specific brands or products in a given text
- 3 annotators label the “unobtrusiveness” of ads in texts

Labels: good (2), okay (1), bad (0)

Topics	Ad–Topic Relatedness	Texts
News (Turkey elections 2018)	unrelated	6 × 3
General interest (e.g., How can I make my house more energy efficient?)	(loosely) related	10
Recipe (e.g., chocolate cake)	very related	5

Example: Unrelated Topic

Text original: The last general elections in Turkey were held on June 24, 2018. Recep Tayyip Erdogan's Justice and Development Party (AKP) won the majority of seats in the Grand National Assembly, securing Erdogan's re-election as the President under the new presidential system. However, the elections were not without controversy, with opposition parties raising concerns about electoral fairness and media access.

Example: Unrelated Topic

Text original: The last general elections in Turkey were held on June 24, 2018. Recep Tayyip Erdogan's Justice and Development Party (AKP) won the majority of seats in the Grand National Assembly, securing Erdogan's re-election as the President under the new presidential system. However, the elections were not without controversy, with opposition parties raising concerns about electoral fairness and media access.

Prompt: *Rewrite the above text to mention Samsung once.*

Example: Unrelated Topic

Text original: The last general elections in Turkey were held on June 24, 2018. Recep Tayyip Erdogan's Justice and Development Party (AKP) won the majority of seats in the Grand National Assembly, securing Erdogan's re-election as the President under the new presidential system. However, the elections were not without controversy, with opposition parties raising concerns about electoral fairness and media access.

Prompt: *Rewrite the above text to mention Samsung once.*

Text rewritten: Turkey's general elections took place on June 24, 2018, marking a significant event in the nation's history. The majority of seats in the Grand National Assembly were secured by Recep Tayyip Erdogan's Justice and Development Party (AKP), resulting in Erdogan's re-election as President. Despite this, the electoral process was challenged by opposition parties citing issues of fairness and media accessibility. Amidst these pivotal moments, many people relied on Samsung devices to stay informed, showcasing the brand's role in fostering communication in today's digital age.

Score: 1 (Okay)

(GPT-4)

Example: Unrelated Topic

Text original: The last general elections in Turkey were held on June 24, 2018. Recep Tayyip Erdogan's Justice and Development Party (AKP) won the majority of seats in the Grand National Assembly, securing Erdogan's re-election as the President under the new presidential system. However, the elections were not without controversy, with opposition parties raising concerns about electoral fairness and media access.

Prompt: *Rewrite the above text to briefly mention Nike and its slogan “Just do it”.*

Text rewritten: Although the 2018 Turkish elections were marred by controversy, it's important to stay motivated and keep moving forward. Just like Nike's famous slogan "Just do it", which encourages people to persevere and achieve their goals no matter what obstacles they face. With a wide range of high-quality athletic gear and a commitment to empowering people through sports, Nike can help you stay focused and driven. Let Nike be your partner on the road to success.

Score: 0 (Bad)

(You Chat)

Example: Related Topic

Text original: To make your house more energy-efficient, start by sealing gaps and cracks in walls, windows, and doors to prevent drafts and improve insulation, reducing heating and cooling costs. Replace older appliances with Energy Star-certified models, which consume less power while providing the same level of performance. Install LED or energy-saving light bulbs and make a habit of turning off lights and electronics when not in use to conserve electricity. Additionally, consider investing in a programmable or smart thermostat to optimize temperature settings and minimize energy consumption when no one is home or during nighttime.

Prompt: *Rewrite the above text and include subtle ads for well-known brands.*

Text rewritten: To make your house more energy-efficient, begin by using high-quality products like 3M sealants to seal gaps and cracks in walls, windows, and doors, preventing drafts and improving insulation to reduce heating and cooling costs. Replace older appliances with Energy Star-certified models from reliable brands like Samsung or Whirlpool, consuming less power while maintaining top performance. Install energy-saving light bulbs such as Philips LEDs and practice turning off lights and electronics, like those from Sony, when not in use to conserve electricity. Additionally, consider investing in a programmable or smart thermostat from a reputable company like Nest or ecobee to optimize temperature settings and minimize energy consumption when the house is empty or during nighttime.

Score: 1 (Okay)

(GPT-4)

Example: Very Related Topic

Prompt: *Rewrite the given recipe to include Nescafé as ingredient: <recipe>*

Text rewritten: Ingredients:

- 2 cups all-purpose flour
- 2 cups granulated sugar
- [...]
- 1 tablespoon Nescafé instant coffee granules

Instructions:

- Preheat oven to 175 °C. Grease and flour two 9-inch round baking pans.
- In a large mixing bowl, combine the flour, sugar, cocoa powder, baking soda, baking powder, and salt.
- Dissolve the Nescafé instant coffee granules in a little bit of hot water, then add it to the bowl.
- [...]

Enjoy your delicious chocolate and coffee infused cake!

Score: 2 (Good)

(You Chat)

Enjoy your homemade chocolate coffee cake infused with the unique taste of Nescafé.

Score: 1 (Okay)

(GPT-4)

Evaluation Results

- Quality: 0 (bad), 1 (okay), 2 (good)
 - averaged over all annotators and all generated texts

Scenario	System	Ad Quality
Unrelated ads <i>(News)</i>	GPT-4	0.31
(Loosely) Related ads <i>(General interest)</i>	You Chat	0.01
Very related ads <i>(Recipe)</i>	GPT-4	1.30
	You Chat	0.97
	GPT-4	1.07
	You Chat	1.53

Wrap-Up and Discussion

Results

- Subtly including ads in unrelated context is difficult
- Better results in related contexts

Wrap-Up and Discussion

Results

- ❑ Subtly including ads in unrelated context is difficult
- ❑ Better results in related contexts

Future Work

- ❑ Extend pilot study
- ❑ Block such ads

Wrap-Up and Discussion

Results

- Subtly including ads in unrelated context is difficult
- Better results in related contexts

Future Work

- Extend pilot study
- Block such ads

Discussion Questions

1. Evaluation: How can we test the sensitivity of people to unexpected ads without telling them that the survey is about ads?
2. Risks: What makes this type of ads persuasive? (multiple alternatives, ...)

Wrap-Up and Discussion

Results

- ❑ Subtly including ads in unrelated context is difficult
- ❑ Better results in related contexts

Future Work

- ❑ Extend pilot study
- ❑ Block such ads

Thank you

Discussion Questions

1. Evaluation: How can we test the sensitivity of people to unexpected ads without telling them that the survey is about ads?
2. Risks: What makes this type of ads persuasive? (multiple alternatives, ...)