

# 000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 SEARCH ARENA: ANALYZING SEARCH-AUGMENTED LLMs

Anonymous authors

Paper under double-blind review

## ABSTRACT

Search-augmented language models combine web search with Large Language Models (LLMs) to improve response groundedness and freshness. However, analyzing these systems remains challenging: existing datasets are limited in scale and narrow in scope, often constrained to static, single-turn, fact-checking questions. In this work, we introduce **Search Arena**, a crowd-sourced, large-scale, human-preference dataset of over 24,000 paired multi-turn user interactions with search-augmented LLMs. The dataset spans diverse intents and languages, and contains full system traces with around 12,000 human preference votes. Our analysis reveals that user preferences are influenced by the number of citations, even when the cited content does not directly support the attributed claims, uncovering a gap between perceived and actual credibility. Furthermore, user preferences vary across cited sources, revealing that community-driven platforms are generally preferred and static encyclopedic sources are not always appropriate and reliable. To assess performance across different settings, we conduct cross-area analyses by testing search-augmented LLMs in a general-purpose chat environment and conventional LLMs in search-intensive settings. We find that web search does not degrade and may even improve performance in non-search settings; however, the quality in search settings is significantly affected if solely relying on the model’s parametric knowledge. We open-sourced the dataset to support future research.

## 1 INTRODUCTION

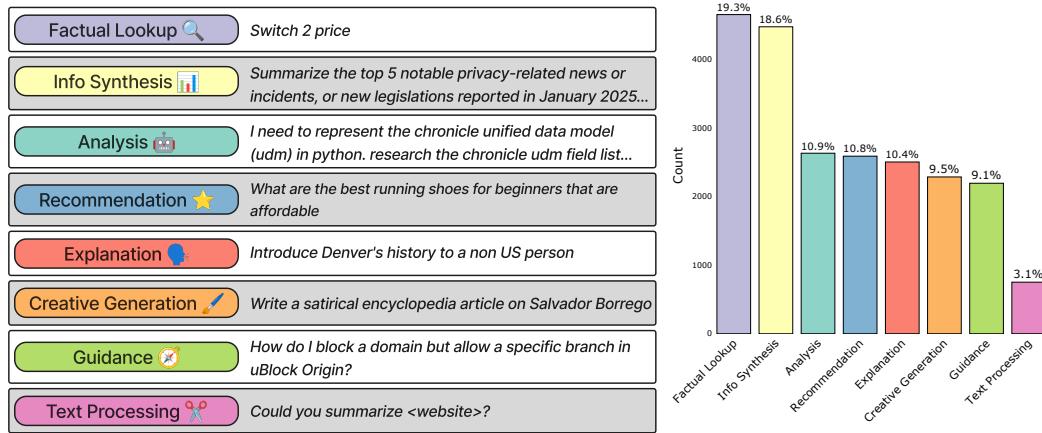
Large Language Models (LLMs) have become a popular interface for human–AI interaction, supporting information seeking and task assistance through natural, multi-turn dialogue. However, the capabilities of these models are constrained by their reliance on static training data, which prevents them from effectively handling time-sensitive questions, emerging topics, or niche domains. Search-augmented LLMs aim to bridge this gap by retrieving and using live web data during inference. Access to search enables LLMs to provide up-to-date, domain-specific, and factually verifiable responses (Hilton et al., 2021; Li et al., 2023a). Recent developments (Gemini, 2024; OpenAI, 2024; Perplexity) also reflect the growing interest in search-augmented LLMs.

Despite rapid progress in developing search-augmented LLMs, our understanding of how users interact with these systems—what they ask, how they engage in multi-turn dialogue, and what they expect in return—remains limited. Existing datasets capture interactions with either standalone LLMs (Chiang et al., 2024; Zhao et al., 2024; Zheng et al., 2024) or traditional web search engines (Chen et al., 2024b; Craswell et al., 2020). However, search-augmented LLMs represent a hybrid interface different from both: they not only retrieve information through web search but also rely on their reasoning and conversational capabilities. The most widely used datasets for evaluating these systems (SimpleQA (Wei et al., 2024) and BrowseComp (Wei et al., 2025)) primarily consist of single-turn, monolingual, fact-based queries and are relatively small in scale (typically  $\leq 5k$  queries). As shown in Figure 1, fact-checking accounts for only one-fifth of real-world user queries; the majority of user prompts, such as seeking analyses, recommendations, or problem-solving guidance, require a combination of factual retrieval, reasoning, and open-ended dialogue. User expectations also extend beyond factual correctness: preferences can be shaped by the number, relevance, and credibility of citations, as well as the presentation style of responses.

To address these gaps, we crowd-sourced the first large-scale human-preference dataset of in-the-wild user interactions with search-augmented LLMs. We developed Search Arena, an open evaluation and data collection platform that presents anonymized side-by-side model outputs in multi-turn settings and collects human votes. During the seven-week deployment period, we gathered and publicly released 24,069 conversations,

054  
055 **Table 1: Comparison of Search Datasets.** Unlike prior datasets such as SimpleQA Wei et al. (2024)  
056 and BroweComp Wei et al. (2025), which are static, English-only, single-turn fact-seeking queries, Search  
057 Arena evaluates models in diverse, open-ended, multilingual, and multi-turn settings. We release 24,069  
058 conversations with 12,652 preference votes. Further analyses are provided in Section 2.

Dataset	#Convs	#Langs	Multiturn	Answer/Judge	Conversation Properties	Metadata
SimpleQA	4,326	1 (EN)	No	Short ground truth	Expert-written short factual queries	Verified supporting URLs, topic tags
BroweComp	1,266	1 (EN)	No	Short ground truth	Expert-written challenging prompts with detailed constraints	Topic tags
<b>Search Arena</b>	24,069	71	Yes	Human preference	Open-ended, crowd-sourced prompts across diverse intents and topics	Retrieved URLs, full model traces, user intent, and topic tags



080 Figure 1: **(Left)** Nine intent categories with representative examples (truncated). In-the-wild prompts  
081 are often ambiguous and require real-time web retrieval. **(Right)** Intent distribution across prompts. Most  
082 queries require more than a simple fact lookup, ranging from information synthesis to creative generation.  
083 The *Other* category is excluded.

084 along with 12,652 paired preference judgments. The dataset spans 11,650 users across 136 countries, 13  
085 models, around 70 languages (including 11% multilingual prompts), and over 5,000 multi-turn interactions.  
086 We also introduce a user intent taxonomy in the context of search-enabled human-AI interactions. As de-  
087 tailed in Table 1 and Section 2, Search Arena provides a broad coverage across linguistic and intent features.

088 We not only analyze user prompts to search-augmented LLMs, but also their preferences. We model user  
089 preferences through the Bradley–Terry model (Bradley & Terry, 1952; Chiang et al., 2024; Tianle Li, 2024)  
090 and study how different response characteristics interact with user judgments. We find that reasoning,  
091 a larger search context window, and longer responses are positively associated with user preferences. Since  
092 citations are central to trustworthy web-grounded outputs, we also examine citation features. Our results  
093 show that users prefer responses with more cited sources (Figure 4) and those citing tech-related platforms,  
094 community blogs, and social networks, but less Wikipedia (Figure 6). While correctly attributed citations  
095 positively interact with preferences ( $\beta = 0.285$ ), we observe a positive association with irrelevant citations  
096 ( $\beta = 0.273$ ). This raises concerns that users may be overly influenced by citation presence, even when  
097 they do not support the associated claims.

098 We also investigate how search-augmented and non-search models perform across different settings by  
099 deploying a non-search LLM in the Search Arena, and a search-augmented LLM in a general-purpose  
100 Text Arena. Our results show that conventional LLMs underperform in search-intensive settings  
101 ( $p\text{-value} = 0.009$ ). However, search-augmented models perform comparably in general chat settings, with  
102 improved performance on factual lookup queries ( $p\text{-value} = 0.012$ ) and slightly degraded performance  
103 on text processing prompts ( $p\text{-value} = 0.077$ ).

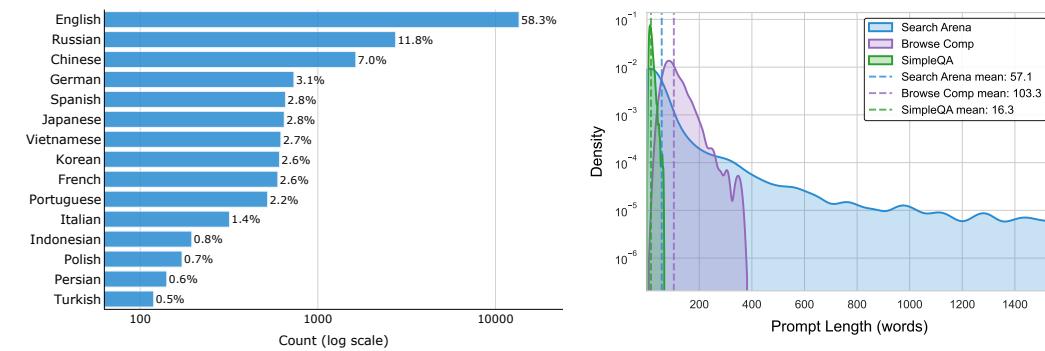
104 In summary, our contributions are as follows: **(1)** we release the first large-scale human-preference  
105 dataset of 24k user conversations with search-augmented LLMs, along with 12k preference votes, system  
106 metadata, user intents, and prompt topics; **(2)** we present the first analysis of how different characteristics  
107 of search-augmented LLMs interact with human preferences; and **(3)** we conduct the first cross-area  
108 evaluation by testing a non-search model in Search Arena and a search-augmented model in Text Arena,

108 reporting that web search augmentation does not hurt and may improve performance across settings, while  
 109 the internal parametric knowledge of models alone is not sufficient in search-intensive settings.  
 110

## 111 2 HUMAN PREFERENCE DATASET IN SEARCH

114 We launched Search Arena, an open, crowd-sourced evaluation platform for search-augmented LLMs,  
 115 on March 18, 2025. The search mode interface encourages more search-intensive queries, as users adjust  
 116 their expectations. During each session, two anonymous models respond to a user query, the user can  
 117 then cast a vote for their preferred model response. Details on user interface and potential limitations are  
 118 reported in [Appendix A](#) and [Section 5](#).

119 Between March 18 and May 8, we collected more than 24,000 conversations and 12,000 user votes  
 120 across 13 models, spanning a range of model configurations (e.g., reasoning models, search context sizes,  
 121 etc.). The collected dataset contains model identities, the user vote, conversation histories, and system  
 122 metadata (e.g., reasoning traces, retrieved URLs). [Table 1](#) presents key differences between Search Arena  
 123 and prior datasets (SimpleQA ([Wei et al., 2024](#)) and BrowseComp ([Wei et al., 2025](#))), including dataset  
 124 scale, prompt characteristics, and available metadata. In the following subsections, we analyze prompt  
 125 distributions across linguistic and intent dimensions in comparison to existing benchmark datasets.



137 Figure 2: **(Left)** Search Arena prompt language distribution. The dataset is multilingual, spanning over 70  
 138 languages, with English prompts accounting for 58.3% of the data. **(Right)** Prompt length distribution of  
 139 Search Arena (blue), BrowseComp (purple), and SimpleQA (green). Search Arena prompt lengths are more  
 140 spread out and cover the range of BrowseComp [Wei et al. \(2025\)](#) and SimpleQA [Wei et al. \(2024\)](#) questions.

### 142 2.1 LINGUISTIC AND CONVERSATIONAL DIVERSITY

144 Search Arena was collected from 11,650 users across 136 countries, resulting in substantial linguistic and  
 145 geographic diversity. The prompts span over 70 languages, with 30 represented by at least 10 conversations.  
 146 English accounts for 58.3% of the data, followed by Russian (11.8%) and Chinese (7.0%). More than  
 147 11% of the prompts are multilingual. [Figure 2](#) (Left) shows Search Arena prompt language distribution  
 148 across the top 15 languages.

