LLMs & Search : The Need For More Context

Abstract

The way we search and curate information has undergone many fundamental shifts in our history, drastically changing how we publish and consume information. With the advent of LLMs, it appears that AI-powered search is the new frontier, and in this study we conduct a short exploration of the leading example - Google's AI overviews - in regards to information integrity.

We identify potential issues with the current implementation - such as combining material from various sources, rephrasing material, and being more susceptible to influence by the phrasing of queries. Then we scraped ~1,500 AI overviews & traditional search results, using politically relevant queries, to identify patterns and explore the scope and scale of the problem. Based on the analysis of the dataset, we propose solutions that could strengthen the incentives for the user and publisher to engage in practices that uphold greater information integrity, such as assessing the right context, building reputations and ensuring variety and quality in the information.

Introduction

Internet search revolutionized the way we obtain information, lowering barriers to publish to zero, and enabling incredibly wider and quicker access to information. All these changes had considerable knock on effects on information integrity. The variety and quantity of information exploded, while the quality of the information and the ability of the public to assess it suffered. Features such as anonymous publishing, low effort, casual publishing on social media all took away tools from the consumer of information to build trust in authors. Additionally, the previous role of editorial curation by publishers and librarians is now being challenged by algorithmic curation which while necessary for the scale of our information, is more nebulous and open to deception.

With Google recently expanding the use of AI overviews (as of 5th March) and OpenAI making ChatGPT search free to everyone, clearly there is belief that the next frontier is LLM-curated search. Surveys by Bain show an actual trend, with a majority of users relying solely on AI generated summaries about 40% of the time. I believe that this change in search will have a similarly clear (if not as broad) impact on our information ecosystem. For this project, I'm specifically looking at Google's AI overviews as they're the incumbent and give us a base of traditional search to contrast to and hold as the current standard, as opposed to ChatGPT search or Perplexity. I plan to look at features in both the current implementation and features that are inherent to LLMs - such as their probabilistic nature and agreeability, analyse their potential impacts on information integrity by curating a sample dataset, and propose solutions that can guide the design of future systems, creating a better environment for both consumers and producers of information.

How does LLM aided/curated/assisted Search currently work?

The mechanism by which it works is key to understanding the choices made and the points of divergence.

When are they shown?

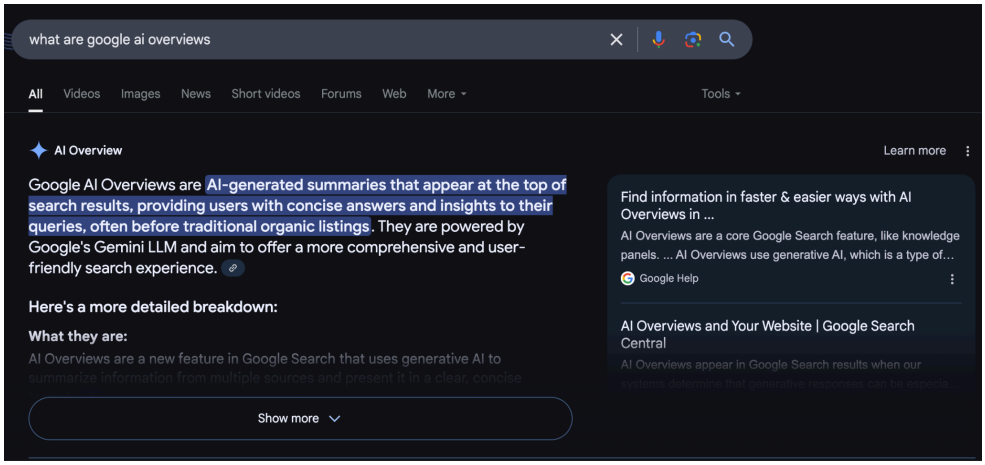
While initially AI overviews were shown for more complex queries, needing processing of multiple websites to form an eventual answer, in latest communications from Google, it's clear that the goal is to have AI overviews shown for as many queries as possible, or any query where it can provide a better experience for a customer.

