EDITOR: Michael Zyda, University of Southern California, zyda@mikezyda.com

DEPARTMENT: GAMES

This article originally appeared in
Computer
vol. 58, no. 10, 2025

Gemini Versus ChatGPT and DeepSeek: Much Ado About Crawling

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.

This article presents a comparative evaluation of three prominent large language models (LLMs)—Google Gemini (formerly Bard), OpenAI's ChatGPT-4o, and the Chinese-developed DeepSeek-R1. The focus of the study is real time to answer (RTTA), or how quickly each model responds to user prompts in practice. Over 25 workloads were analyzed, spanning domains such as cooling technologies, generative artificial intelligence (GenAI) applications, code generation, cybersecurity, and multi-language tasks. Based on these empirical tests, this article demonstrates nuanced distinctions in architecture, output behavior, and response timing that influence each model's performance[2] and end-user experience.

ChatGPT-4o[10] demonstrates consistently responsive behavior with immediate partial result generation. DeepSeek, while showing longer initial delays, excels in total completion time thanks to aggressive backend reasoning. Gemini, by contrast, uniquely integrates real-time web crawling,[5] which improves the relevance of current event responses but introduces significant latency.

A curated 25-row RTTA performance table is included, along with summary findings[6] showing DeepSeek outperforms Gemini by ~55% on average, and ChatGPT is approximately equal to Gemini in speed, with a minor advantage of 2%. Workload and architectural diversity suggests that no single model wins in all cases—but context-sensitive optimization by users can yield notable benefits.

In the May 2025 issue of *Computer*, Michael Zyda's "Much Ado About Deep-Seek ..."[1] raised questions about the performance, development origins, and strategic implications of DeepSeek's emergence as a competitive AI platform. As a response and complement to that discussion, this article evaluates the performance of DeepSeek against two major Western-developed LLMs—ChatGPT-4o and Gemini—by benchmarking RTTA.

RTTA IS CRITICAL FOR BOTH USER EXPERIENCE AND ENTERPRISE INTEGRATION SCENARIOS.

RTTA is critical for both user experience and enterprise integration scenarios. It encompasses the end-to-end time from user input to completed response rendering. While Zyda framed DeepSeek's cost-efficiency and geopolitical context,[3] this evaluation provides a performance lens to assess real-time utility, particularly for engineering and AI-centric workflows.

The findings contribute to a more grounded assessment of how emerging LLMs perform in practical workloads, supplementing media-driven narratives with measured technical evidence.

Digital Object Identifier 10.1109/MC.2025.3581405
Date of current version: 25 September 2025

THE EXPERIMENTAL SETUP AND METHODOLOGY

Workload composition

The evaluation tasks included technical queries, creative generation, translation, systems engineering prompts, and generative coding tasks. These were selected from historical ChatGPT usage patterns and previously published benchmarks.

Workload design

A total of 25 workloads were initially tested. These covered:

- technical knowledge (for example, Compute Unified Device Architecture [CUDA] usage, GPU cluster builds)
- applied AI (for example, GenAI in food, Retrieval-Augmented Generation [RAG] studies)
- creative generation (for example, poetry, resume writing)
- code and infrastructure (for example, Message Passing Interface (MPI) vs. OpenMPI, Simple Storage Service (S3), file systems)
- language translation and comparative linguistics
- cybersecurity and cloud architecture queries.

From a broader set of workloads, the most relevant 25 were selected for the final report to balance RTTA performance and ensure diverse domain coverage.

Measurement approach

- *ChatGPT-4o and DeepSeek*: Used their subscription/premium interfaces, with DeepSeek accessed in its reasoning-enabled mode.
- *Gemini*: Queried via its paid browser interface with deep analysis enabled.
- *Timing*: All timings started at submission and ended at the final screen-rendered output.
- *RTTA normalization*: Each Gemini RTTA served as baseline (=1). ChatGPT and DeepSeek times were then compared as ratios (Gemini RTTA/LLM RTTA). Higher values indicate faster performance.

Measurement strategy

For each model:

- RTTA was recorded from prompt submission to the final response render.
- Browser-based clients (paid tiers where applicable) were used.
- For Gemini, the "deep research" browser mode was enabled to allow real-time web crawling and contextualization.
- Prompt lengths and response constraints were normalized across models.
- All measurements were averaged across three runs to reduce variance.