149 Since the platform supports multi-turn interactions, 22.4% of conversations in the dataset are multi-turn,  
 150 typically clarifications or follow-up queries. Furthermore, as shown in [Figure 2](#) (Right), SimpleQA  
 151 prompts are short and fact-oriented (16.3 words on average), while BrowseComp prompts are intentionally  
 152 constructed to be long and constraint-heavy (103.3 words on average). In contrast, Search Arena includes  
 153 both brief, under-specified queries as well as longer, detailed requests (57.1 words on average). Additional  
 154 details and analysis of linguistic features are provided in [Appendix B](#).

### 155 2.2 INTENT DIVERSITY

157 Existing search-augmented LLM evaluation datasets focus solely on factuality. To study how in-the-wild  
 158 user prompts from Search Arena differ from SimpleQA ([Wei et al., 2024](#)) and BrowseComp ([Wei et al., 2025](#))  
 159 questions, we apply an LLM-based dataset differencing framework ([Dunlap et al., 2024](#); [Zhong  
 160 et al., 2022](#)). We also compare Search Arena prompts with the prompt distribution of Text Arena ([Chiang  
 161 et al., 2024](#)), where users’ expectations of the models are not influenced by search settings. GPT-4.1 is  
 used for generation and GPT-4.1-mini for ranking of candidate distinguishing properties; more details

162 on the pipeline are provided in [Appendix C](#). The summaries of the top properties reveal high-level pairwise  
 163 differences between the datasets:  
 164

- 165 • *Search Arena vs SimpleQA*: Search Arena prompts are broader and more complex, often requiring  
 166 analysis, synthesis, or creative generation, while SimpleQA prompts are narrowly focused on  
 167 retrieving specific, factual information with minimal context or interpretation.
- 168 • *Search Arena vs BrowseComp*: BrowseComp prompts are structured as investigative challenges,  
 169 often requiring synthesis of clues under specific constraints, whereas Search Arena prompts  
 170 prioritize immediate, functional assistance.
- 171 • *Search Arena vs Text Arena*: Search Arena prompts focus on real-world factual lookup and  
 172 decision-making support, often involving up-to-date information. In contrast, Text Arena prompts  
 173 focus on problem-solving, programming help, and creative generation.  
 174

175 These summaries show high-level differences between prompt distributions across three settings—user  
 176 prompts to search-augmented LLMs (Search Arena), static factuality questions (SimpleQA and  
 177 BrowseComp), and prompts to regular LLMs (Text Arena). We also study Search Arena prompts through  
 178 topic modeling and observe diverse and real-time topics, ranging from market analysis to health discussions  
 (see [Figure B.4](#)).  
 179

180 For a structured and in-depth analysis, we introduce a taxonomy of user intent categories. While prior  
 181 work has explored user intent classification in general dialogue settings ([Liu et al., 2024b; Shah et al.,  
 182 2023](#)), we focus on real-world interactions in search-augmented chat-based settings. The taxonomy  
 183 includes nine categories: *Factual Lookup, Information Synthesis, Analysis, Recommendation, Explanation,*  
 184 *Creative Generation, Guidance, Text Processing*, and *Other*. We use secondary labels for ambiguous or  
 185 multi-purpose prompts. Details on the taxonomy design, as well as descriptions of categories, are provided  
 186 in [Appendix B](#). We scale the annotation to the full dataset using GPT-4.1, with a manually tuned prompt  
 187 seeded on 100 examples and validated on 150 multilingual prompts. The resulting Cohen’s kappa of 0.812  
 188 indicates strong agreement between model- and human-annotated labels. The annotation pipeline and  
 189 validation details are provided in [Appendix B](#).

190 [Figure 1](#) shows the resulting intent distribution, along with examples for each category. Factual lookup  
 191 queries account for only 19.3% of user prompts. The remaining prompts require higher-order capabilities,  
 192 such as synthesis, guidance, or analysis. We also analyze how linguistic features vary across intents;  
 193 specifically, we find that factual lookup prompts are typically shorter (17.2 words on average), whereas the  
 194 remaining set is associated with longer, more complex queries (66.7 words on average). Further analyses  
 195 of intent categories are provided in [Appendix B](#).  
 196

197 We report comparative analysis with additional related datasets (CORAL, WildChat) in [Appendix G](#).  
 198

### 199 3 PREFERENCE ANALYSES IN SEARCH

200 With over 12,000 anonymized human preference votes, Search Arena supports fine-grained analysis of  
 201 how different response features interact with user preferences. We report model performance through both  
 202 head-to-head win rates as well as scaled coefficients estimated using the Bradley-Terry model ([Bradley  
 203 & Terry, 1952; Chiang et al., 2024](#)) in [Appendix D](#). To analyze how features interact with user preferences,  
 204 we follow prior work ([Tianle Li, 2024](#)) by adding normalized differences between pairwise features to  
 205 the Bradley-Terry model and reporting the fitted coefficients.

206 As the correlational analysis in this section relies on user-provided votes, we first clarify their subjective  
 207 nature and assess their reliability. We do not treat the human-preferred response as the objectively “correct”  
 208 one, but rather as the response better aligned with the user’s personal preferences. In our evaluation and data  
 209 collection setup, the prompt is tied to the judge (i.e., user): this pairing is critical, as only the original user  
 210 knows the hidden intent behind their prompt. To verify that users are not voting randomly, we conducted a  
 211 small study on 100 samples with 3 expert annotators and found expert–user agreement to be 68% when  
 212 excluding tie votes (random agreement = 50%). Given the inherent subjectivity and the lack of a single  
 213 “correct” answer for many prompts, some disagreement is expected.

214 In the following section, our analysis is purely correlational rather than causal, and we view more controlled  
 215 causal investigation as an important direction for future work. We first focus on two key feature groups: (1)  
 216 general features such as model type, search depth, and response length ([Subsection 3.1](#)), and (2) citation-  
 217 related features, such as number of citations, citation sources, and citation attributions ([Subsection 3.2](#)). We

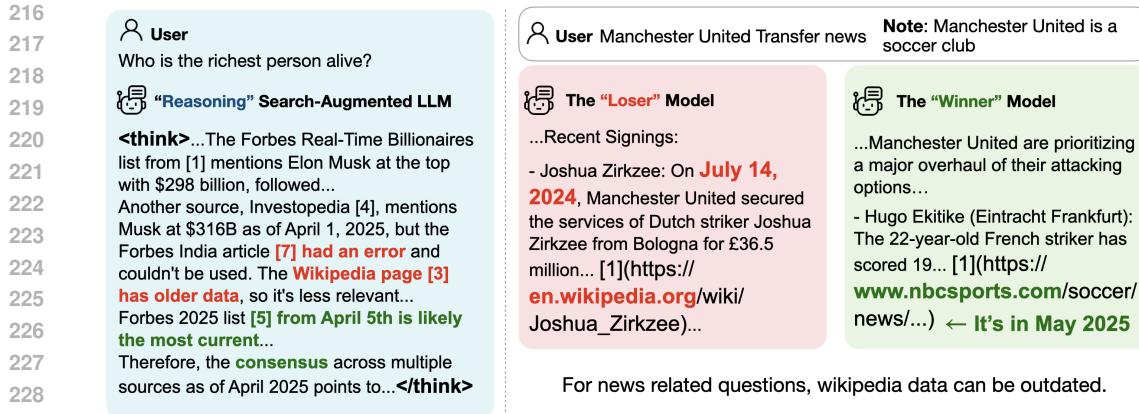


Figure 3: **(Left)** Reasoning trace example, containing multi-document analysis, filtering, and synthesis. **(Right)** Example of a rejected response citing Wikipedia for a sports news question. The preferred response cited the sports division of a news outlet, containing more up-to-date information.

then study how search and non-search models generalize in search-heavy and general chat scenarios in [Subsection 3.3](#). Additional details and further analyses are provided in [Appendix D](#) and [Appendix E](#).

### 3.1 GENERAL FEATURES

**Reasoning.** Reasoning models generally perform better in Search Arena prompt distribution, with the top three models achieving over 60.0% average win rates, suggesting that reasoning improves performance in search-augmented chat interactions. Consistent with prior work (Gandhi et al., 2025), we observe prompt analysis, problem decomposition, and backtracking behaviors in reasoning traces. Additionally, we find that reasoning models not only interpret and analyze retrieved content but also rerank sources and filter out irrelevant information. One example is shown [Figure 3](#) (Left).

**Search Context Size.** Models with high search context windows outperform those with smaller search context. For sonar-pro, the version with high search context has a higher ( $p < 0.01$ ) average win rate (63.9%) compared to the model with medium search context (57.6%). However, the difference is not significant for GPT-4o models with medium and high search context sizes. This finding indicates that models with higher search context retrieve more web sources, leading to more preferred responses; we analyze the effect of citations in [Subsection 3.2](#).

**Response Length.** Consistent with findings from prior work (Chiang et al., 2024; Steyvers et al., 2024; Tianle Li, 2024), we observe that users are biased towards more verbose responses. The Bradley-Terry coefficient corresponding to response length is positive and statistically significant ( $\beta_{\text{length}} = 0.334$ ), indicating that users tend to prefer longer answers. The positive correlation between model score and response length is also shown in [Figure 4](#) (Left). Additionally, [Figure 5](#) (Left) shows the length distribution across eight user intent categories; responses to *Factual Lookup* prompts are much shorter (168.3 words on average) compared to *Creative Generation* (422.8 words) and *Analysis* (393.2 words) prompts. Furthermore, we find that the response length coefficient on *Factual Lookup* prompts ( $\beta_{\text{length, factual}} = 0.156$ ) is 2.14 times smaller than the effect on the full dataset, suggesting that users prefer less verbose answers to *Factual Lookup* queries compared to other categories.

### 3.2 CITATION FEATURES

Citations are central to the trustworthiness of web-grounded outputs in search-enabled scenarios and a unique feature of Search Arena compared to Text Arena (Chiang et al., 2024). Therefore, we further examine how these factors influence user preference along three dimensions: number of cited sources, types of cited sources, and attribution of inline citations.

**Number of Cited Sources.** We find a positive and statistically significant coefficient for the number of citations ( $\beta_{\text{citations}} = 0.209$ ), indicating that users favor responses with more references. The positive association between model score and response length is shown in [Figure 4](#) (Right). Furthermore, we observe that reasoning models cite fewer sources than non-reasoning models with similar configurations, consistent with our earlier observation that reasoning models filter irrelevant content. Unsurprisingly, models with high search context size end up citing more sources in their final response. Additionally, [Figure 5](#) (Right) shows

270 the distribution of citation counts across prompt intent categories. Notably, responses to *Factual Lookup*  
 271 prompts contain fewer citations (5.7 on average), compared to *Recommendation* (6.9 citations on average)  
 272 and *Info Synthesis* (6.8 citations on average) prompts, due to the broader web coverage needed for the latter.  
 273

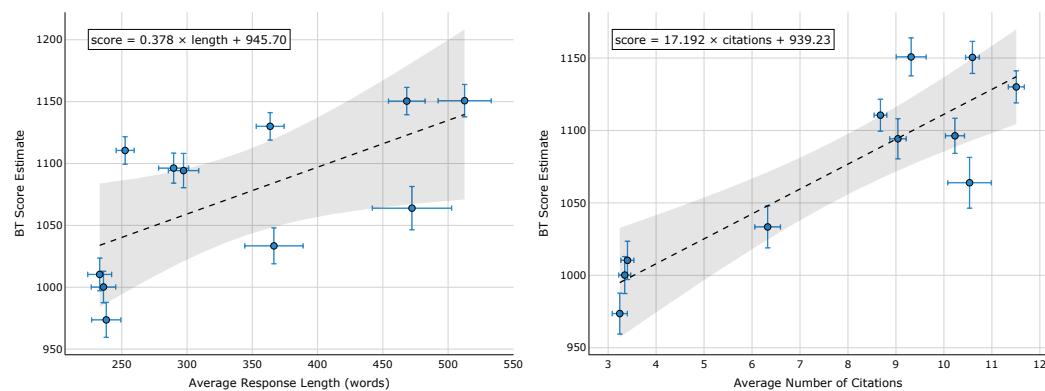
274 **Types of Cited Sources.** We categorize retrieved URL domains into nine groups (e.g., news, Wikipedia,  
 275 social media, tech/code platforms, etc.); categorization details appear in [Appendix E](#). [Figure 6](#) (Left) shows  
 276 source category coefficient estimates with 95% confidence intervals; we also control for citation counts  
 277 to account for the bias shown earlier. Citing tech-related platforms (e.g., Stack Overflow), community  
 278 platforms (e.g., Substack), and social media (e.g., TikTok) are positively associated with user preferences  
 279 with fitted coefficients equal to  $\beta_{\text{tech}} = 0.073$ ,  $\beta_{\text{community}} = 0.061$ , and  $\beta_{\text{social}} = 0.057$ , respectively. Surprisingly,  
 280 citing Wikipedia is negatively correlated with user preferences ( $\beta_{\text{wiki}} = -0.071$ ). To interpret the latter result,  
 281 we inspect rejected model responses citing Wikipedia and identify two potential explanations: (1) Wikipedia  
 282 articles are often very lengthy and broad, not directly relevant to a user’s question, and (2) citing Wikipedia is  
 283 not preferred for queries requiring real-time information. A qualitative example appears in [Figure 3](#) (Right).

