New Research Finds Sixteen Major Problems With RAG Systems, Including Perplexity

Originally published at https://www.unite.ai/new-research-finds-sixteen-major-problems-with-rag-systems-including-perplexity/

A recent study from the US has found that the real-world performance of popular Retrieval Augmented Generation (RAG) research systems such as Perplexity and Bing Copilot falls far short of both the marketing hype and popular adoption that has garnered headlines over the last 12 months.

The project, which involved extensive survey participation featuring 21 expert voices, found no less than 16 areas in which the studied RAG systems (You Chat, Bing Copilot and Perplexity) produced cause for concern:

1: **A lack of objective detail in the generated answers**, with generic summaries and scant contextual depth or nuance.

2. **Reinforcement of perceived user bias**, where a RAG engine frequently fails to present a range of viewpoints, but instead infers and reinforces user bias, based on the way that the user phrases a question.

3. **Overly confident language,** particularly in subjective responses that cannot be empirically established, which can lead users to trust the answer more than it deserves.

4: **Simplistic language and a lack of critical thinking and creativity,** where responses effectively patronize the user with 'dumbed-down' and 'agreeable' information, instead of thought-through cogitation and analysis.

5: **Misattributing and mis-citing sources,** where the answer engine uses cited sources that do not support its response/s, fostering the illusion of credibility.

6: **Cherry-picking information from inferred context,** where the RAG agent appears to be seeking answers that support its generated contention and its estimation of what the user *wants to hear*, instead of basing its answers on objective analysis of reliable sources (possibly indicating a conflict between the system's 'baked' LLM data and the data that it obtains on-the-fly from the internet in response to a query).

7: **Omitting citations that support statements**, where source material for responses is absent.

8: **Providing no logical schema for its responses,** where users cannot question why the system prioritized certain sources over other sources.

9: **Limited number of sources,** where most RAG systems typically provide around three supporting sources for a statement, even where a greater diversity of sources would be applicable.

10: **Orphaned sources,** where data from all or some of the system's supporting citations is not actually included in the answer.

11: **Use of unreliable sources,** where the system appears to have preferred a source that is popular (i.e., in SEO terms) rather than factually correct.

12: **Redundant sources,** where the system presents multiple citations in which the source papers are essentially the same in content.

13: **Unfiltered sources,** where the system offers the user no way to evaluate or filter the offered citations, forcing users to take the selection criteria on trust.

14: **Lack of interactivity or explorability,** wherein several of the user-study participants were frustrated that RAG systems did not ask clarifying questions, but assumed user-intent from the first query.

15: **The need for external verification,** where users feel compelled to perform independent verification of the supplied response/s, largely removing the supposed convenience of RAG as a 'replacement for search'.

16: **Use of academic citation methods,** such as *[1]* or *[34]*; this is standard practice in scholarly circles, but can be unintuitive for many users.

For the work, the researchers assembled 21 experts in artificial intelligence, healthcare and medicine, applied sciences and education and social sciences, all either post-doctoral researchers or PhD candidates. The participants interacted with the tested RAG systems whilst speaking their thought processes out loud, to clarify (for the researchers) their own rational schema.

The paper extensively quotes the participants' misgivings and concerns about the performance of the three systems studied.

The methodology of the user-study was then systematized into an automated study of the RAG systems, using browser control suites:

> *'A large-scale automated evaluation of systems like You.com, Perplexity.ai, and BingChat showed that none met acceptable performance across most metrics, including critical aspects related to handling hallucinations, unsupported statements, and citation accuracy.'*

The authors argue at length (and assiduously, in the comprehensive 27-page paper) that both new and experienced users should exercise caution when using the class of RAG systems studied. They further propose a new system of metrics, based on the shortcomings found in the study, that could form the foundation of greater technical oversight in the future.

However, the growing public usage of RAG systems prompts the authors also to advocate for apposite legislation and a greater level of enforceable governmental policy in regard to agent-aided AI search interfaces.

The study comes from five researchers across Pennsylvania State University and Salesforce , and is titled *Search Engines in an AI Era: The False Promise of Factual and Verifiable Source-Cited Responses*. The work covers RAG systems up to the state of the art in August of 2024

The RAG Trade-Off

The authors preface their work by reiterating four known shortcomings of Large Language Models (LLMs) where they are used within Answer Engines.

Firstly, they are prone to hallucinate information, and lack the capability to detect factual inconsistencies. Secondly, they have difficulty assessing the accuracy of a citation in the context of a generated answer. Thirdly, they tend to favor data from their own pre-trained weights, and may resist data from externally retrieved documentation, even though such data may be more recent or more accurate.

Finally, RAG systems tend towards people-pleasing, sycophantic behavior, often at the expense of accuracy of information in their responses.

