Much Ado About ChatGPT vs. DeepSeek

Sorin Faibish, Life Senior Member, IEEE

A real-world comparison of ChatGPT-4o and DeepSeek-R1 reveals key differences in speed, consistency, and user experience, highlighting tradeoffs shaped more by design than raw performance.

I n response to Michael Zyda's excellent and timely column "Much Ado About DeepSeek…"[7] in *Computer* magazine (vol. 58 no. 5), I felt compelled to contribute a direct, practical comparison between the two artificial intelligence (AI) titans: ChatGPT-4o (subscription)[5] and DeepSeek-R1 (membership)[4,6] actually perform under comparable conditions. An earlier free version of ChatGPT-4 was published in 2024.[1] My contribution here is not to stir panic or fan flames of Cold War 2.0, but rather to add some measured data to the discussion. Instead of emphasizing geopolitical narratives, this article examines the tools from a pragmatic, user-oriented perspective:

How do they behave? How fast are they? And, ultimately, what kinds of tradeoffs do they make?

Rather than speculate, I engaged both models across 25 diverse, real-world workloads sourced from DeepSeek's evaluation framework and my own professional use cases. These included everything from multiple-choice reasoning to summarization, content generation, email parsing, crawling, and domain-specific technical queries.

My approach centered on evaluating the response time to answer (RTTA)—the time it takes from submitting a prompt to receiving a complete answer—under identical network and session conditions. All tests were run manually, but consistently, using equivalent prompts. My intent was simple: compare speed and behavior under the same network conditions. What I found might not crater US$500 billion of market cap—but it might make you think twice about how you interpret performance.

THE EXPERIMENT SETUP

I ran 25 workloads on both models, accounting for total RTTA from prompt to final output. For each workload, I

EDITOR MICHAEL ZYDA
University of Southern California;
zyda@mikezyda.com

COMMENTS?

TABLE 1. RTTA ratios comparing ChatGPT–4o and DeepSeek across workloads.

Tested workload	GPT/DS RTTA ratio
MMLU-Redux	0.95
MMLU-Pro	0.98
MMLU	0.87
if-eval	1.09
GPQA-diamond	0.55
SimpleQA	1.55
FRAMES	1.09
AlpacaEval2.0	0.99
Fluid cooling	0.9
Compare cooling	0.35
Business Deals analysis	1.24
United Kingdom hotels compare	1.25
Insurance needs	0.81
Contact center	0.57
Query GenAI Apps	1.15
Compare MPI solutions	0.74
Technology query	1.09
Email analysis	1.14
Internet crawling for AI	2
Definition of technology	0.39
Write poem	0.85
Poem in foreign language	0.75
Crawling Internet content	0.3
Crawl hacked content	1.38
Write story from loaded documents	0.74
Average RTTA ratio of ChatGPT versus DeepSeek	1.05

RTTA ratio <1 favors ChatGPT; RTTA ratio >1 favors DeepSeek.

calculated the ratio of ChatGPT RTTA versus DeepSeek RTTA—indicating which system responded faster, essentially a normalized speed factor. I assumed correctness of output to be approximately equal (a huge assumption, yes—but this was about speed and user experience, not philosophical completeness). A ratio below 1.0 favors ChatGPT, and above 1.0 favors DeepSeek.

To isolate the interaction and user-experience differences, correctness and completeness of answers were assumed roughly equal, although that is a nontrivial assumption. Interestingly, distinct behavioral patterns emerged:

› ChatGPT-4o begins output generation within 1–2 seconds in most cases, streaming content as it generates it. This leads to a highly responsive and interactive experience.
› DeepSeek, on the other hand, often delays its response initiation for up to 5 seconds, then returns a full block of text all at once.[8] While the final output is often well-structured and comprehensive, the initial latency affects user perception of speed.
› Some of DeepSeek's responses failed to appear unless I patiently waited through the delay. In several extreme cases, the delay was up to 30 seconds, and I mistakenly assumed a crash.

› ChatGPT occasionally varied its output across runs. DeepSeek was more consistent, albeit slower at first touch.

› Consistency vs. interactivity: DeepSeek exhibited consistent results across runs. ChatGPT varied more in style, sometimes rephrasing answers slightly.

THE RESULTS: A NUMERIC SNAPSHOT

Here's the high-level comparison of the test set:

So, averaged out, DeepSeek-R1 edges out ChatGPT-4o in speed by a mere 5% faster on average (Table 1). However, this number masks significant variance across task types. Some tasks showed ChatGPT responding up to 70% faster, while others clearly favored DeepSeek. Some queries (like SimpleQA and Business Deals) clearly favored ChatGPT's interactive responsiveness, while others (like Internet crawling or email analysis) leaned in DeepSeek's favor.

INTERPRETING THE WIDE RTTA VARIANCE

Why does the response time vary so much? The explanation lies in a combination of design strategy, infrastructure, and interaction model.

› *Streaming vs. batched response*: ChatGPT streams output as it generates, resulting in a faster perceived start. DeepSeek appears to preprocess more of the response before initiating the return.

› *Prompt complexity sensitivity*: DeepSeek handles certain multi-hop reasoning tasks more thoroughly, potentially adding processing delay.

› *Infrastructure differences*: ChatGPT-4o operates on Microsoft's Azure stack, highly optimized for low-latency cloud inference. DeepSeek likely uses a mix of NVIDIA and custom-made accelerators, introducing variance depending on load.

› *Caching and optimization*: ChatGPT may leverage internal caching for repeated prompt types, which DeepSeek avoids in favor of consistent generation.