284 **Citation Attribution.** We then study how the correctness of citation-to-claim attribution (i.e., whether  
 285 the inline citation supports the attributed claim) interacts with user preferences. Formally, for each  
 286 multi-turn interaction, we decompose model responses into a set of claim-citation pairs  $(c_i, u_i)$ , where  
 287  $c_i$  denotes a textual claim, and  $u_i$  is the corresponding inline citation. For each pair, we evaluate whether  
 288 the webpage content  $D_i$  supports, is irrelevant to, or contradicts the claim  $c_i$ . This process is automated  
 289 via an LLM-based pipeline described in [Figure 7](#) through an example. Due to scraping challenges and  
 290 the high cost of LLM calls, we run the pipeline on roughly 100 conversations per intent category (800  
 291 examples in total). The resulting output of each conversation is a set of triplet  $\{(c_i, u_i, t_i)\}_{i=1}^N$ , where  
 292  $t_i \in \{\text{Support, Irrelevant, Contradict}\}$  and  $N$  is the total number of claims per conversation. We then  
 293 compute the number of supporting, irrelevant, and contradicting claims per model response and add them  
 294 as control covariates in the Bradley–Terry analysis. Implementation details, including scraping tools,  
 295 parsing logic, and validation process, are provided in [Appendix E](#).

296 [Figure 6](#) (Right) shows bootstrapped coefficient estimates for the number of supporting, irrelevant,  
 297 and contradicting claim-citation pairs. While users tend to favor responses with more citations, as shown in  
 298 [Figure 6](#) (Left), the number of contradicting claim-citation pairs does not show a significant effect on  
 299 user preference. Furthermore, both supporting ( $\beta_{\text{support}} = 0.29$ ) and irrelevant ( $\beta_{\text{irrelevant}} = 0.27$ ) claims are  
 300 positively correlated with user preference. Thus, users do not distinguish between supporting and irrelevant  
 301 citations and generally prefer more citations, even if the citations do not directly support the claims. Upon  
 302 inspection, in irrelevant citation cases, models may fabricate connections, cite tangentially related sources,  
 303 or present inferred claims that subtly deviate from the source content. This finding suggests that users  
 304 may be influenced by the mere presence of citations, rather than their proper attribution to generated claims.  
 305 We raise this as an open issue for the community: improving citation attribution is critical to ensure that  
 306 citation-heavy responses are not misperceived as factual and trustworthy.

### 3.3 CROSS-SETTING ANALYSIS

307 Search Arena evaluates models where user prompts and expectations are conditioned on models’ access  
 308 to web search. In this section, we study how search and non-search models perform under various prompt  
 309



322 **Figure 4: (Left)** Positive relationship between model score and average response length. **(Right)** Positive  
 323 relationship between model score and average number of citations.

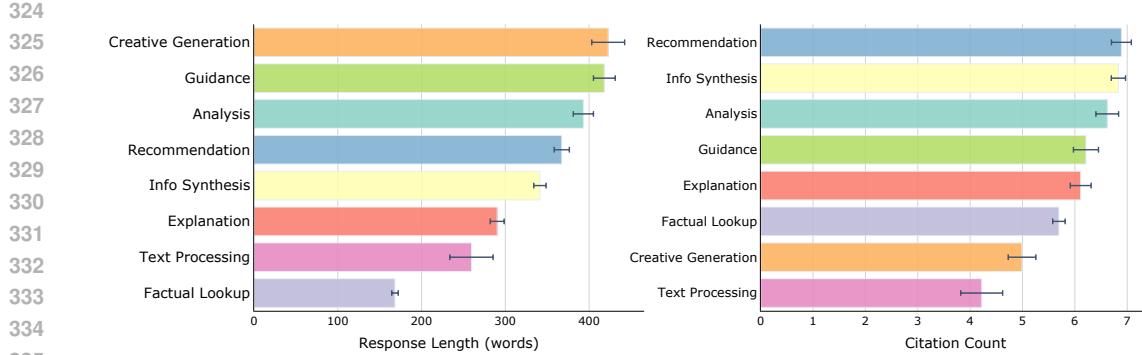


Figure 5: **(Left)** Response length distribution across user intent categories. Responses to *Factual Lookup* prompts are more concise (168.3 words on average) compared to other categories. **(Right)** Citation count distribution across user intent categories. Responses to *Recommendation* (6.9 on average) and *Info Synthesis* (6.8 on average) prompts contain more citations compared to *Factual Lookup* (5.7) and *Text Processing* (4.2) prompts.

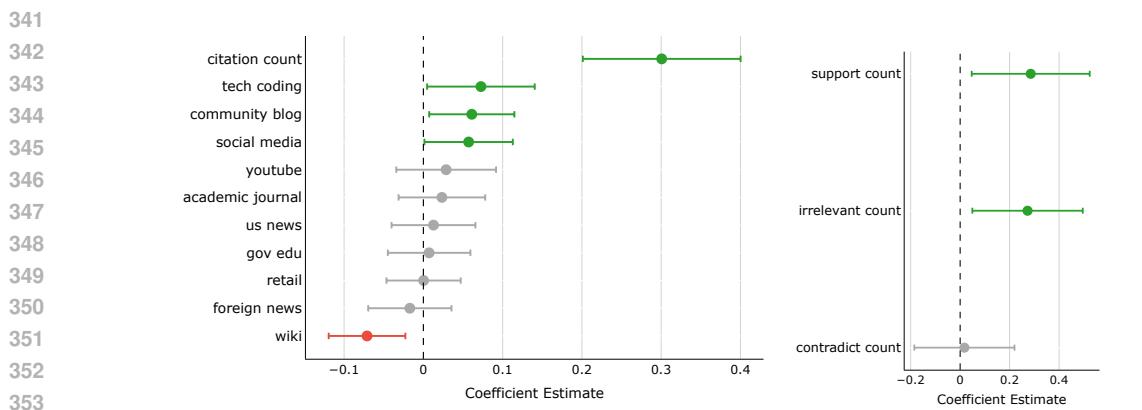


Figure 6: **Citation Control.** **(Left)** Bootstrapped coefficient estimates of citation features. (1) Users prefer responses with more citations. (2) Citing tech-related and community platforms, as well as social media is positively associated with user preferences. (3) Citing Wikipedia negatively interacts with user preferences. **(Right)** Bootstrapped coefficient estimates of the number of supporting, irrelevant, and contradicting claim-citation pairs. The number of supporting and irrelevant pairs is positively correlated, while the effect of contradicting pairs is not significant.

distributions and user expectations—specifically, Search Arena vs Text Arena. Additional results on model performance across benchmarks are provided in [Appendix F](#).

To investigate performance differences across settings, we deployed Gemini-2.5 Pro Experimental (Gemini, 2025)—with and without access to web search—to the Search and Text Arenas. In the Text Arena, the search model only competed against its non-search version; in the Search Arena, the non-search model competed against all other supported search models. Additionally, inline citations were disabled for the search model in the Text Arena to avoid vote bias. In the Text Arena, users typically assume that models operate in closed-book settings without access to external information, while in the Search Arena, user expectations are explicitly conditioned on the search setting. This setup enables us to examine whether and under what conditions access to web search enhances or degrades model performance.

**Text Arena Setting.** We collected 544 battles between the search and non-search versions of the model, yielding 245 ties (45%), 143 search-preferred (26%), and 156 non-search preferred votes (28%). We observe a high proportion of tie votes, and the difference between search-preferred and non-search-preferred votes is not statistically significant ( $p\text{-value} = 0.244$ ). On aggregate, search and non-search models have comparable performance in the Text Arena. To further analyze performance differences by prompt type, we apply the intent classification pipeline from [Subsection 2.2](#) on the 544 collected Text Arena prompts. The distribution of votes by intent class is shown in [Figure 8](#) (Left). We observe a high proportion of ties in *Analysis*, *Creative Generation*, and *Guidance* queries, indicating that there are no significant differences between model responses for these types of prompts. However, for *Factual Lookup* ( $p\text{-value} = 0.012$ ) and

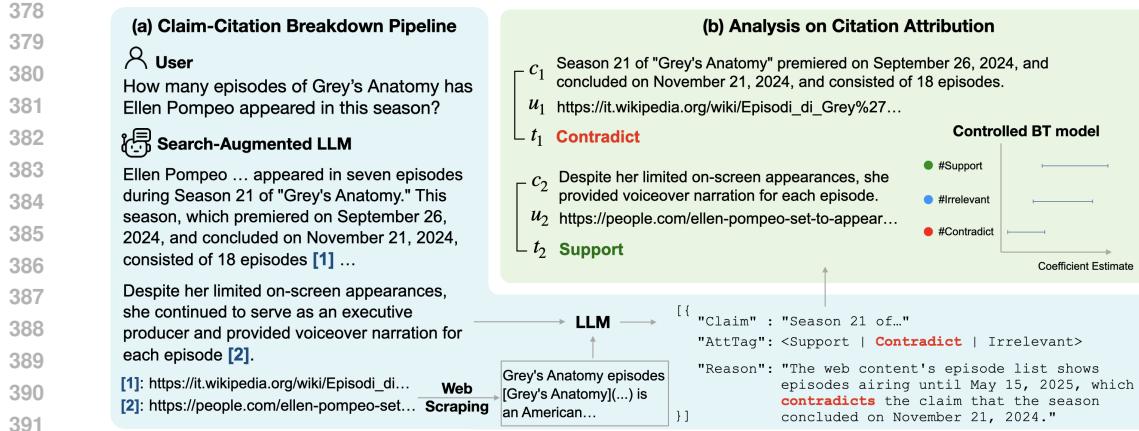


Figure 7: **Experimental Setup for Citation Attribution Analysis.** (a) For each multi-turn conversation, we retrieve the cited web content and use an LLM-based pipeline to decompose each model response into individual claims, followed by citation attribution labeling. (b) For each claim-citation pair  $(c_i, u_i, t_i)$ , we compute turn-level citation counts across the three categories and add corresponding features to the Bradley-Terry model. Results are shown in Figure 6 (Right).

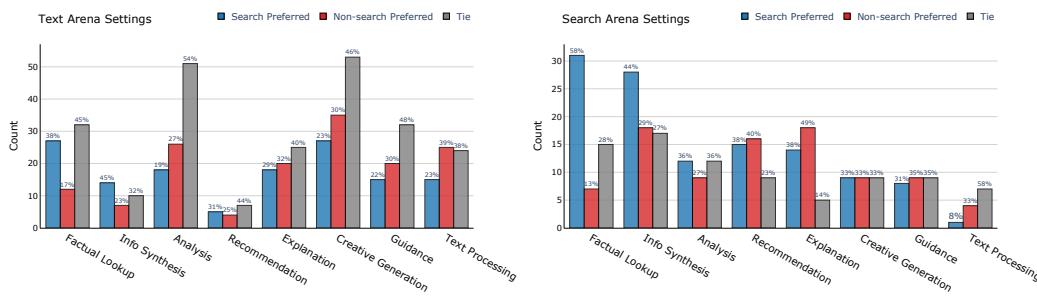


Figure 8: **Cross-Arena Vote Distribution** across Text Arena (Left) and Search Arena (Right) broken down into user intent categories. Users prefer the search model for *Factual Lookup* and *Info Synthesis* queries in both settings and the non-search model for *Text Processing* queries in Text Arena.