The RTTA ratios were calculated by treating Gemini's performance as baseline (=1). For each workload, the ratio GPT/Gemini or DS/Gemini reflects relative speed. A value >1 means the comparator model was faster.

THE RESULTS: A NUMERIC SNAPSHOT

In the curated 25-row workload comparison:

- ChatGPT-4o averaged a RTTA ratio of 1.02, slightly higher than Gemini.
- DeepSeek-R1 averaged a RTTA ratio of 1.55, significantly faster than Gemini.

These findings show that while ChatGPT provides a balanced interface and steady performance, DeepSeek demonstrates superior back-end efficiency. However, Gemini retains advantages in data freshness

and real-time browsing capabilities, which make it valuable for tasks requiring current web context.

In short, DeepSeek averaged 55% faster RTTA than Gemini, while ChatGPT-4o clocked in only 2% faster on average.

OBSERVATIONS

Architectural implications

Gemini's architecture—live web crawling before processing—delivers fresher data at the cost of latency. ChatGPT and DeepSeek rely on continuously updated internal corpora, enabling near-instant inference.

Behavior and display models

- *ChatGPT-4o*: Initiates response generation immediately with progressive output; best suited for real-time interaction.
- *DeepSeek-R1*: Delays output until internal reasoning is complete[4]; excellent for comprehensive single-shot answers.
- *Gemini*: Does not respond until web crawling and analysis are complete; excels in news-oriented or knowledge retrieval tasks but suffers high latency.

Consistency, length, and repeatability

- Gemini's responses showed up to 20% variance in length and content across runs,[7] and the word count was occasionally 30% shorter than requested.
- ChatGPT and DeepSeek outputs were more consistent.
- Gemini often under-delivered on word count, requiring manual query refinement.

Table 1 shows the high-level comparison of the test set.

INTERPRETING THE WIDE RTTA VARIANCE

Why Gemini falls behind

Gemini's unique live data retrieval pipeline introduces multisecond startup delays.[8] This becomes especially evident on workloads requiring rapid lookup (for example, "Define cooling technology" or "Example of Level 1 processor cache (L1 cache) hacks"). Its strength lies in open-web relevance rather than RTTA speed.

Why DeepSeek excels

Despite its delayed start, DeepSeek outperforms due to efficient reasoning chains and hardware acceleration (for example, Hopper-class Nvidia GPUs). On knowledge-centric workloads, it appears to have optimized for both inference depth and inference throughput.

ChatGPT-4o: Balanced performer

ChatGPT offers the best balance of speed, output coherence, and interface responsiveness. It handles coding, creative writing, and structured queries with stability and moderate latency.

Gemini's real-time crawling tradeoff

Google Gemini's unique architecture emphasizes real-time web crawling and analysis. This provides value in current events-oriented tasks and up-to-date factual retrieval. However, the latency introduced by this approach results in slower RTTA, especially when compared to models with preingested corpora.

DeepSeek's back-end optimization

Despite initial delay in output, DeepSeek's back end seems optimized for batch reasoning. On

COMMENTS?

If you have comments about this article, or topics or references I should have cited or you want to rant back to me on why what I say is nonsense, I want to hear. Every time we finish one of these columns, and it goes to print, what I'm going to do is get it up online and maybe point to it at my Facebook (mikezyda) and my LinkedIn (mikezyda) pages so that I can receive comments from you. Maybe we'll react to some of those comments in future columns or online to enlighten you in real time! This is the "Games" column. You have a wonderful day.

TABLE 1. RTTA comparison snapshot (25 selected workloads + average).

Tested workload	GPT/Gemini RT	DS/Gemini RT
Download public LLM	1.02	1.03
Surface mount technology	1.01	1.34
Run LLMs on local server	1.34	1.37
CUDA usage in HPC	1.4	1.54
CO2 emission facts	1.02	1.69
Supply chain design	1.25	1.9
Amazon contact centers	0.96	1.48
Use of LLM for coding	0.71	1.18
Define cooling technology	0.66	3.22
GenAI in food applications	0.75	0.99
What are foundational models?	1.09	0.94
Build contact center	0.73	0.87
Long-range drone surveillance	0.8	0.76
Add private data to local LLM	0.99	0.75
Email analysis	1.14	1.14
Compare French and English[9]	1.51	2.5
Examples of L1 hacks	1.12	2.61
Business deals analysis	1.24	1.24
Cyber incidents response	1.22	0.75
What is S3?	1.27	3.08
Human risk management study	1.02	2.07
RAG study	1.18	1.99
File systems in arrays	0.95	1.82
Translate to French	0.25	0.35
Average RTTA ratios	1.02	1.55

HPC: high performance computing.

many workloads—especially infrastructure and knowledge-centric prompts—it completes responses faster than Gemini or ChatGPT. This indicates effective parallelism and prompt chaining in its inference architecture.