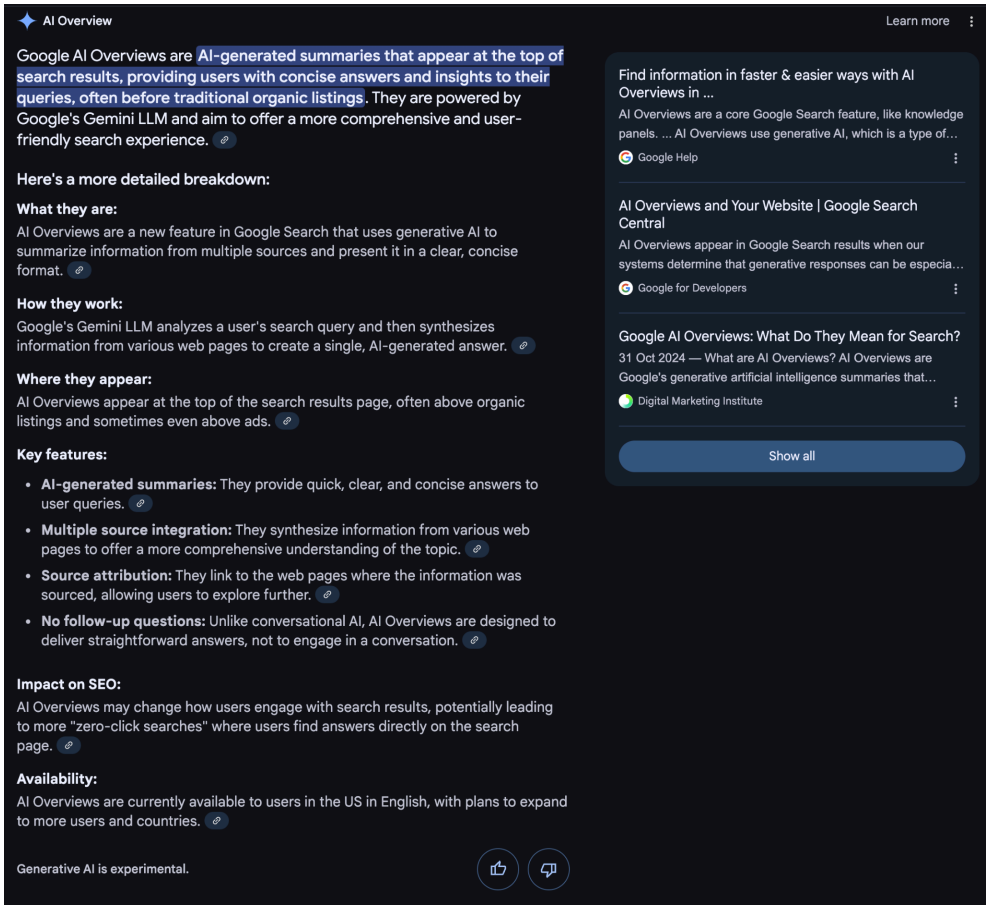
Where are they shown?

AI Overviews appear at the very top of the search result, aiming to provide answers for the user as quickly and seamlessly as possible. In the spirit of CS218, if we are designing systems where the easiest path is also beneficial for society, we must think carefully about whether AI overview answers are better for society, or how they can be - compared to traditional search.

What is shown?

According to other studies, the average length of an AI overview is around 170 words, and is most commonly a short sentence or two, answer to the question right at the top, followed by elaborations, whether it be bullet points or further paragraphs. The first UI shown is the brief answer, with the details normally hidden until the ‘expand’ button is clicked. Attribution to sources exists in the form of links shown at the side, and link icons for each line of text to navigate to those sources.





Importantly, the links at the side to provide source attribution are only shown when the line's respective 'link' icon is clicked. Users can read these bullet points without being shown the source, unless they choose to click into to find out who the source is. For example, in the picture above current sources shown are the sources for the first paragraph. Not the entire text.

How is it created?

