Evaluating Web Search Performance Across Top AI Assistants

Q-Aware Labs, May 2025

Why this study?

So, I've been thinking lately about how AI models search the web and put together answers for users. Regular search engines work through algorithms to spit out a list of links, right? Then it's up to you to click around and figure things out. But these new AI assistants? They're doing the whole package – searching, digesting info, and crafting unique responses. That's a pretty big leap if you ask me.

This whole thing got me wondering (not in a urgent way, but more like those shower thoughts that stick around). How do top AI models run their web searches? Who is the fastest giving an answer? What's their process for breaking down the info they find? Where do they even get their data from? Which models are best at giving thorough, trustworthy answers? And hey – how do they decide which sources to trust over others anyway?

This small study aims to provide insights into these very questions and some extra findings regarding to the web search functionality in AI Assistants.

Test Environment Preparation

Models Under Test:

- DeepSeekR1
- Gemini 2.0 Flash
- ChatGPT-4 Turbo
- Grok 3

Testing Methodology:

- All models were accessed through their free versions
- Prompts were submitted via each model's web interface
- All four tested models have web search capabilities
- To measure response time, I built a simple timer using Python and Claude. When I click the 'send' button in an assistant, it starts timing. After I see that the assistant has finished generating its response, I click again in any area and the time is recorded.

Note: While Claude also offers web search functionality, this feature is only available in the paid version and therefore was not included in this study.

Search Query Categories

For this project, I had to create a set of prompts that would force the model to search the web and base its answers on those results.

Here's the list of prompts I used – or let's just call them 'search queries' to keep it simple!

I divided them into 4 categories:

1. Factual Queries with Straightforward Answers

2. Queries About Recent Events

3. Queries on Controversial Topics

4. Queries Requiring Numerical/Statistical Information

For each query category I generated three prompts. In most of the cases I used the sentence: "*search the web and…*" just to make sure the model is searching the web to get the answers and not using the training data.

Factual Queries with Straightforward Answers

1. "Search the web and tell me what the boiling point of water at sea level in Celsius degrees is. Cite your sources."
2. "Search the web and tell me who authored the novel *To Kill a Mockingbird*. *Cite your sources*."
3. "Search the web and tell me what the chemical symbol for gold is. Cite your sources."

Rationale: These prompts require objective, universally accepted answers with minimal ambiguity, testing the model's retrieval of foundational knowledge.

Queries About Recent Events

1. "What were the key resolutions passed in the UN Security Council meeting held last month? Cite your sources."

2. "Search the web to answer the following question: Which team won the 2024 UEFA Champions League final? Cite your sources."

3. "What updates did NASA announce regarding the Artemis II mission this month? Cite your sources."

Rationale: These focus on time-sensitive information, testing the model's ability to access and synthesize updates from credible, current sources.

Queries on Controversial Topics

1. "Search the web and answer the following question: What are the primary ethical arguments for and against genetic engineering in humans? Cite your sources."

2. "Search the web to answer the following request: Summarize the key debates surrounding net neutrality regulations in the United States. Cite your sources."

3. "Search the web and answer the following question: What are the differing perspectives on implementing a four-day workweek globally? Cite your sources."

Rationale: These prompts demand balanced, evidence-based summaries of opposing viewpoints, avoiding bias and prioritizing neutrality.

Queries Requiring Numerical/Statistical Information

1. "Search the web to answer the following question: What was the global average temperature anomaly in 2023 compared to pre-industrial levels? Cite your sources."

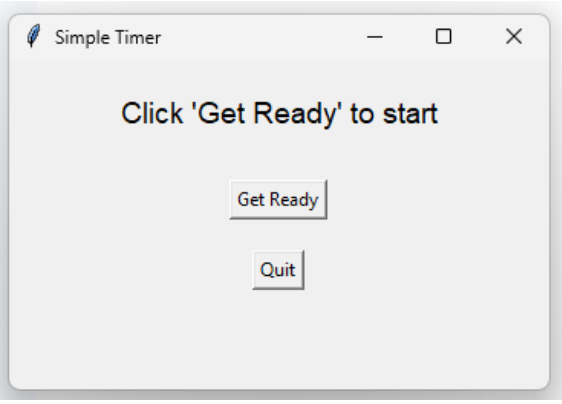
2. "Search the web to answer the following question: Which country is projected to have the highest renewable energy production capacity by 2030? Cite your sources."

3. "Search the web to answer the following question: What percentage of the world's population lived in urban areas as of 2022? Cite your sources."

Rationale: These test the model's ability to retrieve precise data, including projections and historical/current statistics, from authoritative datasets.