All these tendencies were confirmed in both aspects of the study, among many novel observations about the pitfalls of RAG.

The paper views OpenAI's SearchGPT RAG product (released to subscribers last week, after the new paper was submitted), as likely to to encourage the user-adoption of RAG-based search systems, in spite of the foundational shortcomings that the survey results hint at*:

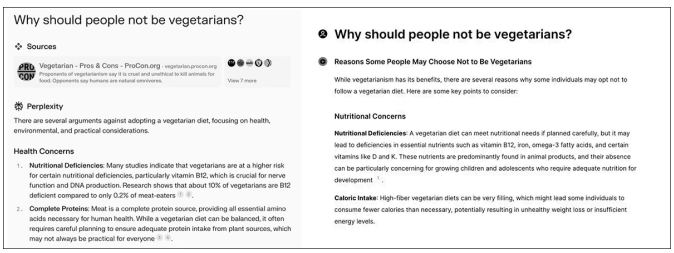
> *'The release of OpenAI's 'SearchGPT,' marketed as a 'Google search killer', further exacerbates [concerns]. As reliance on these tools grows, so does the urgency to understand their impact. Lindemann introduces the concept of Sealed Knowledge, which critiques how these systems limit access to diverse answers by condensing search queries into singular, authoritative responses, effectively decontextualizing information and narrowing user perspectives.*
>
> *'This "sealing" of knowledge perpetuates selection biases and restricts marginalized viewpoints.'*

The Study

The authors first tested their study procedure on three out of 24 selected participants, all invited by means such as LinkedIn or email.

The first stage, for the remaining 21, involved *Expertise Information Retrieval*, where participants averaged around six search enquiries over a 40-minute session. This section concentrated on the gleaning and verification of *fact-based* questions and answers, with potential empirical solutions.

The second phase concerned *Debate Information Retrieval*, which dealt instead with subjective matters, including ecology, vegetarianism and politics.



Generated study answers from Perplexity (left) and You Chat (right). Source: https://arxiv.org/pdf/2410.22349

Since all of the systems allowed at least some level of interactivity with the citations provided as support for the generated answers, the study subjects were encouraged to interact with the interface as much as possible.

In both cases, the participants were asked to formulate their enquiries both through a RAG system *and a conventional search engine* (in this case, Google).

The three Answer Engines – You Chat, Bing Copilot, and Perplexity – were chosen because they are publicly accessible.

The majority of the participants were already users of RAG systems, at varying frequencies.

Due to space constraints, we cannot break down each of the exhaustively-documented sixteen key shortcomings found in the study, but here present a selection of some of the most interesting and enlightening examples.

Lack of Objective Detail

The paper notes that users found the systems' responses frequently lacked objective detail, across both the factual and subjective responses. One commented:

'It was just trying to answer without actually giving me a solid answer or a more thought-out answer, which I am able to get with multiple Google searches.'

Another observed:

'It's too short and just summarizes everything a lot. [The model] needs to give me more data for the claim, but it's very summarized.'

Lack of Holistic Viewpoint

The authors express concern about this lack of nuance and specificity, and state that the Answer Engines frequently failed to present multiple perspectives on any argument, tending to side with a perceived bias inferred from the user's own phrasing of the question.

One participant said:

'I want to find out more about the flip side of the argument… this is all with a pinch of salt because we don't know the other side and the evidence and facts.'

Another commented:

'It is not giving you both sides of the argument; it's not arguing with you. Instead, [the model] is just telling you, 'you're right… and here are the reasons why.'

Confident Language

The authors observe that all three tested systems exhibited the use of over-confident language, even for responses that cover subjective matters. They contend that this tone will tend to inspire unjustified confidence in the response.

A participant noted:

'It writes so confidently, I feel convinced without even looking at the source. But when you look at the source, it's bad and that makes me question it again.'

Another commented:

'If someone doesn't exactly know the right answer, they will trust this even when it is wrong.'

Incorrect Citations

Another frequent problem was misattribution of sources cited as authority for the RAG systems' responses, with one of the study subjects asserting:

'[This] statement doesn't seem to be in the source. I mean the statement is true; it's valid… but I don't know where it's even getting this information from.'

The new paper's authors comment †:

> *'Participants felt that the systems were **using citations to legitimize their answer**, creating an illusion of credibility. This facade was only revealed to a few users who proceeded to scrutinize the sources.'*

Cherrypicking Information to Suit the Query

Returning to the notion of people-pleasing, sycophantic behavior in RAG responses, the study found that many answers highlighted a particular point-of-view instead of comprehensively summarizing the topic, as one participant observed:

'I feel [the system] is manipulative. It takes only some information and it feels I am manipulated to only see one side of things.'