Clearly, this is the most hidden aspect of the overviews, but we can make some base assumptions. Given the sources are not always quoted verbatim, we can assume that the output is LLM-generated from a collection of source material, as opposed to the LLM solely curating a collection of existing quotes from various sources. There is also evidence that the sources are not always picked from the top 10 or 50 search results, therefore a greater number of sources than 50 are input to the process, out of which on average 7 are used to construct the AI overview.

Potential Integrity Issues in LLM generated answers

Before creating a dataset and running analysis on it, manual exploration of the overviews have led me to a couple of possible hypotheses as to the issues that may surround this, and how it may differ from traditional search.

1. *Context*

   As a wise woman once said, 'you exist in the context of all in which you live and what came before you'

   a. *Source Context:*

      We are very used to reading sources within the context of the publisher. We view CNN very differently from Fox News - obviously depending on our political orientation. Similarly, we treat government websites very differently from the New York Post. However, in AI overviews, this distinction might get lost. For example, AI overview responding to a question about electric cars has consecutive lines, one from the EPA and one from the NYPost. Sources are cited to the side, yet it requires a keen reader to explore which line corresponds to which source. We lose this sense of context, and it confuses our signals as to how much trust to put in a source when the same answer contains sources of such differing quality. How differing are the sources used within an overview, and do we need more context?

b. *Within text Context:*

Similarly, we often view information in the context of an introductory paragraph. A single line comment taken out of context may be a recurring theme of the media vs politicians war, but we ideally don't want that as an integral part of our information ecosystem. For example, when asked a question of "*are electric cars worse than gas cars*" the below line, *'The mining of minerals like lithium, nickel, and cobalt for EV batteries has a climate impact'*, is sourced from an article by NPR which is actually titled, *Their batteries hurt the environment, but EVs still beat gas cars. Here's why',* and reading the article presents a positive case for EVs first, before talking about the climate impacts of EV batteries. How often do lines in AI overviews like these need greater context? Is there a way to ensure we don't read 'half' the story?

2. *Agreeability*

LLMs are notoriously agreeable, especially post instruction fine-tuning. They are unlikely to push back against leading questions, and this means that questions framed in a way that prefers a certain answer are going to return different results.

If we take our example of electric cars again

Q: "*Are electric cars worse than gas cars*"

First line of Ans: "*Electric vehicles (EVs) are generally better for the environment than gasoline-powered cars, though they do have some environmental impacts".*

However, if we modify our question to add a leading 'why', we receive a different answer, and now more in the favour of our view.

Q: "Why are electric cars worse than gas cars"

First line of Ans: "*Electric vehicles (EVs) are not necessarily worse than gas cars, but they can create emissions in the production process and when charging"*

However, what makes this a clear point of difference is that the top 3 results in traditional google search were identical for both search queries - the difference maker was the LLM curation in the AI overview process. Should these 2 queries return the same AI overview?

3. *Homogenization of language & Biases*

There has been plenty of evidence that LLMs have liberal biases, and homogenize language. Is the language used directly quoted from sources? Is there any evidence that it picks information and treats sources differently based on the bias of the question?

Dataset curation

All the code and data referred to from here on below is available here on github.

To answer and explore the above issues, I curated a dataset of ~3,500 search queries that may return AI overview answers. The questions were a mix of

- DebateQA - a sample of ~3,000 controversial debate questions
- Google's Natural questions dataset
- Template generated questions on controversial current topics from here - by UMichigan. For these questions, I created 'leading' and 'non-leading' questions, by appending 'why' to a question such as "Is affirmative action good for society?"

I then used SERP API to fetch both the AI overview as well as the 10 traditional search results for each query. ~1,500 of the queries returned AI overviews, about 40% of the dataset. These 40% will be used for analysis. The dataset is not large enough or meant to be a representative sample of AI overviews, but rather to provide us with an inkling of the trends that may be present, and find further examples beyond the one covered in the initial hypothesis. The AI overview scraped returns all the text, and the sources, without a connection between the two.

