Search Arena: Analyzing Search-Augmented LLMs

Mihran Miroyan*, Tsung-Han (Patrick) Wu*, Logan King, Tianle Li, Jiayi Pan, Xinyan Hu
Wei-Lin Chiang, Anastasios Nikolas Angelopoulos, Trevor Darrell, Narges Norouzi, Joseph E. Gonzalez

Project Page

ICLR International Conference On Learning Representations

Search-Augmented LLMs

Search-augmented LLMs will become a new search interface.

Problem: No large-scale, real-world data from natural user interactions.

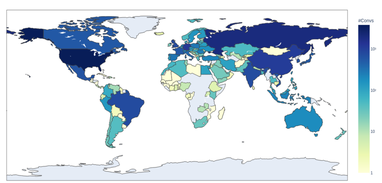
Our Contribution: Large-scale data collection for analyses and training (e.g., Olmo3) of models in search-augmented chat settings.

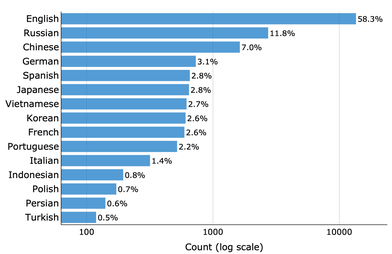
Important Takeaways

1. Users' behavior changes in search-enabled setting from pure LLMs.
2. Users prefer more citations and certain sources but are not sensitive to relevance and quality of the cited sources.
3. Search usually does not hurt performance and even may improve in non-search live chat settings.

Crowd-sourced Data Collection

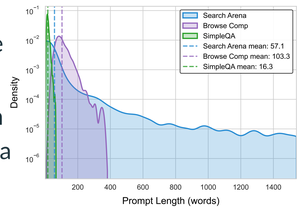
- **24,069** real-world conversations, with **12,652** preference votes.
- **11,650+** unique users across **136** countries.
- **71** languages with **11%** multilingual prompts.
- Metadata (retrieved URLs, reasoning traces, intent/topic tags, .etc).
- All open-sourced on Huggingface.

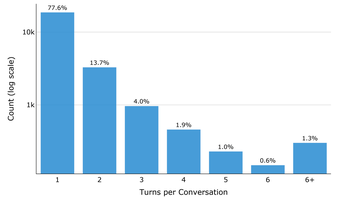


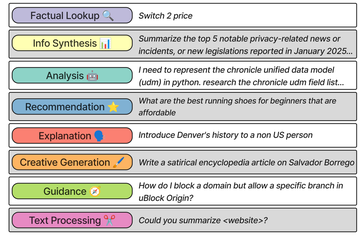


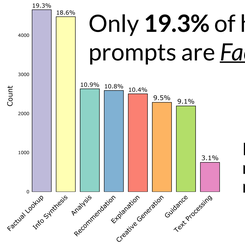
How Do Users Engage with Search-Augmented LLM Systems?

- **22.4%** of Search Arena prompts are multi-turn.
- Our prompt length distribution spans a wide range.







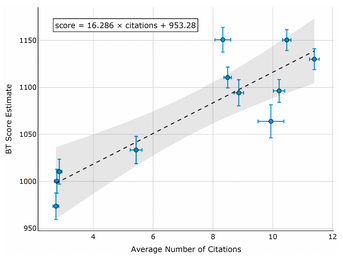


Only **19.3%** of human prompts are *Factual Lookup*.

Prior benchmarks might focus too much on this!

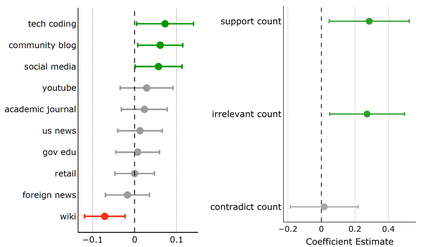
What Are user Preferences in Search Settings?

1. Users prefer more citations.



2. Users prefer certain sources.

(community-driven platforms and social media)



3. Users prefer reasoning models and dislike Wikipedia.

(reasoning models for document filtering and re-ranking)

User Who is the richest person alive?

"Reasoning" Search-Augmented LLM

<think>...The Forbes Real-Time Billionaires list from [1] mentions Elon Musk at the top with $298 billion, followed...
Another source, Investopedia [4], mentions Musk at $316B as of April 1, 2025, but the Forbes India article **[7] had an error** and couldn't be used. The **Wikipedia page [3] has older data**, so it's less relevant...
Forbes 2025 list **[5] from April 5th is likely the most current**...
Therefore, the **consensus** across multiple sources as of April 2025 points to...**</think>**

User Manchester United Transfer news **Note:** Manchester United is a soccer club

The "Loser" Model

...Recent Signings:

- Joshua Zirkzee: On **July 14, 2024**, Manchester United secured the services of Dutch striker Joshua Zirkzee from Bologna for £36.5 million... [1](https://**en.wikipedia.org**/wiki/Joshua_Zirkzee)...

The "Winner" Model

...Manchester United are prioritizing a major overhaul of their attacking options…

- Hugo Ekitike (Eintracht Frankfurt): The 22-year-old French striker has scored 19... [1](https://**www.nbcsports.com**/soccer/news/...) ← **It's in May 2025**

For news related questions, wikipedia data can be outdated.

Should Web Search always Be On?

Question: Would search-enabled models always be better, even on language arena where user don't expect the model would do web search?

Our Experiments: We did a cross-arena deployment and found:

1. Search usually won't hurt and may improve non-search settings
2. Search is required in search settings.