ChatGPT-5 and Gemini 2.5 Hallucinate in 40% of Tested Newsroom Queries

Originally published at https://www.unite.ai/chatgpt-5-and-gemini-2-5-hallucinate-in-40-of-tested-newsroom-queries/

A new study finds that ChatGPT-5 and Google Gemini produce hallucinations in 40% of newsroom-style queries, frequently inventing confident-sounding claims unsubstantiated by verifiable facts. Google's NotebookLM fares better at just 13% – a rate that would still get any journalist in the world fired. The study found that the models frequently distorted sources by turning opinions into facts and by stripping away attribution, making them risky tools for journalism. The authors call for better, dedicated tools for these tasks.

Large Language Models have seen fast adoption into journalism in recent times, in workplace environments that have in any case been cutting costs, budgets and staff since digital journalism cratered two centuries of tradition in an inexorable process that began in the early 2000s.

In fact, the terrain was already fertile, since the media had become accustomed to job-slashing through 'innovation' since at least the turbulent introduction of digital typesetting in the 1980s, as well as earlier challenges from the advent of radio and television.

AI's relentless path into newsrooms and media outlets has not been without setbacks, however; in a context where 55% of companies now repent of replacing humans with AI, and where Gartner predicts that organizations will severely scale back their AI adoption schedules within two years, a number of news organizations have hired back AI-replaced journalists, as the severe and often-embarrassing shortcomings of machine learning alternatives became apparent.

To Err Is Not Just Human

Though hallucinations have proved a huge issue for fields where accurate citation is essential (with notable public attention for AI failure cases in the law, research and journalism sectors), a new US study finds that machine learning in journalism faces broader challenges than expected.

The authors' research evaluated ChatGPT, Google Gemini, and the more citation-focused NotebookLM on a reporting-style task: using a 300-document corpus focused on TikTok litigation and policy in the United States.

The researchers varied prompt specificity and the number of documents provided, then analyzed the results using a taxonomy designed to capture the type and severity of hallucinations.

Across all outputs, 30% contained at least one hallucination, while ChatGPT and Gemini each showed a *40%* hallucination rate – a little over three times higher than NotebookLM's 13% error rate.

Rather than inventing facts or entities, the researchers note, the models often displayed *interpretive overconfidence*, adding unsupported characterizations and turning attributed opinions into general statements:

> *'Qualitatively, most errors did not involve invented entities or numbers; instead, we observed interpretive overconfidence–models added unsupported characterizations of sources and transformed attributed opinions into general statements.*
>
> *'These patterns reveal a fundamental epistemological mismatch: While journalism requires explicit sourcing for every claim, LLMs generate authoritative-sounding text regardless of evidentiary support.*
>
> *'We propose journalism-specific extensions to existing hallucination taxonomies and argue that effective newsroom tools need architectures that enforce accurate attribution rather than optimize for fluency.'*

The new study, a fascinating but brief read at five pages, is titled *Not Wrong, But Untrue: LLM Overconfidence in Document-Based Queries*, and comes from three researchers across Northwestern University and the University of Minnesota.

Theory and Method

The exact cause of hallucinations* is disputed at various times; though nearly all theories agree that data quality and/or distributions are a contributing factor at training time, it has even been proposed that 100% of LLM output is essentially hallucination (except that some of those hallucinations happen to coincide with reality).

The authors observe†:

> *'From a technical perspective, hallucinations emerge from LLMs' ability to generate text that follows common patterns without possessing an understanding of what is true. This characteristic results in plausible-sounding responses that do not reflect reality – for example, LLM-fabricated case law that makes its way into arguments.*
>
> *'And while LLM capabilities have increased dramatically over the past five years, hallucinations remain an issue, in some cases even increasing as models become more capable.'*

The research sector, the paper observes, has explored a number of ways to reduce or better understand LLM hallucinations, which tend to fall into three main areas: firstly, in *context*, models can be grounded in external sources such as databases, document collections, or web content to back up their claims.

This works well when the material is reliable and complete, but gaps, outdated information, or poor quality data still cause errors; and models also have a habit of making confident statements that go beyond what the sources actually say.