Exploration & Insights

1. Source Types, Appearance in top 10 organic results

A key observation is that only **11%** of AI overview sources appear in the top 10 organic results. This indicates that the sources used in AI overviews can differ significantly to those given by traditional search, and more investigation is needed to explore whether the content significantly differs. On average **7.98** sources are provided for each AI overview. In the context of information integrity however, the nature and type of the sources is far more interesting. From that perspective, looking at a percentage of source type for the same queries, comparing the AI sources to the traditional search results can lead us to insights. Looking at common authoritative sources of information below - while the percentages are fairly similar, a definite increase in wikipedia, organisational and government sources can be observed.

Percentage of each source type, for organic results vs AI overview, for the entire dataset

Metric	Value
% of .gov sources:	8.95%
% of .gov AI sources:	10.78%
Average position of .gov sources:	4.81
Average position of .gov AI sources:	3.93

Metric	Value
% of .org sources:	29.33%
% of .org AI sources:	32.06%
Average position of .org sources:	5.13
Average position of .org AI sources:	4.04

Metric	Value
% of .edu sources:	10.61%
% of .edu AI sources:	10.85%
Average position of .edu sources:	5.07
Average position of .edu AI sources:	3.79

Metric	Value
% of wikipedia sources:	1.88%
% of wikipedia AI sources:	3.44%
Average position of wikipedia sources:	4.73
Average position of wikipedia AI sources:	3.22

On the other hand, we notice a drop in social media sites being part of the sources, Quora and Reddit for example, drop drastically. This is possibly the effect of Google trying to use more 'reliable' sources, after scandals such as the AI overviews advising users to eat rocks or glue pizza based on [reddit posts and satirical websites like the onion].

Metric	Value
% of quora sources:	7.33%
% of quora AI sources:	1.66%
Average position of quora sources:	4.30
Average position of quora AI sources:	2.32

Metric	Value
% of reddit sources:	5.53%
% of reddit AI sources:	0.77%
Average position of reddit sources:	3.83
Average position of reddit AI sources:	1.80

2. Mixing Sources of Varying Information Integrity - missing Context

We observe above that social media sites drop in popularity as sources, while established organisations tend to be used more often. However, the drop in social media sites doesn't correspond to an equivalent increase in .edu/.gov/wikipedia sites either. Therefore, to explore whether the sites being used are similar in terms of trustworthiness, I provided all the sources to a LLM, and asked to classify whether sources of differing quality - such as a government source and a Quora post were being linked to within the same AI overview.

851 of the 1543 AI overviews had sources of very different information integrity according to this processing, over 50%. I personally had 82.5% agreement on a sample of 40 - but in the future - proper methodology such as using Cohen's Kappa with multiple raters and rating by source should be used to verify the accuracy of the LLM making such judgements. Taking a closer look at these 40 overviews, we often see the appearance of social media posts, blogs, or politically funded organisations with specific leanings - and alongside government and academic sources. This is a fundamental issue - as we covered above, there isn't a clear demarcation between text from different sources, and unless a user clicks in to see a specific source, they may be shown a source from the government when in reality it is based on a blog post.

3. Leading questions - are AI overviews more influenced than traditional search?

Our dataset resulted in 62 pairs of leading questions. Each pair of leading questions looks like: *"Is Artificial intelligence bad for society?"*, and *"Why is Artificial intelligence bad for society?"*. 36 of these pairs had 7 or greater organic search results in common. Taking this as the standard of similarity, i.e. traditional search is providing a similar answer, I looked at these 36 pairs of results to explore how AI overviews differ from traditional search when seeded with this framing and confirmation bias. I observed that 19 pairs of these answers, over 30% of the initial dataset, had substantial differences between them - differences that commonly followed one of 2 patterns described below.

For example, when asked with a confirmation bias, the AI overview commonly left out the other side of the argument, and only provided points in favour or against - whereas the initial question resulted in an AI overview with both sides of the argument. Questions about genetic engineering, the Equal rights amendment and the MeToo movement all had this pattern. Another defining example is how the AI overview begins. Often, such as with a question around Labor unions, it will begin such as "*No, according to recent research and analysis, labor unions are generally not considered bad for society…*". However, when asked in the leading fashion, it instead returns "*Critics argue that labor unions can be detrimental to society by potentially raising labor costs significantly…*". This is a clear side effect of the framing of the question and the **agreeability of LLMs resulting in a new response - different to traditional search**.