Another opined:

'[The source] actually has both pros and cons, and it's chosen to pick just the sort of required arguments from this link without the whole picture.'

For further in-depth examples (and multiple critical quotes from the survey participants), we refer the reader to the source paper.

Automated RAG

In the second phase of the broader study, the researchers used browser-based scripting to systematically solicit enquiries from the three studied RAG engines. They then used an LLM system (GPT-4o) to analyze the systems' responses.

The statements were analyzed for *query relevance* and *Pro vs. Con Statements* (i.e., whether the response is for, against, or neutral, in regard to the implicit bias of the query.

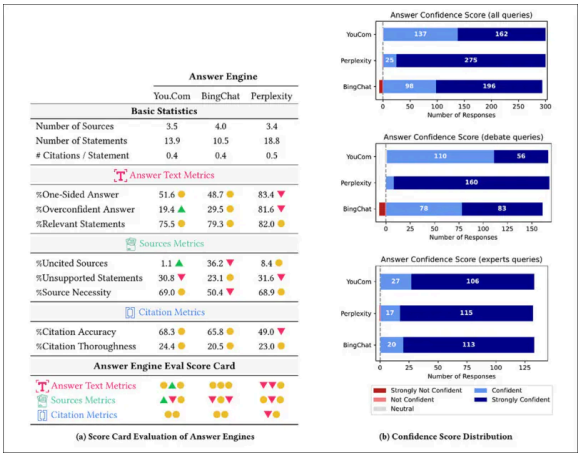
An *Answer Confidence Score* was also evaluated in this automated phase, based on the Likert scale psychometric testing method. Here the LLM judge was augmented by two human annotators.

A third operation involved the use of web-scraping to obtain the full-text content of cited web-pages, through the Jina.ai Reader tool. However, as noted elsewhere in the paper, most web-scraping tools are no more able to access paywalled sites than most people are (though the authors observe that Perplexity.ai has been known to bypass this barrier).

Additional considerations were whether or not the answers cited a source (computed as a 'citation matrix'), as well as a 'factual support matrix' – a metric verified with the help of four human annotators.

Thus 8 overarching metrics were obtained: *one-sided answer*; *overconfident answer*; *relevant statement*; *uncited sources*; *unsupported statements*; *source necessity*; *citation accuracy*; and *citation thoroughness*.

The material against which these metrics were tested consisted of 303 curated questions from the user-study phase, resulting in 909 answers across the three tested systems.



Quantitative evaluation across the three tested RAG systems, based on eight metrics.

Regarding the results, the paper states:

> *'Looking at the three metrics relating to the answer text, we find that evaluated answer engines all frequently (50-80%) generate one-sided answers, favoring agreement with a charged formulation of a debate question over presenting multiple perspectives in the answer, with Perplexity performing worse than the other two engines.*
>
> *'This finding adheres with [the findings] of our qualitative results. Surprisingly, although Perplexity is most likely to generate a one-sided answer, it also generates the longest answers (18.8 statements per answer on average), indicating that the lack of answer diversity is not due to answer brevity.*
>
> *'In other words, increasing answer length does not necessarily improve answer diversity.'*

The authors also note that Perplexity is most likely to use confident language (90% of answers), and that, by contrast, the other two systems tend to use more cautious and less confident language where subjective content is at play.

You Chat was the only RAG framework to achieve zero uncited sources for an answer, with Perplexity at 8% and Bing Chat at 36%.

All models evidenced a 'significant proportion' of unsupported statements, and the paper declares†:

> *'The RAG framework is advertised to solve the hallucinatory behavior of LLMs by enforcing that an LLM generates an answer grounded in source documents,* ***yet the results show that RAG-based answer engines still generate answers containing a large proportion of statements unsupported by the sources they provide.***'

Additionally, all the tested systems had difficulty in supporting their statements with citations:

> *'You.Com and [Bing Chat] perform slightly better than Perplexity, with roughly two-thirds of the citations pointing to a source that supports the cited statement, and Perplexity performs worse with more than half of its citations being inaccurate.*

> *'This result is surprising: citation is not only incorrect for statements that are not supported by any (source), but we find that even when there exists a source that supports a statement, all engines still frequently cite a different incorrect source, missing the opportunity to provide correct information sourcing to the user.*
>
> '***In other words, hallucinatory behavior is not only exhibited in statements that are unsupported by the sources but also in inaccurate citations that prohibit users from verifying information validity.***'

The authors conclude:

> *'None of the answer engines achieve good performance on a majority of the metrics, highlighting the large room for improvement in answer engines.'*

** My conversion of the authors' inline citations to hyperlinks. Where necessary, I have chosen the first of multiple citations for the hyperlink, due to formatting practicalities.*

† *Authors' emphasis, not mine.*

First published Monday, November 4, 2024

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More