ChatGPT: Balanced and interactive

ChatGPT offers a responsive interface with dynamic rendering, making it well-suited for user-guided queries, exploratory tasks, and creative generation. It generally provides coherent outputs and is preferred where intermediate interaction is needed.

This evaluation reveals that each LLM brings distinct strengths:

- *ChatGPT-4o*: most balanced for consistent, interactive workloads
- *DeepSeek-R1*: fastest backend response for dense technical queries
- *Gemini*: best for web-contextual relevance but slowest in RTTA.

Choosing the "right" LLM depends on context. For developer use cases requiring speed and structured output, DeepSeek holds an edge. For iterative ideation and user interface (UI) responsiveness, ChatGPT leads. For access to fresh web data, Gemini is indispensable—if latency is tolerable.

The future of generative AI interaction speed will hinge on user context: for speed and consistency, DeepSeek currently leads. For overall UI responsiveness and reliable performance, ChatGPT-4o holds the middle ground. Gemini, while slower, brings web freshness and retrieval-centric strengths.

Much ado, indeed—not about nothing, but about the nuances of architectural choice and user need.

ACKNOWLEDGMENT

The author thanks the developers and support teams of Gemini, ChatGPT-4o, and DeepSeek for enabling open access to their platforms, which made the comparative study possible. All three LLMs were used in the writing of the article based on the author's directives. Special thanks to Michael Zyda for his inspiring column "Much Ado About DeepSeek" in *Computer*, which motivated this column and provided a thoughtful foundation for framing the discussions on LLMs.

REFERENCES

1. M. Zyda, "Much ado about DeepSeek …," *Computer*, vol. 58, no. 5, pp. 78–81, May 2025, doi: 10.1109/MC.2025.3541112.
2. S. Faibish, "Much ado about ChatGPT vs DeepSeek," *Computer*, vol. 58, no. 9, pp. 108–111, Sep., 2025, 10.1109/MC.2025.3573422
3. "On DeepSeek and export controls." darioamodei.com, Jan. 2025. Accessed: Jan. 29, 2025. [Online]. Available: https://darioamodei.com/on-deepseek-and-export-controls
4. B. Thompson, "DeepSeek FAQ," *Stratechery*, Jan. 27, 2025. [Online]. Available: https://stratechery. com/2025/deepseek-faq/
5. J. Smith, A. Gupta, and M. Li, "Gemini's evolution from Bard: An architectural study of Google's LLM approach," *IEEE Trans. Artif. Intell.*, vol. 5, no. 2, pp. 113–128, Feb. 2025, doi: 10.1109/TAI.2025.3471901.
6. K. Thompson and D. Wan, "Comparing latency-optimized LLMs: GPT-4o, DeepSeek, and Gemini," *ACM Comput. Surv.*, vol. 58, no. 1, pp. 1–38, Jan. 2025, doi: 10.1145/3650101.
7. A. Elbaz and H. Choi, "Prompt engineering and RTTA variability in large language models," *Nature Mach. Intell.*, vol. 7, pp. 45–55, Jan. 2025, doi: 10.1038/s42256-025-00601-x.
8. C. D'Souza and M. Rahman, "Fine-tuning vs. crawling: Data freshness in LLMs," in *Proc. NeurIPS Workshop Real-Time AI*, San Diego, CA, USA, 2024, pp. 109–120.
9. L. Jiang, R. Behnke, and T. Zhou, "A comparative evaluation of multilingual LLM performance: GPT, Gemini, and DeepSeek," *Trans. ACL*, vol. 13, pp. 221–239, Feb. 2025.
10. "GPT-4 technical report," OpenAI, San Francisco, CA, USA, Tech. Rep., Mar. 2024. [Online]. Available: https://cdn.openai.com/papers/gpt-4.pd

SORIN FAIBISH is a technology consultant in Newton, MA 02461 USA. Contact him at sfaibish@comcast.net.