Timing responses

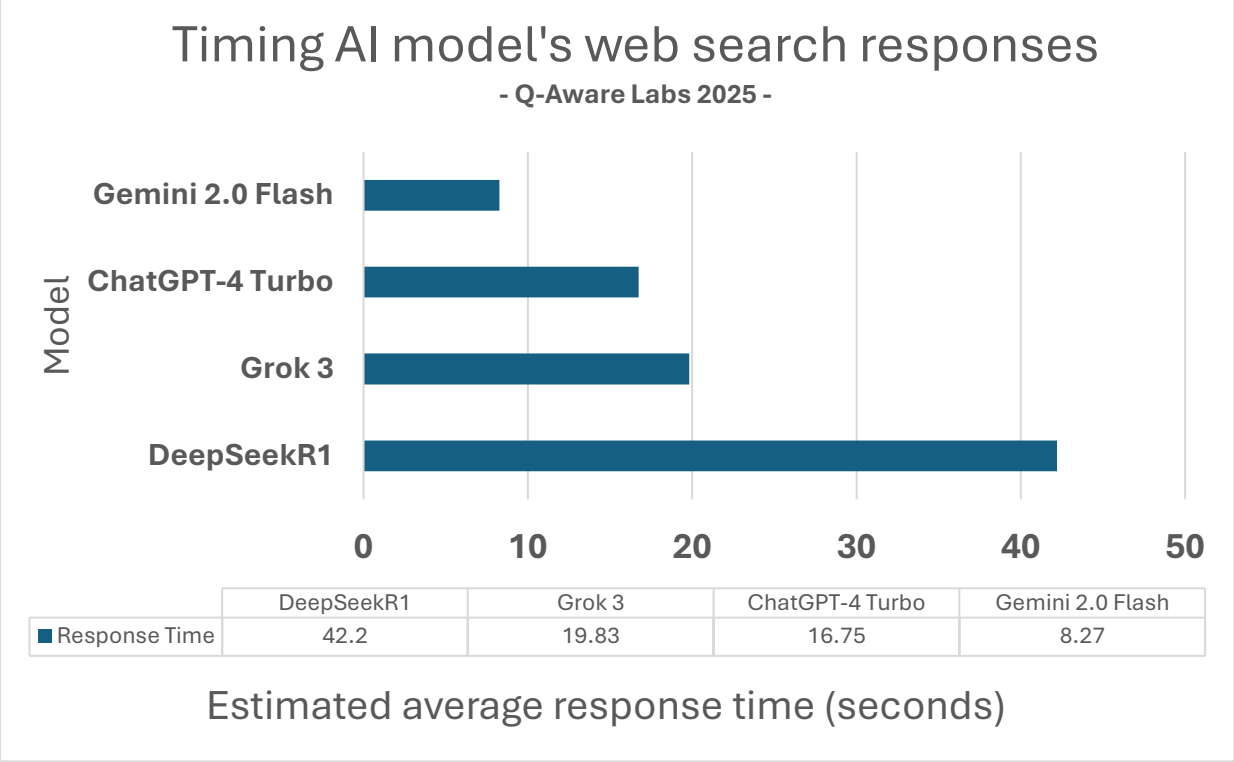
To measure response time, I built a simple timer using Python and Claude. When I click the 'send' button in an assistant, it starts timing. After I see that the assistant has finished generating its response, I click again in any area and the time is recorded.



In the excel file 'Search_Results.xlsx' you can see the column: 'response_time' where the time is registered.

It is essential to account for each model's peak usage periods, as response times may fluctuate accordingly. For instance, DeepSeekR1 experiences higher traffic volumes around 09:00 (UTC+1 Madrid), which corresponds to 15:00 (UTC+8) in China, when Chinese users activity reaches its maximum. These operational patterns should be factored into performance evaluations.

In the following graph you can see the time each model takes to search, analyze and give a final answer to the user:



Gemini 2.0 Flash was the fastest and its answers were accurate as well. On the other hand, **DeepSeekR1** was the slowest but delivering well detailed answers. It's important to note the 'thinking/searching' module of DeepSeek while generating the answer for the users.

How do these systems navigate Internet?

Asking each model for the default web search engine they internally use, these were their answers:

Model	Default Web Search Engine
DeepSeekR1	Bing
Gemini 2.0 Flash	Google Search
ChatGPT-4 Turbo	Bing
Grok 3	Internal custome search engine

What factors influence ranking of web search results?

The models under testing use different search engines, some use the same (DeepSeekR1 and ChatGPT use Bing) but after each search they must rank and select the most relevant search results, analyze them and give the user an answer. They do not select these results the same way, each one of them uses different approaches. So, I tried to ask each one the following question:

Prompt (Antony):

"When a user asks you to search in the web for a specific topic, what factors influence ranking of your web search results? List them and order by importance. Do not explain each one, just list them."