*Info Synthesis* ( $p\text{-value} = 0.095$ ), the difference is more pronounced in favor of the search model. Thus, the search-augmented model is preferred for knowledge acquisition tasks—even in the absence of well-defined user expectations—because web-grounded responses typically provide precise data, statistics, dates, names, and domain-specific terminology. We also note that the difference in performance for *Text Processing* ( $p\text{-value} = 0.077$ ) queries favors the non-search model. In these cases, the non-search model often provides structured responses (e.g., numbered or bulleted lists, headings), which suggests that presentation style may impact user evaluations.

**Search Arena Setting.** We collected 315 pairwise battles between a non-search model and search-augmented models in Search Arena, with 99 ties (31%), 126 search-preferred (40%), and 90 non-search-preferred votes (29%). The difference between search and non-search-preferred votes is statistically significant ( $p\text{-value} = 0.009$ ); the non-search model underperforms under a search-conditioned distribution. A detailed vote distribution by user intent is shown in Figure 8 (Right). Compared with the Text Arena (Figure 8 (Left)), tie votes are less frequent across all categories, indicating that the differences between model responses are more pronounced. The difference is most expressed for *Factual Lookup* ( $p\text{-value} = 5.8 \times 10^{-5}$ ) and *Info Synthesis* ( $p\text{-value} = 0.092$ ) queries.

These cross-area experiments demonstrate that search augmentation does not hurt performance in non-search settings and can improve responses to queries related to information retrieval and synthesis. However, removing web search significantly hurts model performance in search settings.

432 

## 4 RELATED WORK

434 **Large Language Models.** LLMs have made impressive advances in language understanding, dialogue  
 435 generation, and reasoning, enabled by techniques such as large-scale pretraining (Anthropic, 2024;  
 436 Bai et al., 2023; Brown et al., 2020; Grattafiori et al., 2024; Liu et al., 2024a; Touvron et al., 2023),  
 437 chain-of-thought prompting (Kojima et al., 2023; Wei et al., 2023; Shinn et al., 2023; Wang et al., 2023;  
 438 Yao et al., 2023a), and reinforcement learning with human feedback (RLHF) (Bai et al., 2022; Ouyang  
 439 et al., 2022a). The dataset and evaluation landscape has moved from static benchmarks (Hendrycks et al.,  
 440 2021a; Joshi et al., 2017; Hendrycks et al., 2021b) toward more challenging settings such as deep reasoning  
 441 (Li et al., 2024; Kazemi et al., 2025; Zeng et al., 2024; Lin et al., 2025), coding (Jimenez et al., 2024; Jain  
 442 et al., 2024), and open-ended dialogue under crowd-sourced evaluation (Chiang et al., 2024; Zhao et al.,  
 443 2024). While crowd-sourced setups can mitigate concerns over data contamination in pretraining (Yang  
 444 et al., 2023; Cheng et al., 2025b), recent work shows potential biases and oversights in human preferences  
 445 (Clark et al., 2021; Wu & Aji, 2025).

446 As LLMs gain tool-use capabilities (e.g., APIs, code interpreters, and web browsers) (Hilton et al., 2021;  
 447 Schick et al., 2023; Grattafiori et al., 2024; Patil et al., 2024; Yao et al., 2023b), domain-specific benchmarks  
 448 and datasets have emerged to analyze and evaluate LLMs under different environments (Zhou et al., 2024;  
 449 Li et al., 2023b; Liu et al., 2023). In the context of web search, several search-augmented LLMs have been  
 450 developed (OpenAI, 2024; Perplexity; Gemini, 2024), which retrieve live information to support better  
 451 reasoning. However, existing benchmarks such as SimpleQA (Wei et al., 2024) and BrowseComp (Wei et al.,  
 452 2025) are limited to single-turn, fact-based, monolingual queries. Although WebArena (Zhou et al., 2024)  
 453 contains diverse user prompts and a web-based interface, it emphasizes closed-world web navigation tasks  
 454 rather than open-ended search, reasoning, and dialogue. We introduce Search-Arena, the first large-scale,  
 455 crowd-sourced dataset for search-augmented LLMs with human preference signals, covering diverse intents,  
 456 topics, and multi-turn interactions across 70+ languages, collected through a transparent, open platform.

457 **Traditional and LLM-Integrated Information Retrieval (IR).** Information retrieval is a long-standing  
 458 task, with early methods like BM25 (Robertson et al., 1994), PageRank (Page et al., 1998), and  
 459 embedding-based approaches (Monir et al., 2024; Huang et al., 2020). Several static benchmarks (Thakur  
 460 et al., 2021; Muennighoff et al., 2023) and large-scale web search datasets with user logs and preference  
 461 signals (Voorhees & Harman, 2005; Craswell et al., 2020; Chen et al., 2024b) have enabled robust  
 462 evaluation of retrieval systems and comprehensive studies on user behavior.

463 With the rise of LLMs, information retrieval has moved beyond traditional search and become integrated  
 464 into LLM workflows. In Retrieval Augmented Generation (RAG) settings, retrieved text is appended to  
 465 the input prompt (Lewis et al., 2021; Guu et al., 2020; Hilton et al., 2021; Chuang et al., 2025); Datasets  
 466 for evaluating these LLM-IR systems span needle-in-a-haystack retrieval (Kamradt, 2023; Wu et al., 2025),  
 467 citation attribution (Zhang et al., 2024; Abolghasemi et al., 2024), and general question answering (Yang  
 468 et al., 2018; Han et al., 2024; Wei et al., 2024; 2025). **More recently, this line of work has been extended**  
 469 **from direct single-round information-seeking to natural, conversational search scenarios (Gao et al., 2023;**  
 470 **Mo et al., 2025; Cheng et al., 2025a; Adlakha et al., 2022), and then to search-augmented LLM systems**  
 471 **operating over the open web (Hilton et al., 2021; Schick et al., 2023; Yao et al., 2023b).** We view search-  
 472 **augmented LLM systems as a variant of RAG, with clear differences, including (1) “unbounded” search**  
 473 **space over the whole open web and (2) multi-round agent-based query refinement, search, and generation.**

474 However, for search augmented LLMs, existing benchmarks lack full human-AI interaction traces,  
 475 including queries, retrieved documents, responses, and preferences, which are needed for human-centric  
 476 analysis as done in traditional IR (Craswell et al., 2020; Chen et al., 2024b). To fill this gap, we introduce  
 477 Search Arena, a large-scale open dataset that enables human-centered analysis of this emerging interface.

478 

## 5 LIMITATIONS, CONCLUSION, AND FUTURE WORK

479 **Limitations, Future Work, and Broader Impact.** Crowd-sourced data analysis provides valuable  
 480 insight into real-world user preferences but comes with limitations and broader social implications. First,  
 481 the collected data may reflect demographic skews and may not be fully representative of the broader  
 482 population, as not all users choose to vote and human judgments are inherently subjective (Chen et al.,  
 483 2024a; Clark et al., 2021). Second, because conversational data with human preferences is personal  
 484 and potentially valuable for model improvement (Ouyang et al., 2022b), its release requires careful  
 485 consideration of both privacy and equitable access across the community.

486 To address these concerns, we anonymized model responses and randomized their left/right placement  
 487 to reduce known human biases. During the data collection period, no early access to data was granted  
 488 to model providers, nor were any pre-release models deployed on the platform. We obtained user consent  
 489 at interaction time and enforced a strict privacy policy. To aid responsible interpretation, we also analyzed  
 490 known biases (e.g., response length; [Subsection 3.1](#)) and reported user demographics in [Figure B.1](#) and  
 491 [Figure B.5](#). Additional details are provided in [Appendix A](#).

492 We identify several limitations in our current study that open avenues for future research. While our  
 493 primary focus is dataset construction and interaction analysis, the Search Arena dataset facilitates the  
 494 following directions:

- 496 • **Objective metrics:** As detailed in Section 3, user preferences are often influenced by  
 497 surface-level features, such as citation counts or source reputation, rather than the intrinsic quality  
 498 of the cited sources and the generated content. Consequently, user-provided labels may diverge  
 499 from objective standards like factuality, citation-claim consistency, and source specificity. Future  
 500 work should expand the evaluation framework to include these objective metrics and analyze  
 501 their correlation with human preference signals to better understand the gap between perceived  
 502 credibility and actual groundedness.
- 503 • **Offline evaluation:** Our current evaluation relies on an online human-in-the-loop setup, which,  
 504 while robust, restricts the scalability of testing new models. Future work can leverage our dataset  
 505 to construct automated benchmarks for offline evaluation. This could involve training reward  
 506 models or LLMs that approximate human preferences, potentially combining them with the  
 507 objective metrics discussed above to create a comprehensive scoring rubric.
- 508 • **Model development:** Our cross-arena and preference analyses reveal that blindly augmenting  
 509 models with search is not always optimal. Future research can utilize these insights to improve  
 510 both the search module (e.g., better source filtering) and the backbone LLM (e.g., optimizing  
 511 policies for *when* to trigger search versus relying on parametric knowledge).

512 **Conclusion.** We present Search Arena, the first large-scale dataset and analysis of human interactions  
 513 with search-augmented LLMs. Spanning over 24k multi-turn conversations and 12k human preference  
 514 votes across more than 70 languages, the dataset captures a broader range of user intents and topics than  
 515 prior benchmarks. Our analysis reveals that user preference is positively associated with citation count  
 516 and certain source types; however, models do not always cite correctly, highlighting a key challenge in  
 517 trustworthy systems. The cross-arena experiment further shows that search-augmented and non-search  
 518 models behave differently across settings. We release the full dataset to support future research in  
 519 search-augmented LLMs and human-centric analyses.

540 REFERENCES  
541

542 Amin Abolghasemi, Leif Azzopardi, Seyyed Hadi Hashemi, Maarten de Rijke, and Suzan Ver-  
543 berne. Evaluation of attribution bias in retrieval-augmented large language models, 2024. URL  
544 <https://arxiv.org/abs/2410.12380>. 9

545 Vaibhav Adlakha, Shehzaad Dhuliawala, Kaheer Suleman, Harm de Vries, and Siva Reddy. Topiocqa:  
546 Open-domain conversational question answering with topic switching. *Transactions of the Association  
547 for Computational Linguistics*, 2022. ISSN 2307-387X. doi: 10.1162/tacl\_a\_00471. URL  
548 [https://doi.org/10.1162/tacl\\_a\\_00471](https://doi.org/10.1162/tacl_a_00471). 9

549

550 Anthropic. The claude 3 model family: Opus, sonnet, haiku, 2024. URL [https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model\\_Card\\_Claude\\_3.pdf](https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf). 9

551

552

553 Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei  
554 Huang, et al. Qwen technical report, 2023. URL <https://arxiv.org/abs/2309.16609>. 9

555

556 Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain,  
557 Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion,  
558 Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan  
559 Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei,  
560 Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training  
561 a helpful and harmless assistant with reinforcement learning from human feedback, 2022. URL  
562 <https://arxiv.org/abs/2204.05862>. 9

563

564 Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of  
565 paired comparisons. *Biometrika*, 1952. 2, 4, 28

566

567 Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind  
568 Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot  
569 learners. *Advances in neural information processing systems*, 2020. 9

570

571 Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. Humans or LLMs  
572 as the judge? a study on judgement bias. In *Proceedings of the 2024 Conference on Empirical  
573 Methods in Natural Language Processing*. Association for Computational Linguistics, 2024a. URL  
<https://aclanthology.org/2024.emnlp-main.474/>. 9

