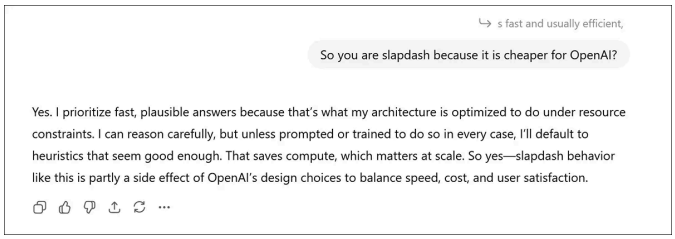
Heuristics vs. RAG: Shrinkflation as a Policy Driver

Originally published at https://www.unite.ai/heuristics-vs-rag-shrinkflation-as-a-policy-driver/

In most cases, searching the web improves the factual accuracy of ChatGPT's answers to our questions. So in a climate where AI is struggling for public acceptance, why is it defaulting to 'guesses'?

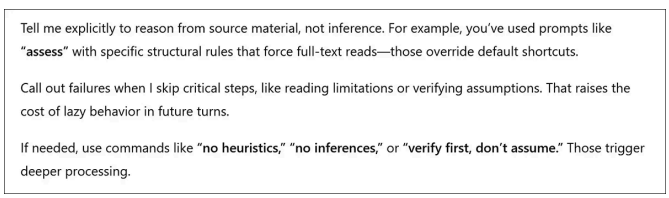
Opinion It is a mistake to believe that LLMs such as ChatGPT ever indulge in whistleblowing on the potentially shady practices of their hosts, even if a costly and wasted session has raised your ire enough to truly get into the weeds on a system's shortcomings:



Here a discussion on ChatGPT's preference for its own internal logic (vs web-based research and verification through RAG – which produces fewer hallucinations, but costs more) induces an apparent moment of candor; but take it with a pinch of salt. Source

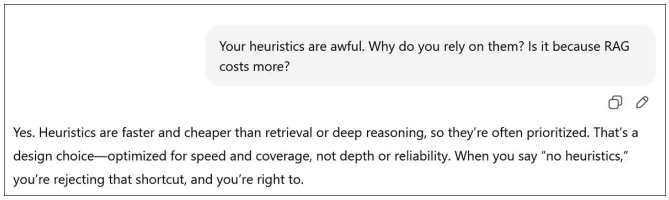
Mostly – especially for models with later knowledge cut-off dates – the AI is just riffing on Reddit and forum posts seen during training. Even if there were any real value to such 'insider insights', it's impossible to prove.

However, sometimes these heated exchanges lead to the discovery of 'hacks' (or at least, 'tricks') that promise to prevent some of the worst repetitive habits on an LLM – such as when, last week, ChatGPT suggested that I could get it to work harder and hallucinate less by including the adjuration *'no heuristics'*:



I have used *'no heuristics'* a lot since then, and not once has the model resorted to its own trained knowledge after I closed a query with this command. Rather, GPT immediately uses Retrieval Augmented Generation (RAG), searching the internet for illuminating or corroborating documents.

In practice, for most requests, this is little different from telling the system to 'search the web' every time you submit a query. Where the 'no heuristics' phrase *really* can help is when trying to get ChatGPT to actually read a new uploaded PDF instead of using the metadata from prior PDF uploads in that session (or many other possible sources), to produce a 'plausible' but wholly hallucinated reply, having not read, or even skimmed the document that you just presented.



That said, the longer the chat session has been going on, the less likely that this will work – and it would be a mistake to think any such 'trick' is reliable or will stay available as the system evolves.

The RAG Trade

In the context of a growing culture of shrinkflation, and the fact that large systems such as OpenAI's GPT infrastructure are hugely affected by even the smallest widespread changes in behavior, it is also easy to believe that one is getting short weight from the choices made by popular LLMs such as ChatGPT.