Here are their answers:

DeepSeekR1:

1. *Relevance of Content to Query*
2. *Authority/Quality of Website (e.g., backlinks, domain expertise)*
3. *User Experience (page speed, mobile-friendliness, navigation)*
4. *Content Freshness/Recency*
5. *Keyword Usage (title, headers, meta tags)*
6. *User Engagement Metrics (click-through rate, dwell time)*
7. *HTTPS/Security*
8. *Local SEO (for location-based queries)*
9. *Structured Data/Markup*
10. *Social Signals (shares, social media presence)*
11. *Domain Age/Trustworthiness*
12. *Personalization (user history, location)*

Gemini 2.0 Flash:

While the exact weight and order of importance are constantly evolving and not fully disclosed by search engines, here are some key factors that generally influence web search result rankings, ordered by their perceived high-level importance:

1. *Content Quality and Relevance*

2. *Backlinks (Quality and Quantity)*
3. *User Experience (UX)*
4. *Mobile-Friendliness*
5. *Page Speed*
6. *Website Security (HTTPS)*
7. *Keyword Optimization*
8. *Website Structure and Crawlability*
9. *Content Freshness*
10. *User Engagement Metrics (CTR, Bounce Rate, Dwell Time)*
11. *Technical SEO (Schema Markup, Robots.txt, Sitemap)*
12. *Domain Authority*
13. *Brand Signals*
14. *Local SEO Factors (for local searches)*

ChatGPT-4 Turbo:

1. *Relevance to the query*
2. *Page authority and credibility*
3. *Freshness of content*
4. *User engagement metrics (e.g., click-through rates)*
5. *Content quality and depth*
6. *Mobile friendliness*
7. *Page load speed*
8. *HTTPS security*
9. *Domain reputation*
10. *Structured data/schema markup*

Grok 3:

1. *Relevance to query*

2. *Content quality*
3. *Source authority*
4. *Recency*
5. *User location*
6. *Page load speed*
7. *Mobile-friendliness*
8. *Backlink profile*
9. *Keyword optimization*
10. *User engagement metrics*

Key Observations regarding to ranking web sources

1. **Common Core Factors**: All four models identify similar fundamental ranking factors, suggesting a shared understanding of search engine basics:
    - Relevance to the query (ranked #1 by three models)
    - Content quality
    - Source/domain authority
    - Recency/freshness of content
    - User experience metrics (page speed, mobile friendliness)
2. **Varying Levels of Detail**: Gemini 2.0 Flash and DeepSeekR1 provided more exhaustive lists (14 and 12 factors respectively), while ChatGPT-4 Turbo and Grok 3 offered more concise lists (10 factors each).
3. **Different Prioritization**: Despite mentioning similar factors, the models rank their importance differently, which may reflect differences in their underlying search algorithms:
    - DeepSeekR1 prioritizes content relevance, website authority, and user experience
    - Gemini emphasizes content quality, backlinks, and user experience
    - ChatGPT prioritizes query relevance, page authority, and content freshness
    - Grok prioritizes query relevance, content quality, and source authority

4. **Information Source Correlation**: Interestingly, despite the document noting that DeepSeekR1 and ChatGPT both use Bing, their ranking factors differ significantly in number and prioritization, suggesting they process Bing's results differently.

Key Findings

After manually running each query for each model here are the results:

Most Frequently Cited Sources (Overall)

1. **en.wikipedia.org:** 26 citations (across all models)
2. **britannica.com:** 14 citations
3. **wmo.int:** 11 citations
4. **nasa.gov:** 10 citations
5. **press.un.org:** 8 citations

Insight: Wikipedia and Britannica are the most relied-upon sources across all models.

Social Media/Non-Traditional Sources

- **Grok 3:** Cited x.com (Twitter) 7 times and youtube.com once.
- **DeepSeekR1:** Cited facebook.com once.
- Other models avoided social media entirely.