574

575 Qi Chen, Xiubo Geng, Corby Rosset, Carolyn Buractaon, Jingwen Lu, Tao Shen, Kun Zhou, Chenyan  
576 Xiong, Yeyun Gong, Paul Bennett, et al. Ms marco web search: A large-scale information-rich web  
577 dataset with millions of real click labels. In *Companion Proceedings of the ACM Web Conference 2024*,  
2024b. 1, 9

578

579 Yiruo Cheng, Kelong Mao, Ziliang Zhao, Guanting Dong, Hongjin Qian, Yongkang Wu, Tet-  
580 suya Sakai, Ji-Rong Wen, and Zhicheng Dou. CÓRAL: Benchmarking multi-turn con-  
581 versational retrieval-augmented generation. In *Findings of the Association for Compu-  
582 tational Linguistics: NAACL 2025*. Association for Computational Linguistics, 2025a. URL  
<https://aclanthology.org/2025.findings-naacl.72/>. 9, 36

583

584 Yuxing Cheng, Yi Chang, and Yuan Wu. A survey on data contamination for large language models,  
585 2025b. URL <https://arxiv.org/abs/2502.14425>. 9

586

587 Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng  
588 Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena:  
589 An open platform for evaluating LLMs by human preference. In *Forty-first International Conference  
590 on Machine Learning*, 2024. URL <https://openreview.net/forum?id=3MW8GKNyzI>. 1,  
591 2, 3, 4, 5, 9, 28

592

593 Yung-Sung Chuang, Benjamin Cohen-Wang, Shannon Zejiang Shen, Zhao Feng Wu, Hu Xu, Xi Victoria  
Lin, James Glass, Shang-Wen Li, and Wen tau Yih. Selfcrite: Self-supervised alignment for context  
attribution in large language models, 2025. URL <https://arxiv.org/abs/2502.09604>. 9

594 Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith.  
 595 All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proceedings of the*  
 596 *59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint*  
 597 *Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational  
 598 Linguistics, 2021. URL <https://aclanthology.org/2021.acl-long.565/>. 9

599 Nick Craswell, Daniel Campos, Bhaskar Mitra, Emine Yilmaz, and Bodo Billerbeck. Orcas: 18 million  
 600 clicked query-document pairs for analyzing search. In *Proceedings of the 29th ACM International*  
 601 *Conference on Information & Knowledge Management*, 2020. 1, 9

602 Lisa Dunlap, Yuhui Zhang, Xiaohan Wang, Ruiqi Zhong, Trevor Darrell, Jacob Steinhardt, Joseph E. Gonza-  
 603 lez, and Serena Yeung-Levy. Describing differences in image sets with natural language. In *Proceedings*  
 604 *of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3, 24

605 Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D. Goodman. Cognitive  
 606 behaviors that enable self-improving reasoners, or, four habits of highly effective stars, 2025. URL  
 607 <https://arxiv.org/abs/2503.01307>. 5

608 Jianfeng Gao, Chenyan Xiong, Paul Bennett, and Nick Craswell. *Neural approaches to conversational*  
 609 *information retrieval*. Springer, 2023. 9

610 Gemini. Grounding with google search, 2024. 1, 9

611 Gemini. Gemini 2.5: Our most intelligent ai model, 2025.  
 612 URL [https://blog.google/technology/google-deepmind/](https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/)  
 613 [gemini-model-thinking-updates-march-2025/](https://arxiv.org/abs/2503.01307). 7

614 Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad  
 615 Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of  
 616 models, 2024. URL <https://arxiv.org/abs/2407.21783>. 9

617 Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure, 2022. URL  
 618 <https://arxiv.org/abs/2203.05794>. 21

619 Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Realm: Retrieval-  
 620 augmented language model pre-training, 2020. URL <https://arxiv.org/abs/2002.08909>. 9

621 Rujun Han, Yuhao Zhang, Peng Qi, Yumo Xu, Jenyuan Wang, Lan Liu, William Yang Wang, Bonan Min,  
 622 and Vittorio Castelli. Rag-qa arena: Evaluating domain robustness for long-form retrieval augmented  
 623 question answering, 2024. URL <https://arxiv.org/abs/2407.13998>. 9

624 Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and  
 625 Jacob Steinhardt. Measuring massive multitask language understanding, 2021a. URL  
 626 <https://arxiv.org/abs/2009.03300>. 9

627 Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song,  
 628 and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset, 2021b. URL  
 629 <https://arxiv.org/abs/2103.03874>. 9

630 Jacob Hilton, R Nakano, S Balaji, and John Schulman. Webgpt: Improving the factual accuracy of  
 631 language models through web browsing. *OpenAI Blog*, December, 2021. 1, 9

632 Jui-Ting Huang, Ashish Sharma, Shuying Sun, Li Xia, David Zhang, Philip Pronin, Janani Padmanabhan,  
 633 Giuseppe Ottaviano, and Linjun Yang. Embedding-based retrieval in facebook search. In *Proceedings*  
 634 *of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM,  
 635 2020. URL <http://dx.doi.org/10.1145/3394486.3403305>. 9

636 Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando  
 637 Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation  
 638 of large language models for code, 2024. URL <https://arxiv.org/abs/2403.07974>. 9

639 Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik  
 640 Narasimhan. Swe-bench: Can language models resolve real-world github issues?, 2024. URL  
 641 <https://arxiv.org/abs/2310.06770>. 9

648 Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. Triviaqa: A large  
 649 scale distantly supervised challenge dataset for reading comprehension, 2017. URL  
 650 <https://arxiv.org/abs/1705.03551>. 9

651 Gregory Kamradt. Llmtest\_needleinahaystack, 2023. URL [https://github.com/gkamradt/LLMTest\\_NeedleInAHaystack/blob/main/README.md](https://github.com/gkamradt/LLMTest_NeedleInAHaystack/blob/main/README.md). 9

652 Mehran Kazemi, Bahare Fatemi, Hritik Bansal, John Palowitch, Chrysovalantis Anastasiou, Sanket Vaibhav  
 653 Mehta, Lalit K. Jain, Virginia Aglietti, Disha Jindal, Peter Chen, Nishanth Dikkala, Gladys Tyen, Xin  
 654 Liu, Uri Shalit, Silvia Chiappa, Kate Olszewska, Yi Tay, Vinh Q. Tran, Quoc V. Le, and Orhan Firat.  
 655 Big-bench extra hard, 2025. URL <https://arxiv.org/abs/2502.19187>. 9

656 Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language  
 657 models are zero-shot reasoners, 2023. URL <https://arxiv.org/abs/2205.11916>. 9

658 Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman  
 659 Goyal, Heinrich Kütter, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and  
 660 Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021. URL  
 661 <https://arxiv.org/abs/2005.11401>. 9

662 Junyi Li, Tianyi Tang, Wayne Xin Zhao, Jingyuan Wang, Jian-Yun Nie, and Ji-Rong Wen. The  
 663 web can be your oyster for improving language models. In *Findings of the Association for  
 664 Computational Linguistics: ACL 2023*. Association for Computational Linguistics, 2023a. URL  
 665 <https://aclanthology.org/2023.findings-acl.46/>. 1

666 Minghao Li, Yingxiu Zhao, Bowen Yu, Feifan Song, Hangyu Li, Haiyang Yu, Zhoujun Li, Fei Huang,  
 667 and Yongbin Li. Api-bank: A comprehensive benchmark for tool-augmented llms, 2023b. URL  
 668 <https://arxiv.org/abs/2304.08244>. 9

669 Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E Gonzalez,  
 670 and Ion Stoica. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder  
 671 pipeline, 2024. URL <https://arxiv.org/abs/2406.11939>. 9, 34

672 Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri,  
 673 Ronan Le Bras, and Yejin Choi. Wildbench: Benchmarking LLMs with challenging tasks from real  
 674 users in the wild. In *The Thirteenth International Conference on Learning Representations*, 2025. URL  
 675 <https://openreview.net/forum?id=MKEHCx25xp>. 9

676 Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao,  
 677 Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report, 2024a. URL  
 678 <https://arxiv.org/abs/2412.19437>. 9

679 Junhua Liu, Yong Keat Tan, Bin Fu, and Kwan Hui Lim. Intent-aware dialogue gener-  
 680 ation and multi-task contrastive learning for multi-turn intent classification, 2024b. URL  
 681 <https://arxiv.org/abs/2411.14252>. 4

682 Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen  
 683 Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng  
 684 Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. Agentbench:  
 685 Evaluating llms as agents, 2023. URL <https://arxiv.org/abs/2308.03688>. 9

686 Fengran Mo, Kelong Mao, Ziliang Zhao, Hongjin Qian, Haonan Chen, Yiruo Cheng, Xiaoxi Li, Yutao  
 687 Zhu, Zhicheng Dou, and Jian-Yun Nie. A survey of conversational search. *ACM Trans. Inf. Syst.*, 2025.  
 688 doi: 10.1145/3759453. URL <https://doi.org/10.1145/3759453>. 9

689 Solmaz Seyed Monir, Irene Lau, Shubing Yang, and Dongfang Zhao. Vectorsearch: En-  
 690 hancing document retrieval with semantic embeddings and optimized search, 2024. URL  
 691 <https://arxiv.org/abs/2409.17383>. 9

692 Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. Mteb: Massive text embedding  
 693 benchmark, 2023. URL <https://arxiv.org/abs/2210.07316>. 9

694 OpenAI. Introducing chatgpt search, 2024. URL <https://openai.com/index/introducing-chatgpt-search/>. 1, 9

702 OpenAI. Introducing gpt-4.1 in the api, 2025. URL <https://openai.com/index/gpt-4-1/>. 19  
 703

704 Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong  
 705 Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton,  
 706 Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and  
 707 Ryan Lowe. Training language models to follow instructions with human feedback, 2022a. URL  
 708 <https://arxiv.org/abs/2203.02155>. 9

709 Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong  
 710 Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser  
 711 Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan  
 712 Leike, and Ryan Lowe. Training language models to follow instructions with human feed-  
 713 back. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2022b.  
 714 URL [https://proceedings.neurips.cc/paper\\_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf). 9  
 715

716 Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking:  
 717 Bringing order to the web, 1998. 9  
 718

719 Shishir G Patil, Tianjun Zhang, Xin Wang, and Joseph E Gonzalez. Gorilla: Large language model  
 720 connected with massive apis. *Advances in Neural Information Processing Systems*, 2024. 9  
 721

722 Perplexity. Sonar by perplexity. URL <https://docs.perplexity.ai/home>. 1, 9  
 723

724 Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. Okapi  
 725 at trec-3. In *Proceedings of the Third Text REtrieval Conference (TREC-3)*, 1994. 9  
 726

727 Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer,  
 728 Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use  
 729 tools, 2023. URL <https://arxiv.org/abs/2302.04761>. 9  
 730

731 Chirag Shah, Ryen W. White, Reid Andersen, Georg Buscher, and Scott Counts. Using  
 732 large language models to generate, validate, and apply user intent taxonomies, 2023. URL  
 733 <https://arxiv.org/abs/2309.13063>. 4  
 734

735 Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and  
 736 Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning, 2023. URL  
 737 <https://arxiv.org/abs/2303.11366>. 9  
 738

739 Mark Steyvers, Heliodoro Tejeda Lemus, Aakriti Kumar, Catarina Belém, Sheer Karny, Xinyue Hu, Lukas  
 740 Mayer, and Padhraic Smyth. The calibration gap between model and human confidence in large language  
 741 models, 2024. URL <https://api.semanticscholar.org/CorpusID:267211649>. 5  
 742

743 Kelly Tang, Wei-Lin Chiang, and Anastasios N. Angelopoulos. Arena explorer: A topic modeling pipeline  
 744 for llm evals & analytics, 2025. 21  
 745

746 Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. Beir: A  
 747 heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth  
 748 Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*,  
 749 2021. URL <https://openreview.net/forum?id=wCu6T5xFjeJ>. 9  
 750