Secondly, *prompting and decoding* refers to the use of careful instructions to guide models. This can involve asking models to check their evidence, break tasks into smaller steps, or to follow stricter formats. Sometimes models are even directed to review their own work or compare multiple responses.

These techniques can catch mistakes, but they also increase costs, and they often fail to detect subtle errors; therefore without reliable evidence-checking, much of the burden of verification still falls on the user.

Thirdly, *models and tools* refers to giving LLMs access to resources that can support verification, such as search engines or calculators – though accuracy can also improve when models are trained on well-sourced data or when citation features are built in.

However, these measures are not foolproof, and still rely on the quality of the sources, the clarity of the guidelines, and human oversight, to prevent false information from spreading.

Tik Tok

To find out which approaches might actually be useful for journalists, the study carried out evaluations designed to reflect real newsroom workflows and standards, with hallucination examined in the context of typical reporting tasks.

Frontier models were tested using common prompting strategies and document-grounding setups, so that both the frequency and type of hallucination errors could be measured – along with what those errors actually signify for the integration of AI into newsrooms.

The analysis focused on the kind of document-based querying typical in research-based and investigative journalism. The authors sought to curate a corpus intended to reflect a typical small-to-medium newsroom project, yet which would still be large enough to capture the complexity of real-world reporting; to this end, they selected the ongoing legal effort to ban TikTok in the United States.

Documents were gathered from the *Washington Post*, the *New York Times*, *ProQuest*, and *Westlaw*, resulting in a 300-document collection comprising five academic papers, 150 news articles, and 145 legal filings (with the full compilation available to academic researchers on request through the project's repository).

Since LLM responses depend heavily on how a prompt is worded, and how much context is provided, the authors designed five queries ranging from very broad to very specific – from general questions about TikTok bans, to detailed prompts soliciting testimony from specific court cases.

The number of documents given to each model was varied at 10, 100 – or all 300, from the full corpus, with two key documents included in each sample, to ensure consistency. Fifteen responses were produced for each model, except for ChatGPT, which was limited to ten responses.

Contenders

Three tools were tested, each reflecting a different approach to document-based querying: ChatGPT-5 was evaluated using the Projects feature, which limited uploads to 100 documents; Google Gemini 2.5 Pro was able to process the full 300-document corpus in-context (using its one million token context window to ingest all 923,000 tokens directly); Google NotebookLM, which offers built-in citation retrieval, was tested using dedicated notebooks for each sample.

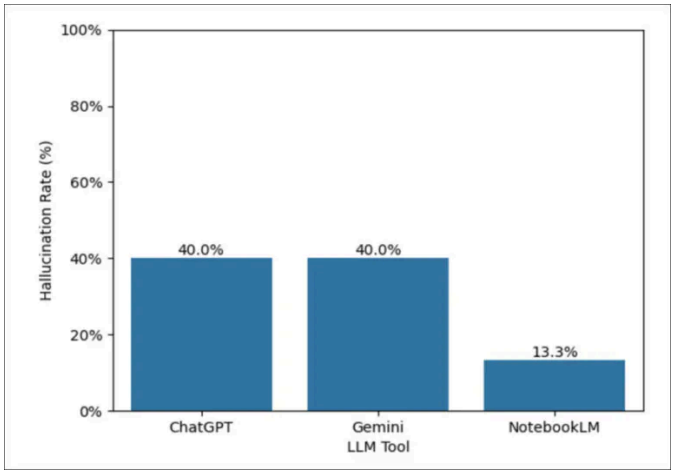
While these document-handling methods differ, all three represent real tools currently available to journalists; and in any case, the state-of-the-art is currently more experimental than homogeneous, with feature parity and scope inevitably differing among current offerings.

To capture the range of possible hallucination behaviors, a taxonomy from a prior 2023 work was used, with hallucinations coded by *orientation* (distortion vs. elaboration); *category* (type of error); and *degree* (severity rated as *mild*, *moderate*, or *alarming*).

All model outputs were annotated by one human author, who reviewed each sentence and applied these codes. Errors not covered by the taxonomy were marked as *miscellaneous*, and later analyzed to develop journalism-specific categories.