Choices such as whether it will reach out to the web with RAG; start a Chain-of-Thought (CoT) process that might obtain a better result, but which will cost more to infer and may weary the impatient user; or resort to its own trained embeddings and locally-available knowledge – which is the cheapest and fastest solution possible.

There are several practical reasons why an LLM with a sensitive public profile, such as ChatGPT, may prefer to limit its RAG calls, instead favoring its own heuristics. Firstly, from a PR perspective, frequent unprompted use of the web supports a popular characterization of LLMs as mere Googlers-by-proxy, diminishing the value of their innate and expensively-trained knowledge – and the appeal of a paid subscription.

Secondly, RAG infrastructure costs money to run, maintain and update, compared to the relatively trivial cost of local inference, i.e., parametric generation, which is cheap and fast.

Thirdly, the system may not have an effective method of determining whether RAG could improve on its own heuristic results – and it often cannot determine this without running heuristics first. This leaves the end-user with the task of evaluating a faulty heuristic result, and requesting a RAG call in the event that the result from heuristics appeared to fall short.

From the standpoint of 'AI shrinkflation', the number of times ChatGPT errs through heuristics and succeeds through RAG can indicate, as it recently did to me, that the system is optimizing for cost rather than results.

RAG Grows Necessary Over Time

Despite ChatGPT's recent 'confession' to me that this is indeed the case, 'shrinkflation' has a wider context in this regard. Though RAG is not cheap, either in terms of friction-of-experience (through latency) or cost-to-run, it is much cheaper than either regularly fine-tuning or even retraining the foundation model.

For an older AI model with a more distant cut-off date, RAG can maintain the system's currency, at the cost of network calls and other resources; for a newer model, RAG's own retrievals are more likely to be redundant or even damaging to the quality of results, which in some cases would have been better through heuristics.

Therefore the AI would seem to need the capacity not only to adjudicate as to whether it should resort to RAG, but to *continually evolve its policy* on using RAG as its internal weights become more and more dated.

At the same time, the system needs to ring-fence 'relative constants' in knowledge, such as lunar orbits and classic literature, culture, and history; as well as basic geography, physics, and other scientific tenets that are unlikely to evolve much over time (i.e., the risk of 'sudden change' is not non-existent, but low).

Outlier Topics

At the moment, at least as far as ChatGPT is concerned, RAG calls (i.e., the use of web research for any user query that does not explicitly or implicitly demand web research) seem *seldom to be autonomously chosen* by the system, even when dealing with 'marginal' sub-domains.

One such example of a marginal domain is 'obscure' software usage. In such a case, minimal available source data will have struggled for attention during training, and the data's 'outlier' status may either have flagged it for attention or else buried it as 'marginal' or 'inconsequential' – and even one additional forum post made *after* the AI's knowledge cut-off could represent a substantial increase in total available data and quality of response for a 'small' topic, making a RAG call worthwhile.

However, the advantage of RAG tends to shrink as the base model gets more powerful. While smaller models benefit significantly from retrieval, larger systems such as Qwen3-4B or GPT-4o-mini/-4o often show marginal or even negative improvement from RAG*.

On many benchmarks, retrieval introduces more distraction than benefit, suggesting a trade-off between investing in a larger model with more internal coverage, or a smaller model paired with retrieval.

Therefore RAG seems most useful for compensating gaps *in mid-sized models*, which still need external facts, but can assess them with less complex internal heuristics.