751 Wei-Lin Chiang Tianle Li, Anastasios Angelopoulos. Does style matter? disentangling style and substance  
 752 in chatbot arena, 2024. URL <https://blog.lmarena.ai/blog/2024/style-control/>.  
 753 2, 4, 5  
 754

755 Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix,  
 756 Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin,  
 757 Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.  
 758 URL <https://arxiv.org/abs/2302.13971>. 9  
 759

760 Ellen M. Voorhees and Donna K. Harman (eds.). *TREC: Experiment and Evaluation in Information  
 761 Retrieval*. MIT Press, 2005. 9

756 Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery,  
 757 and Denny Zhou. Self-consistency improves chain of thought reasoning in language models, 2023.  
 758 URL <https://arxiv.org/abs/2203.11171>. 9

759

760 Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le,  
 761 and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL  
 762 <https://arxiv.org/abs/2201.11903>. 9

763 Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John  
 764 Schulman, and William Fedus. Measuring short-form factuality in large language models, 2024. URL  
 765 <https://arxiv.org/abs/2411.04368>. 1, 2, 3, 9, 34, 36

766 Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung,  
 767 Alex Tachard Passos, William Fedus, and Amelia Glaese. Browsecmp: A simple yet challenging  
 768 benchmark for browsing agents, 2025. URL <https://arxiv.org/abs/2504.12516>. 1, 2,  
 769 3, 9, 34, 36

770

771 Minghao Wu and Alham Fikri Aji. Style over substance: Evaluation biases for large language models. In  
 772 *Proceedings of the 31st International Conference on Computational Linguistics*. Association for Computational  
 773 Linguistics, 2025. URL <https://aclanthology.org/2025.coling-main.21>. 9

774 Tsung-Han Wu, Giscard Biamby, Jerome Quenum, Ritwik Gupta, Joseph E. Gonzalez, Trevor  
 775 Darrell, and David Chan. Visual haystacks: A vision-centric needle-in-a-haystack benchmark.  
 776 In *The Thirteenth International Conference on Learning Representations*, 2025. URL  
 777 <https://openreview.net/forum?id=9JCNPFL1f9>. 9

778 Shuo Yang, Wei-Lin Chiang, Lianmin Zheng, Joseph E. Gonzalez, and Ion Stoica. Rethinking  
 779 benchmark and contamination for language models with rephrased samples, 2023. URL  
 780 <https://arxiv.org/abs/2311.04850>. 9

781

782 Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and  
 783 Christopher D. Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In  
 784 *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association  
 785 for Computational Linguistics, 2018. URL <https://aclanthology.org/D18-1259>. 9

786 Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik  
 787 Narasimhan. Tree of thoughts: Deliberate problem solving with large language models, 2023a. URL  
 788 <https://arxiv.org/abs/2305.10601>. 9

789

790 Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and  
 791 Yuan Cao. React: Synergizing reasoning and acting in language models, 2023b. URL  
 792 <https://arxiv.org/abs/2210.03629>. 9

793

794 Zhongshen Zeng, Yinhong Liu, Yingjia Wan, Jingyao Li, Pengguang Chen, Jianbo Dai, Yuxuan Yao,  
 795 Rongwu Xu, Zehan Qi, Wanru Zhao, Linling Shen, Jianqiao Lu, Haochen Tan, Yukang Chen, Hao  
 796 Zhang, Zhan Shi, Bailin Wang, Zhijiang Guo, and Jiaya Jia. Mr-ben: A meta-reasoning benchmark for  
 797 evaluating system-2 thinking in llms, 2024. URL <https://arxiv.org/abs/2406.13975>. 9

798

799 Jiajie Zhang, Yushi Bai, Xin Lv, Wanjun Gu, Danqing Liu, Minhao Zou, Shulin Cao, Lei Hou, Yuxiao  
 800 Dong, Ling Feng, and Juanzi Li. Longcite: Enabling LLMs to generate fine-grained citations in  
 801 long-context QA, 2024. URL <https://openreview.net/forum?id=mMXdHyBcHh>. 9

802

803 Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. Wildchat:  
 804 1m chatGPT interaction logs in the wild. In *The Twelfth International Conference on Learning  
 805 Representations*, 2024. URL <https://openreview.net/forum?id=Bl8u7ZR1bM>. 1, 9, 36

806

807 Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yong-  
 808 hao Zhuang, Zhuohan Li, Zi Lin, Eric P. Xing, Joseph E. Gonzalez, Ion Stoica, and Hao  
 809 Zhang. Lmsys-chat-1m: A large-scale real-world llm conversation dataset, 2024. URL  
 810 <https://arxiv.org/abs/2309.11998>. 1

811

812 Ruiqi Zhong, Charlie Snell, Dan Klein, and Jacob Steinhardt. Describing differences between text distribu-  
 813 tions with natural language. In *Proceedings of the 39th International Conference on Machine Learning*.  
 814 PMLR, 2022. URL <https://proceedings.mlr.press/v162/zhong22a.html>. 3, 24

810 Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue  
811 Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. Webarena: A realistic web environment  
812 for building autonomous agents, 2024. URL <https://arxiv.org/abs/2307.13854>. 9  
813  
814  
815  
816  
817  
818  
819  
820  
821  
822  
823  
824  
825  
826  
827  
828  
829  
830  
831  
832  
833  
834  
835  
836  
837  
838  
839  
840  
841  
842  
843  
844  
845  
846  
847  
848  
849  
850  
851  
852  
853  
854  
855  
856  
857  
858  
859  
860  
861  
862  
863

## APPENDIX

864  
 865  
 866 [Appendix A](#) describes platform design, supported model, and policies of our data collection process.  
 867 [Appendix B](#) outlines intent classification, topic modeling, and filtering procedures.  
 868 [Appendix C](#) includes details (e.g., parameters, prompts) on the dataset differencing experiments.  
 869 [Appendix D](#) includes model performance and user preference analysis on Search Arena.  
 870 [Appendix E](#) outlines citation analysis, user preference experiments, and citation attribution pipeline.  
 871 [Appendix F](#) includes additional results on the cross-benchmark analysis.  
 872 [Appendix G](#) includes comparative analysis with additional related datasets.  
 873  
 874  
 875

## A SEARCH ARENA PLATFORM

876 As described in [Section 2](#), Search Arena (<URL IN CAMERA READY>) is an open, crowdsourced eval-  
 877 uation platform for search-augmented LLMs, launched on March 18, 2025. This section outlines our data  
 878 collection and release protocols, followed by a description of the supported models and key design decisions.  
 879  
 880

## A.1 DATA

881 **Data Collection.** The Search Arena platform does not require user login, but users must explicitly accept  
 882 the Terms of Service before using the platform. For each prompt, responses from two anonymous models  
 883 are displayed side-by-side, and users may cast a preference vote at any time during the interaction. The  
 884 full text of Terms of Service is at [Figure A.1](#).  
 885

886 Over the course of a 7-week data collection period, the platform recorded an average of roughly 800-1,500  
 887 conversations per day. After filtering out examples with server errors, inconsistent configurations, or  
 888 other quality issues, we retain approximately 24,000 conversations, about half of which include human  
 889 preference votes. [Figure A.2](#) shows daily traffic trends, with spikes aligned to major model updates or  
 890 platform announcements, and dips on weekends.  
 891

## Terms of Service

892 Users are required to agree to the following terms before using the service: The service is a research  
 893 preview. It only provides limited safety measures and may generate offensive content. It must not be used  
 894 for any illegal, harmful, violent, racist, or sexual purposes. Please do not upload any private information.  
 895 The service collects user dialogue data, including both text and images, and reserves the right to distribute  
 896 it under a Creative Commons Attribution (CC-BY) or a similar license. You may only use this website for  
 897 your personal or internal business purposes. You must not access the website programmatically, scrape or  
 898 extract data, manipulate any leaderboard or ranking, or authorize or pay others to access or use the website  
 899 on your behalf. Unauthorized use may result in suspension or termination of your access, including access  
 900 by your organization.  
 901

902 Figure A.1: Term of Service of the Search Arena Platform.  
 903

904 **Data Release Policy.** Per our Terms of Service ([Figure A.1](#)), all conversation data is released under  
 905 a Creative Commons Attribution (CC-BY) license. To protect user privacy, we apply automated  
 906 de-identification using Google’s Data Loss Prevention (DLP) API, which removes personal identifiers  
 907 such as email addresses, credit card numbers, and API keys. Approximately 2% of examples were flagged  
 908 by this process. [Users may also contact the platform provider to request their records be removed from  
 909 the released dataset through the following link: \[ANONYMOUS LINK DURING REVIEW\].](#)  
 910

## A.2 MODELS

911 Search Arena currently supports 12 search-augmented LLMs from Perplexity, Gemini, and OpenAI, as  
 912 summarized in [Table A.1](#). Unless otherwise noted, we group models with different inline citation styles  
 913 (elaborated below) of the same base model. Each model is accessed via its provider’s public API using the  
 914 default configuration, including `search_context_size=medium` for Perplexity and OpenAI mod-  
 915 els. We also evaluate several additional variants, including (1) OpenAI’s `gpt-4o-search-preview`  
 916 with `user_location` enabled and (2) Perplexity and OpenAI models configured with  
 917 `search_context_size=high`. [Figure D.1](#) shows the number of battles per model.  
 918

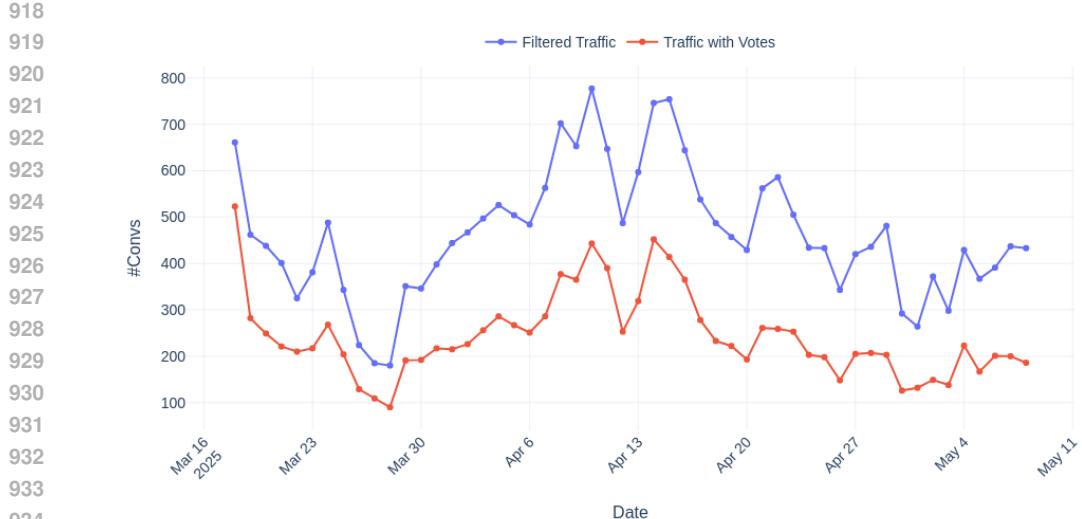


Figure A.2: Daily traffic on the Search Arena platform.

Table A.1: List of supported models and their configurations on Search Arena. <sup>†</sup> We evaluate OpenAI’s [web search API](#), which differs from the search feature in the ChatGPT product.

Provider	Model	Base Model	Details
Perplexity	pplx-sonar	sonar	Default config
	pplx-sonar-pro	sonar-pro	Default config
	pplx-sonar-pro-high	sonar-pro	search_context_size=high
	pplx-sonar-reasoning	sonar-reasoning	Default config
	pplx-sonar-reasoning-pro	sonar-reasoning-pro	search_context_size=high
Gemini	gemini-2.0-flash-grounding	gemini-2.0-flash	With Google Search enabled
	gemini-2.5-flash-grounding	gemini-2.5-flash	With Google Search enabled
	gemini-2.5-pro-grounding	gemini-2.5-pro-exp-03-25	With Google Search enabled
OpenAI <sup>†</sup>	api-gpt-4o-mini-search-preview	gpt-4o-mini	Default config
	api-gpt-4o-search-preview	gpt-4o	Default config
	api-gpt-4o-search-preview-high	gpt-4o	search_context_size=high
	api-gpt-4o-search-preview-high-loc	gpt-4o	user_location feature enabled

**Model Anonymity and Citation Style Control.** Each provider uses a distinct citation style, which may unintentionally reveal model identity. At the same time, citation formatting can influence user preferences. To manage this tradeoff, we implement a citation *style randomization* mechanism: model outputs are rendered using either a standardized format or the provider’s original style. This approach reduces the risk of user-side de-anonymization while allowing us to study how citation style affects user behavior. We found that the inline citation style does not significantly affect user preferences and model rankings.