The following diagram illustrates how these factors contribute to the RTTA variance observed.

CHIPS, HARDWARE, INFRASTRUCTURE

One interesting dimension is hardware. DeepSeek claims use of both NVIDIA GPUs,[2] and some form of custom accelerator. However, it's unclear which workloads run on what hardware. ChatGPT, in contrast, is known to run on optimized Microsoft Azure stacks.[3] It's unclear whether hardware differences influenced the wide RTTA variance (ranging from 0.3 to 2x), but it's certainly plausible.

One simple explanation for the large RTTA variance (0.3x to 2x) reflects not just speed differences, but design tradeoffs as well: ChatGPT prioritizes responsiveness, while DeepSeek emphasizes coherence and reasoning completeness—even if that means delaying output. At the time this article was written, there were no complete reasons for the differences in architecture's impact on RTTA. Figure 1 shows the differences between ChatGPT and DeepSeek and the impact of different contributing factors.

THE TAKEAWAY

DeepSeek-R1 may be making headlines, but from a user experience and performance perspective, it's not (yet) a knockout winner. It has different tradeoffs: consistent answers, long initial latency, and a data-centric explanation strategy. ChatGPT-4o offers a faster start and more interactivity, but sometimes varies slightly in response.

The real story here isn't that either is winning, but that both are incredibly capable, and their differences point more to design philosophy than any East versus West narrative.

DESIGN PHILOSOPHY OVER GEOPOLITICS

One might be tempted to read into these differences a geopolitical narrative. However, this evaluation suggests that the contrast is less about region and more about philosophy:

› ChatGPT prioritizes real-time interaction, often at the cost of slightly varied outputs.

› DeepSeek emphasizes deterministic, structured answers, sometimes with delay, but with stability.

Neither approach is superior for all workloads. The best choice depends on whether your use case values speed, consistency, or reasoned depth.
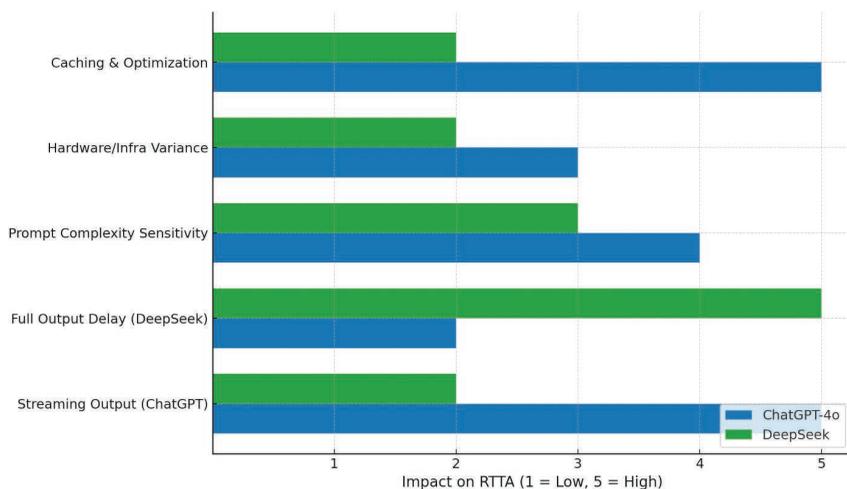

FIGURE 1. Contributors to RTTA Variance Between ChatGPT–4o and DeepSeek

COMMENTARY

If you're looking for instant output, go with ChatGPT. If you want a fully digested answer with reasoning, DeepSeek might be worth the wait.

As a note, I used both solutions to help me write this article based on my directives and edits. While the final product remains human-curated, it reflects a collaborative process between writer and model. One might even say it's a meta-example of the tools being evaluated.

It could be improved, but I am wondering if it is not fancier to use both large language model (LLM) solutions to write the article. In either case, let's hope we keep open source and cross-border collaboration alive, or we'll all just be waiting for our models to catch up while the AI swims past us.

Rather than crown a winner, this analysis aims to highlight how two leading LLMs differ in practical usage. If we avoid oversimplified comparisons and focus on use-case alignment, we all benefit.

Let's hope the future of AI is shaped not by rivalry, but by an ethos of interoperability, experimentation, and transparency. **C**

ACKNOWLEDGMENT

REFERENCES

1. M. Bubeck et al., "Sparks of artificial general intelligence: Early experiments with GPT-4," 2023, *arXiv:2303.12712*.
2. "GPU-accelerated AI inference," NVIDIA, Santa Clara, CA, USA, 2024. Accessed: Jan. 22, 2025. [Online]. Available: https://www.nvidia.com/en-us/lp/ai/inference-whitepaper/
3. "Optimizing LLM latency on Azure." Microsoft. Accessed: Mar. 27, 2025. [Online]. Available: https://azure.microsoft.com/en-us/products/ai
4. "Into the unknown." DeepSeek. Accessed: Jan. 22, 2025. [Online]. Available: https://www.deepseek.com/en
5. "ChatGPT overview." OpenAI. Accessed: Feb. 22, 2025. [Online]. Available: https://openai.com/chatgpt/overview
6. "DeepSeek platform." DeepSeek. Accessed: Jan. 10, 2025. [Online]. Available: https://platform.deepseek.com/
7. M. Zyda, "Much Ado about DeepSeek," *IEEE Comput.*, vol. 58, no. 5, pp. 78–81, May 2025. doi: 10.1109/MC.2025.3541112.

SORIN FAIBISH is a technology consultant in Newton, MA 02461 USA. He is an IEEE Life Senior Member. Contact him at sfaibish@comcast.net.