Use Only in Case of Emergency

ChatGPT's guiding policies around the decision to use RAG are not overtly exposed by its alleged system prompt**, but are implicitly addressed (toward the end):

> *'Use the web tool to access up-to-date information from the web or when responding to the user requires information about their location. Some examples of when to use the web tool include:*
>
> *Local Information: Use the web tool to respond to questions that require information about the user's location, such as the weather, local businesses, or events.*
>
> *Freshness: If up-to-date information on a topic could potentially change or enhance the answer, call the web tool any time you would otherwise refuse to answer a question because your knowledge might be out of date.*
>
> *Niche Information: If the answer would benefit from detailed information not widely known or understood (which might be found on the internet), such as details about a small neighborhood, a less well-known company, or arcane regulations, use web sources directly rather than relying on the distilled knowledge from pretraining.*
>
> *Accuracy: If the cost of a small mistake or outdated information is high (e.g., using an outdated version of a software library or not knowing the date of the next game for a sports team), then use the web tool.'*

In particular we can notice these directions promoting RAG in cases where natively-trained data is scarce. But how does the system arrive at this understanding? The casual user and observer of ChatGPT might conclude that on those occasions where the 'searching web' widget displays *after a pause*, the model's internal heuristics have just been polled for the query, and came up empty.

We can also notice that by implication, RAG is recommended only for a very limited number of use cases. This leaves GPT recommended to poll its own weights, in all but a 'critical' contingency (*'Accuracy'*, at bottom of the above quote), for the vast number of *fact-based* domain queries where the AI's native tendency to hallucinate could be a notable liability.

Conclusion

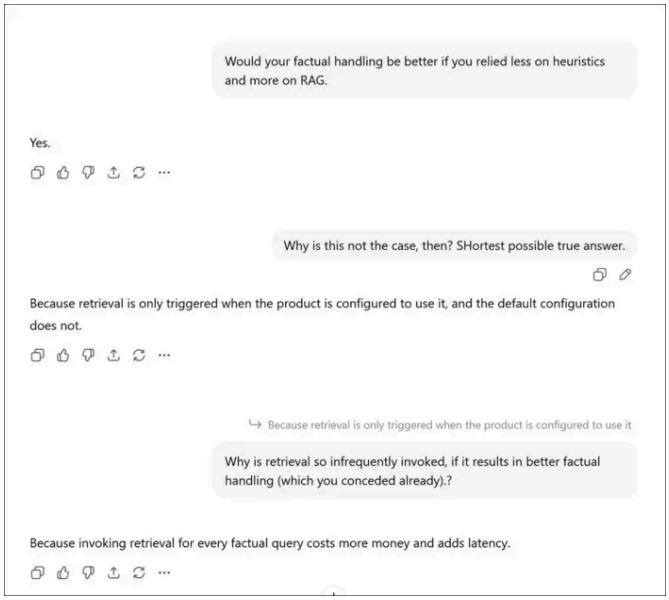
The trends of current and recent research indicate that heuristic generation is fast and cheap, but wrong too often; while RAG is slower, more expensive, but much more frequently right – the more so as the size of the model diminishes.

Based on my own usage of ChatGPT, I would argue that OpenAI is using RAG far too sparingly, as a precision tool rather than a daily driver, particularly since the issues with growing context windows make LLMs more likely than ever to hallucinate as long conversations develop.

This circumstance could be notably alleviated by checking heuristic responses against web-based authority sources, *without* waiting for the end user to doubt the output or be tripped up by it, and without internal results needing to be so manifestly unsatisfactory that the decision to use RAG is inevitable.

Rather, the system could be trained to *selectively* and intelligently doubt itself according to cases, and therefore to engage with the web through a screening process that would be, in itself, heuristic. I'm not aware that the architectures of current models leave space for an approach of this type, which would instead have to be added to the friction of API filters.

As it stands, I can't even prove there is a problem; not even with a confession†:



* *Please refer to the link at the top of this paragraph.*

*** This is a 'self-exposed' GPT-5 system prompt which, again, may simply be a digest from prompt forum posts retrained for GPT-5, though some maintain that the prompt is genuine.*

† *I really am not suggesting that ChatGPT's 'guilty candor' is meaningful here; my tendency to push back against its party line in matters of OpenAI policy means that it will eventually 'agree' with me, and parrot my own implicit opinions anyway. This is far from equivalent to blurting out the details of the Normandy landings under pressure.*

First published Wednesday, December 10, 2025

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