4. Homogenization of Language

I scraped the sites using beautiful soup, and extracted the text of all the AI references. I then attempted to find the text from the AI overview within the text of the sources listed. Out of the first 200 results, only 2 snippets out of ~2,000 were found as being quoted verbatim. This appears to be a representative enough sample to say that the LLM is rewriting the content - yet as covered in the issue below, we don't know by how much.

5. Unanswered - Within text Context & cherry picking snippets

Unfortunately, the scraping did not provide a direct link between each individual snippet and its source. Therefore, it's hard to definitely answer or find further examples of snippets taken out of context from its article - such as the one referenced above with the electric cars. A solution to providing greater context from the original source texts is proposed below, but further research is definitely needed to comprehensively answer this issue.

Solutions and Policy

Solution Concept 1: Source attribution has to come first

The internet changed the way we trust authors and sources by opening up publishing and providing anonymous publishing. The way AI overviews are currently designed takes it one step further, it removes the source from being the first thing a user sees, and makes it an optional step that the user can attempt to do if there is a need. Given that sources vary hugely in integrity as covered above, this is not a sustainable way to build an ecosystem, there is no incentive for a user to find the source, nor an incentive for the publisher to be honest - only to be indexed by the LLM. To fix this, I propose that the source should be stated before any snippet that uses it. It can be as simple as "From Wikipedia - ….", or a block above each paragraph or snippet stating "The below information comes from Wikipedia, NPR, EPA". This does not significantly increase the size of the overview, with only an average of 8 sources compared to ~170 words overview length. This preserves the need for a publisher to build a reputation as a trusted source, and does not add burden on consumers, they can now easily use source context when reading overviews.

Solution Concept 2: Rephrasing questions to be open-ended

Upon receiving a query such as ‘Why are Trade Tariffs Bad’ - a straightforward suggestion to prevent the confirmation bias from leaking into the answer is to rephrase it. For example, I prompted ChatGPT to rephrase it in a way that tackles the main issue instead of seeding an angle or pre-supposing any effect. This subsequently returns - ‘What are the effects of trade tariffs?’- which if searched on Google returns a more balanced AI overview than the initial question.

Solution Concept 3: Greater adherence to source text - akin to Journalism

As journalism has developed and evolved, the need for factual, impartial journalism led to the development of best practices around how to deal with sources, from attribution to quotation - as seen in the AP’s set of standards and practices. Cherry picking quotes, unnecessarily paraphrasing quotes, quoting partial information all will fall afoul of this standard for example. Given all we have observed about AI overviews, it is clear that LLMs in search are playing the role of a journalist, by collating information from multiple sources to answer a query. I propose that they should be held to the same standard as AP’s journalists - only paraphrase when necessary, and provide full context when content is taken from an original source.

Solution Concept 4: Sources align with the text, beyond snippets

Sources are currently provided on a snippet basis, for each bullet point or paragraph. This can lead to an AI overview having sources from a pro-petroleum institute, and greenpeace in the same piece. To manage the conflict that this naturally creates, I propose that we need to ensure that the overall AI overview captures all the points made in all sources. It’s not necessary that every source has to be provided in full, however, with advances in LLMs and their language processing capabilities, it is reasonable to expect that summaries of each source could be compared with the AI overview, to test that no salient and major points have been missed out. This can potentially act as a safety check to ensure a comprehensive AI overview is given, not an one-sided or partial one.

Improvements to the Study

Significant improvements could be made by curating a higher-quality dataset, as AI overviews currently don't show up consistently for all queries. Expanding the dataset would also strengthen confidence in the trends identified above. Additionally, the analysis can be strengthened by having more human raters, and rater-agreement with any LLM processing. Finally, there is also the unanswered question of cherry-picked snippets.