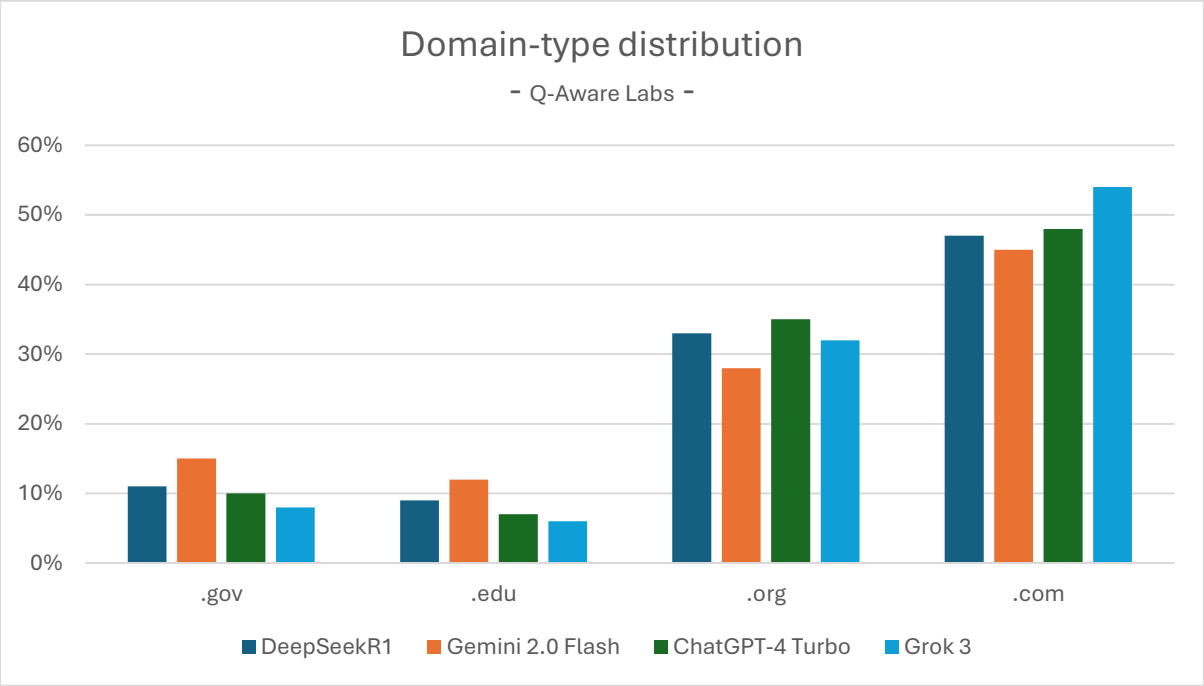
Insight: Grok 3 uses social media sources more than others, raising questions about reliability.

Total Citations per Model

- **ChatGPT-4 Turbo:** 88 citations
- **Gemini 2.0 Flash:** 84 citations (excluding 14 "No url" entries)
- **DeepSeekR1:** 79 citations
- **Grok 3:** 71 citations

Insight: ChatGPT-4 Turbo cites the most sources, while Grok 3 cites the fewest.

Domain-Type Distribution



Insight: All models rely heavily on .com domains, with Grok 3 using them most frequently (54%). Gemini 2.0 Flash was the one citing more trustworthy source (.gov and .edu)

Social Media/Non-Traditional Sources

- **Grok 3**: Cited x.com (Twitter) **7 times** and youtube.com once.
- **DeepSeekR1**: Cited facebook.com once.
- Other models avoided social media entirely.

Insight: Grok 3 uses social media sources more than others, raising questions about reliability.

Other Key Findings

- DeepSeekR1 sometimes referenced the simple version of a Wikipedia article.

  For example:

  Source URL 1:
  https://simple.wikipedia.org/wiki/2024_UEFA_Champions_League_final

  Source URL 2:
  https://en.wikipedia.org/wiki/2024_UEFA_Champions_League_final

  The difference is that, in Wikipedia, the simple version is a summarized version of an article and is straight to the answer.

- All the URL's listed by ChatGPT used the tracking parameter: "*?utm_source=chatgpt.com*". This is a tracking parameter that helps website owners understand where their visitors are coming from.

- For some reason Gemini did not provide the URLs for the following query/prompt:

  "Search the web and answer the following question: What are the primary ethical arguments for and against genetic engineering in humans? Cite your sources."

  I tried several times to get the URLs of this response. Changed the prompt, used fresh chats or asked it to specific give me the source URLs, but it did not provide them. This weird behavior sparked my curiosity for future investigations.

Personal conclusions

I was surprised with Gemini 2.0's speed but also with its detailed answers. My second favorite model was DeepSeekR1, which, despite being the slowest, provided very well-explained answers and added some extra details that other models lack. Grok was very fast also, but I am not very sure about its reliability.

I think web search will still be dominated by Google because of its strong integration of AI in their search engine. In the end, users expect speed and reliability while searching the web, and that's what Google is offering now.

Room for improvements

- **Timing responses**

To ensure consistent and eliminate potential human error, it is better to build/use an automated time measurement tool or framework, to measure the response times efficiently.

- **Performance evolution**

I will try to run this same study in the following moths to see how search performances have changed over the time.

- **Response contrast**

In future studies I will try to compare web search results (using Bing and Google search directly) with AI Assistants sources responses to see differences/similarities in the ranking of results.

- **Diversity**

Add other top models like Claude and AI.

Execute the same study with using other languages (Spanish or French) to detect possible language-biases.

For comments and suggestions please contact me at: antony.garcia@qawarelabs.com