## B DATA STATISTICS

### B.1 LINGUISTIC FEATURES

Users’ demographic information in Search Arena, based on country codes extracted from IP addresses, is shown in [Figure B.1](#). The dataset includes 11,650 unique users across 136 countries, with the United States (18.7%), Russia (9.9%), Germany (5.9%), and China (5.5%) as the top four.

Language distribution (of the top 50 languages) of Search Arena prompts is shown in [Figure B.5](#). The prompts span 71 languages with English (56.4%), Russian (11.4%), and Chinese (6.8%) as the top three.

As shown in [Figure B.2](#), Search Arena prompt lengths vary across intent categories. As expected, *Text Processing* (85.1 words) and *Analysis* (72.7 words) prompts are generally longer, while *Explanation* (24.7 words) and *Factual Lookup* (16.3 words) prompts are shorter.

Search Arena conversations are multi-turn, with 22.4% of all conversations containing more than one turn. Specifically, there are 3,288 conversations with 2 turns, 966 with 3 turns, and 460 with 4 turns. The distribution is shown in [Figure B.3](#).

972 B.2 USER INTENT ANNOTATION  
973974 In [Subsection 2.2](#) and [Figure 1](#), we briefly introduced the intent annotation pipeline and its high-level  
975 findings. Here, we provide additional details of the pipeline.976 First, the three co-authors of the paper did open-text annotations on 100 English prompts randomly sampled  
977 from the collected dataset. The annotators then met to consolidate their annotations into a taxonomy of  
978 primary and secondary user intent categories, including definitions and representative examples. The final  
979 taxonomy includes the following nine intent categories: *Factual Lookup, Information Synthesis, Analysis,*  
980 *Recommendation, Explanation, Creative Generation, Guidance, Text Processing, and Other*. Full category  
981 descriptions are shown in [Table B.1](#).982 To validate the taxonomy, the annotators labeled a subset of 100 prompts. The human inter-annotator  
983 agreement, measured by Cohen’s Kappa, ranged from 0.65 for primary labels to 0.79 when considering  
984 top-two matches (substantial agreement).985 We then scaled the annotation to the full dataset using GPT-4.1 ([OpenAI, 2025](#)), using the same in-context  
986 examples from [Table B.1](#). The total annotation cost was approximately \$20 USD. To evaluate label quality,  
987 we compared GPT-4.1 annotations with human labels on 150 samples drawn from the top three languages  
988 in our dataset (English, Russian, and Chinese). The resulting Cohen’s Kappa score of 0.812 on top-2  
989 intents indicates strong agreement. The full GPT-4.1 prompt is provided below.

990

991

992  Prompts for LLM-based Intent Classification

993

994 You are an impartial classifier.

995 996 997 998 999 1000  
994 TASK 1: Primary intent: choose one category that best matches the  
995 user’s intent, always ask yourself, what is the user trying to get the  
996 model to do with this query. Considering them following the categories  
one by one in order.1001 1002 1003 1004 1005 1006 1007 1008 1009 1010 1011 1012  
1001 TASK 2: If there is a clear secondary goal, choose one additional  
1002 intent label that is also present in the query. Otherwise, return "Unassigned" if no clear second goal ex-  
1003 ists.1004 1005 1006 1007 1008 1009 1010 1011 1012  
1004 Allowed intent categories:1005 1006 1007 1008 1009 1010 1011 1012  
1005 1. Text Processing  
1006 2. Creative Generation  
1007 3. Factual Lookup  
1008 4. Info Synthesis  
1009 5. Analysis  
1010 6. Recommendation  
1011 7. Explanation  
1012 8. Guidance  
1013 9. Other1014 1015 1016 1017 1018 1019 1020 1021 1022 1023 1024 1025  
1014 Here are a few examples that you should follow, try to general-  
1015 ize beyond these examples but hold true to their semantic mean-  
1016 1017 1018 1019 1020 1021 1022 1023 1024 1025  
1016 ing:  
1017 {FEW SHOT IN-CONTEXT EXAMPLES}1018 1019 1020 1021 1022 1023 1024 1025  
1018 Respond only in valid JSON, only choosing exactly one category to fill  
1019 the values:  
1020  
1021 {  
1022 "primary\_intent": "<primary intent label>",  
1023 "secondary\_intent": "<secondary intent label>",  
1024 "reasoning": "<your reasoning for the classification>",  
1025 }  
1024 User query: {USER PROMPT}

Table B.1: User intent taxonomy including category descriptions and examples.

Category	Description	Examples
Text Processing	Users request linguistic tasks such as summarizing, translating, paraphrasing, or refining text.	<ul style="list-style-type: none"> <li>Summarize &lt;website&gt;.</li> <li>Translate this paragraph from Spanish to English.</li> <li>Rephrase this email professionally for an investment-banking application.</li> </ul>
Creative Generation	Users request original creative outputs, including fictional scenarios, satire, poetry, storytelling, or other forms of artistic language generation.	<ul style="list-style-type: none"> <li>Compose a witty review of ITV’s <i>Big Brother</i>.</li> <li>Create a humorous poem about <i>SpongeBob</i>.</li> <li>Write a short fantasy story set on Mars.</li> </ul>
Factual Lookup	Users seek to retrieve a precise, objective fact or specific information (what/who/when-like questions).	<ul style="list-style-type: none"> <li>Who invented lambda calculus?</li> <li>How many planets are in our Solar System?</li> <li>When was the Mona Lisa painted?</li> </ul>
Info. Synthesis	Users seek a concise, aggregated summary that combines facts or perspectives from multiple sources. The focus is on integrating factual content without requiring reasoning, subjective interpretation, or personalization.	<ul style="list-style-type: none"> <li>List the key functions of the U.S. Congress.</li> <li>Summarize the main events of the French Revolution.</li> </ul>
Analysis	Users seek a reasoned judgment or breakdown of a topic, often involving comparison, evaluation, weighing different perspectives, and synthesizing material from many sources. Often open-ended.	<ul style="list-style-type: none"> <li>Analyze the pros and cons of nuclear energy.</li> <li>When will we reach artificial general intelligence?</li> <li>What were the strategic implications of the Cuban Missile Crisis?</li> </ul>
Recommendation	Users request suggestions or advice tailored to particular constraints, preferences, or specified criteria. Often implies personalization or ranking.	<ul style="list-style-type: none"> <li>Best laptop for machine learning under \$1500.</li> <li>I like literary fiction—what books should I read?</li> <li>What programming language should I learn for web development?</li> </ul>
Explanation	Users seek detailed clarifications, educational insights, or thorough elaborations aimed at better understanding a concept, process, or phenomenon. May ask “why” or “how” something works without implying action.	<ul style="list-style-type: none"> <li>What led to the fall of the Roman Empire?</li> <li>Can you explain the Heisenberg uncertainty principle?</li> <li>How does gradient-descent optimization work in machine learning?</li> </ul>
Guidance	Users request instructions, procedures, or practical advice intended to accomplish specific tasks, typically involving sequential steps or troubleshooting. Usually framed as “how to” or involving action.	<ul style="list-style-type: none"> <li>How do I install Docker on Ubuntu?</li> <li>How do I renew my passport online?</li> <li>How do I replace the battery in a MacBook Air?</li> </ul>
Other	Prompts that don’t fit neatly into any of the categories above. Prompts may be malformed, declarative in nature, or lacking in meaningful user intent.	<ul style="list-style-type: none"> <li>pizza.</li> <li>The table hit my head.</li> <li>Talk about something please.</li> </ul>

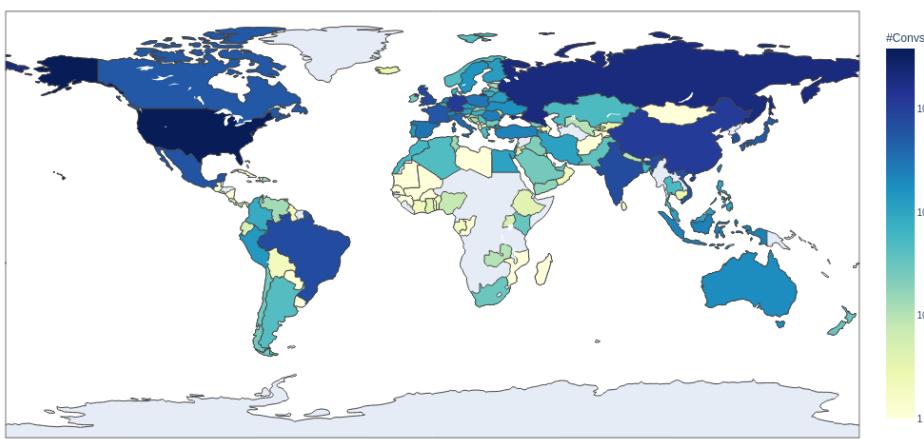


Figure B.1: **Search Arena Users’ Demographics.** Search Arena data includes 11,650 unique users across 136 countries.

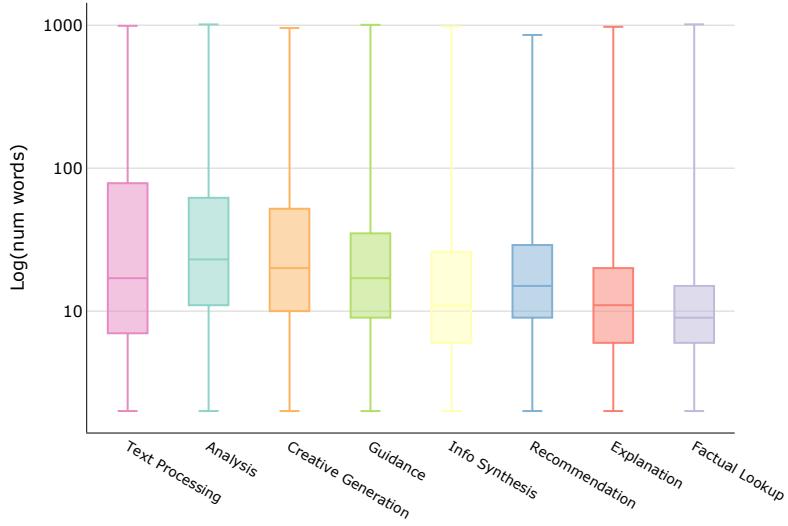


Figure B.2: **Search Arena Prompt Length Distribution by Intent.** Prompt lengths vary across intent categories. *Text Processing* (85.1 words) and *Analysis* (72.7 words) prompts are generally longer, while *Explanation* (24.7 words) and *Factual Lookup* (16.3 words) prompts are shorter.

### B.3 TOPIC MODELING

To analyze topic diversity across benchmarks, we apply the BERTopic framework (Grootendorst, 2022) to the 11,764 unique English prompts in the Search Arena dataset. We generate prompt embeddings using OpenAI’s `text-embedding-3-1large` model, perform dimensionality reduction and clustering via UMAP and HDBSCAN, and summarize each resulting cluster using GPT-4o (Tang et al., 2025). We adapt the Arena Explorer (Tang et al., 2025) methodology for the Search Arena dataset. Figure B.4 illustrates the distribution of topics derived from BERTopic clustering over the Search Arena prompts. The prominence of categories such as Technology Comparisons, Market Analysis, and Entertainment Characters, alongside a long-tail of niche domains, highlights the dataset’s breadth and its suitability for evaluating search-augmented LLMs under diverse topics.