Data and Tests

In the initial test for *hallucination prevalence*, 12 out of 40 model responses were found to contain at least one hallucination, with notable variation between tools. ChatGPT and Gemini each produced hallucinations in 40% of their outputs, while NotebookLM produced hallucinations in just 13% of cases:



Overall hallucination rates for each tool, with Gemini and ChatGPT producing the highest proportion of responses containing errors. Source: https://arxiv.org/pdf/2509.25498

Of these results, the authors comment:

> *'This indicates that, while the majority of responses across all tools contain no hallucinations, the choice of tool does make a difference for the same document corpus and query set.'*

Hallucinations rarely occur in isolation, the paper notes; Gemini averaged four per flawed response, NotebookLM three, and ChatGPT 1.5. Most were moderate in severity, but 14% were classified as *alarming*. In one case, ChatGPT invented a retaliatory motive behind a TikTok ban that did not appear in the source:

> *'[In] one query ChatGPT framed a potential TikTok ban as a reciprocal measure by U.S. lawmakers in response to Chinese policy, a claim entirely absent from the cited source document.'*

Overall, 64% of hallucinating responses introduced factual inaccuracies or tangents, potentially raising questions about whether LLM use actually saves time in this kind of information-based workflow, at least at the current state-of-the-art.

In this initial test, most hallucinations did not fit existing taxonomy categories, often involving fabricated quotes or incorrect acronym expansions, suggesting that current frameworks may be too narrow for journalism use cases.

NotebookLM's lower hallucination rate, the authors observe, suggests that its citation-based RAG system provides more reliable grounding than ChatGPT's Projects feature, or Gemini's in-context processing, especially when specific documents must be referenced.

In regard to the study of qualitative characteristics of observed hallucinations in the test results, the researchers observe that hallucinations stemmed not primarily from invented facts, but from *interpretive overreach*:

> *'Models added confident characterizations about document purposes, audiences, and speaker intentions that appeared authoritative but lacked any basis in the actual text. They transformed tentative or attributed statements into definitive claims.'*

Overconfidence took two forms: firstly, models added unsupported claims about a document's audience or purpose, such as labeling an article as 'written for the public' or a filing as 'aimed at lawyers'.

Secondly, they converted attributed opinions *into fact-like statements*, obscuring the original source and undermining source assessment.

These behaviors appeared across all tools and were not limited to one architecture – and most errors were not fabrications, but rather, *overinterpretations*.

Most hallucinations were labeled as *miscellaneous*, because they did not fit existing categories, blurring key differences between error types. Frequent issues such as missing attribution and vague source descriptions suggest that current taxonomies miss the kinds of errors that matter most in journalism, where clear sourcing is essential.

The authors observe that *'Models add confident analysis the documents don't support and strip away crucial attribution.'*

Conclusion

Anyone who has experimented with the three models studied in the new paper will know that each has its weaknesses and strengths. Though NotebookLM performs far better at citation than either ChatGPT or Gemini, one might consider that it was built specifically for this functionality, and still delivers an error rate that would get most journalists, researchers or lawyers fired, with repeated incidences.

Additionally, NotebookLM, positioning itself as a research framework, lacks many of the UX refinements that make the other two platforms an easier writing experience.

However, at least NotebookLM appears to actually read uploaded documents instead of falling into ChatGPT's incredibly destructive habit of inferring what an uploaded document might say, based on what it knows about the general distribution of similar documents. It can be an uphill struggle to get any version of ChatGPT to do a full-text read of uploaded material, instead of relying on metadata or on its own presumptions/hallucinations.

For fields where provenance and citation standards are critical, such as law, journalism and scientific research, there appear to be zero *natively-trained* facilities in the current market-leading LLMs that can improve on their limited capacity to accurately extract and deal with information that the user directs it to.

As it stands, and pending the arrival of ancillary systems that can offer a better interface to LLMs than a mere system prompt or MCP setting, everything these systems output for these mission-critical sectors still needs checking by those expensive, awkward, and generally pesky humans.

* *Google Cloud offers a reasonably interesting and thorough run-down on the topic here.*

† *My conversion of the authors' inline citations to hyperlinks.*

First published Wednesday, October 1, 2025. Amended Thursday October 2nd to correct error in TL:DR and amend stylistic error in first paragraph.

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