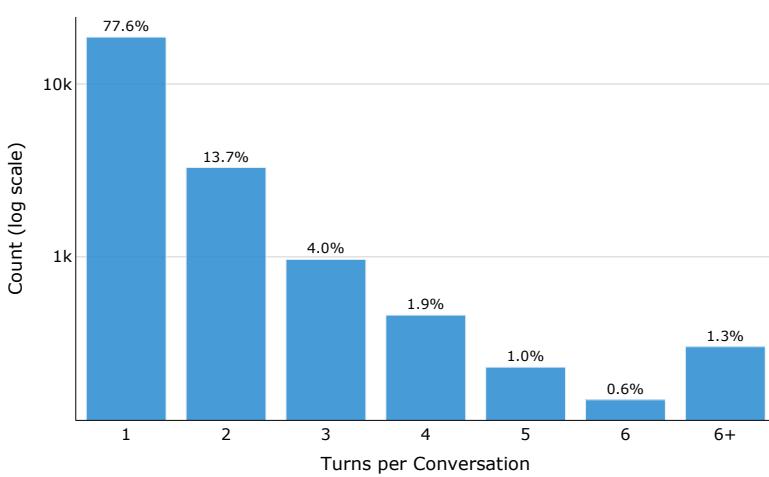


Figure B.3: **Search Arena Conversation Length (Number of Turns) Distribution.** Search Arena chats are multi-turn with 22.4% of conversations containing more than 1 turn.

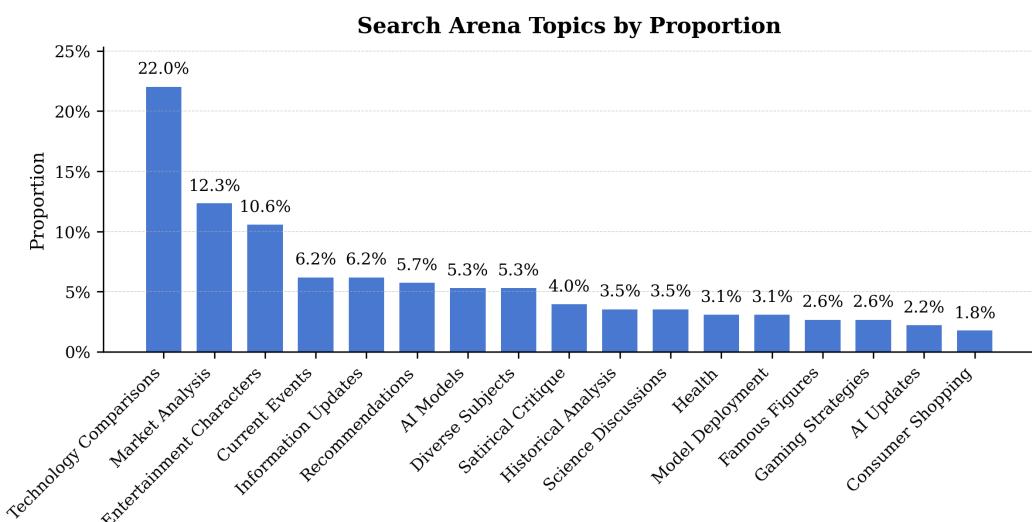
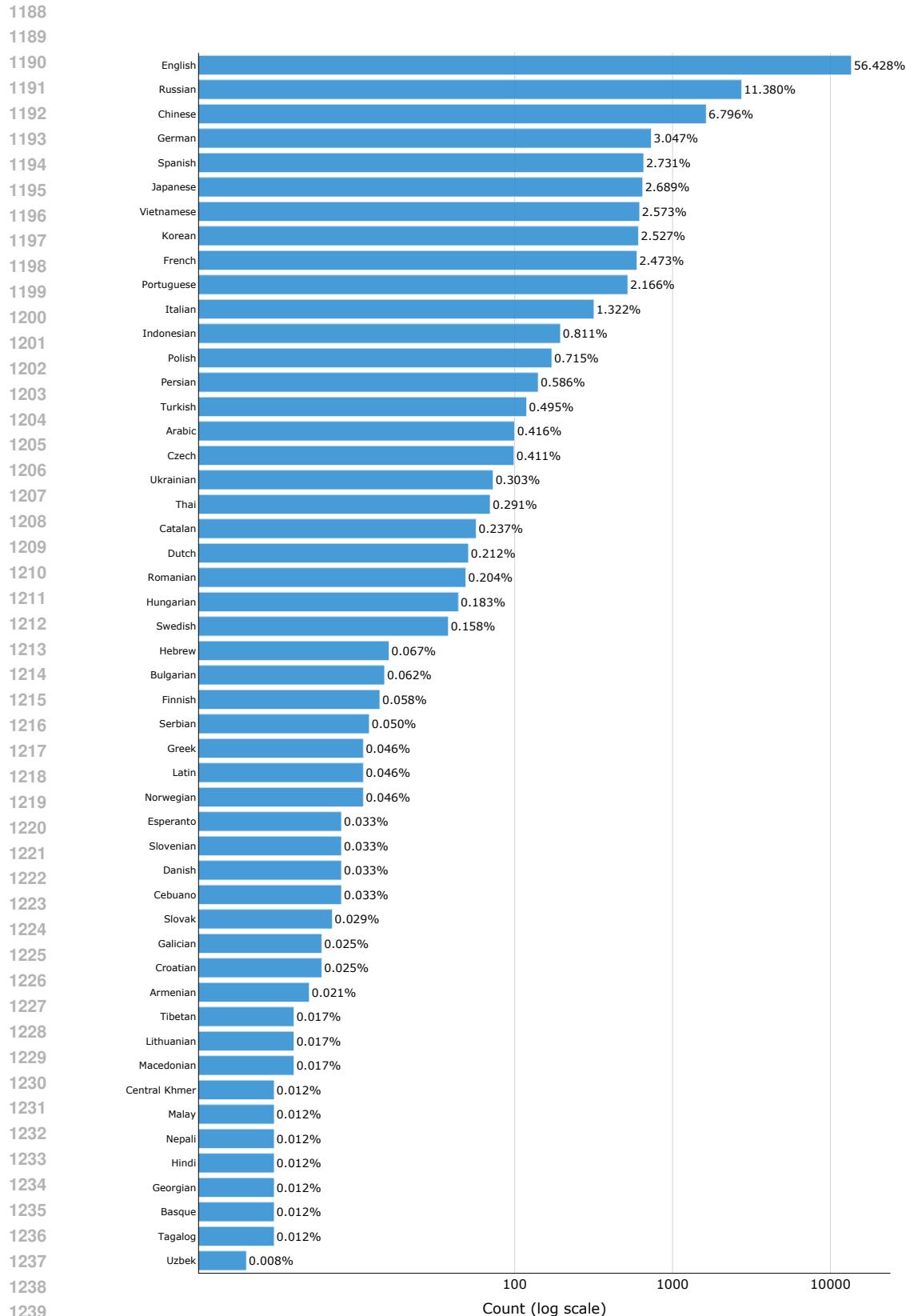


Figure B.4: **Top Topic Categories in Search Arena.** This figure shows the distribution of topic clusters. The most prevalent topics include *Technology Comparisons* (22.0%), *Market Analysis* (12.3%), and *Entertainment Characters* (10.6%). Less frequent but still diverse topics include health and shopping. The long-tail distribution reflects the breadth of real-world usage of search-augmented LLMs.

Figure B.5: **Search Arena Language Distribution (top 50).** Search Arena prompts span 71 languages.

1241

---

## 1242 C DESCRIBING DIFFERENCES

1244 To extract interpretable differences across two text corpora (e.g., prompts, responses), we use LLM-based  
 1245 dataset differencing methods (Zhong et al., 2022; Dunlap et al., 2024).

### 1247 C.1 PROMPT DIFFERENCES

1249 To describe differences in prompt distributions between Search Arena and other datasets (see analysis  
 1250 in [Subsection 2.2](#)), we use GPT-4.1 to propose properties (hypotheses) across 16 rounds, with 32 samples  
 1251 per group in each round. We then use o3 to filter the top five properties and GPT-4.1-mini to re-rank them.  
 1252 The prompts are shown below. We use the default temperature parameter of 1.0.

1253 We extract the following properties ( $p < 0.001$ ) for each dataset pair and report the proportion of validation  
 1254 set examples (100 samples) satisfying each property.

#### 1255 **Search Arena vs SimpleQA:**

- 1256 Requests in-depth explanations, analyses, or step-by-step guidance rather than single factual  
 1257 answers (28% vs 0%).
- 1258 Requests recent or real-time information, including current events, product features, or online  
 1259 research (35% vs 7%).
- 1260 Seeks technical help, troubleshooting, or comparative evaluations for software, programming,  
 1261 or digital tools (17% vs 1%).
- 1262 Asks for creative content generation, rewriting, or stylistic transformations (stories, poems,  
 1263 satirical pieces, etc.) (13% vs 0%).

#### 1265 **SimpleQA vs Search Arena:**

- 1266 Requests exact factual details such as specific names, dates, numbers, or titles (97% vs 41%).
- 1267 Expects a single objective, verifiable answer rather than explanations or analysis (97% vs 43%).
- 1268 Avoids subjective, open-ended, or creative requests, limiting queries to factual retrieval (99%  
 1269 vs 52%).
- 1270 Focuses on niche or lesser-known historical, scientific, or cultural topics (51% vs 7%).
- 1271 Presents concise, narrowly scoped questions with minimal background information (91% vs 56%).

#### 1275 **Search Arena vs BrowseComp:**

- 1276 User messages are brief and provide minimal contextual detail (80% vs 23%).
- 1277 Responses expected are immediate functional outputs such as lists, summaries, or code snippets  
 1278 (50% vs 20%).
- 1279 Prompts seek practical advice or step-by-step instructions for real-world tasks (29% vs 1%).
- 1280 Queries focus on well-known, mainstream topics, products, or services (75% vs 48%).

#### 1282 **BrowseComp vs Search Arena:**

- 1283 Frames the query as a deductive puzzle or investigative challenge (80% vs 3%).
- 1284 Provides explicit temporal or geographic constraints that narrow the solution space (99% vs 25%).
- 1285 Requires synthesizing multiple detailed clues from different sources to deduce the answer (92%  
 1286 vs 36%).

#### 1289 **Search Arena vs Text Arena:**

- 1290 Requests for current factual information about real-world entities, events, or products (61% vs  
 1291 25%).
- 1292 Requests for technical comparisons, evaluations, or purchasing guidance on devices, software,  
 1293 or services (16% vs 4%).

#### 1295 **Text Arena vs Search Arena:**

1296     • Seeks analytical, step-by-step solutions to mathematical, logical, or technical puzzles and  
 1297        explanations (21% vs 4%).  
 1298     • Requests programming assistance, such as debugging code, generating scripts, or explaining  
 1299        programming concepts (23% vs 8%).  
 1300     • Requests creative writing outputs, including stories, poems, jokes, or fictional scenarios (20%  
 1301        vs 5%).  
 1302

1303     **LLM Prompt for Proposing Distinguishing Properties between Prompt Sets**

1305     The following are a two separate lists of prompts that users have asked  
 1306        to a chatbot:  
 1307        <START>  
 1308        {text}  
 1309        <END>  
 1310        I am a machine learning researcher trying to figure out the major differ-  
 1311        ences between these two groups of prompts. This is a very small portion  
 1312        of the data, so I want the differences to be general.  
 1313        Please provide a list of the top three differences (separated by bullet  
 1314        points "\*") between the prompts in Group A and Group B.  
 1315  
 1316        Follow the detailed instructions below:  
 1317        - Identify properties that are more common in Group A prompts compared to  
 1318        Group B prompts.  
 1319        - The property descriptions should be simple and detailed. Do not pro-  
 1320        duce vague and generic descriptions.  
 1321        - Start the description of each property with "User". Do not use key-  
 1322        words like "group A", "group B", "more common", "less common", "fre-  
 1323        quently", "occasionally" in your response.  
 1324        - Do not combine multiple properties / features in a single bullet point.  
 1325        - An example of the desired output format:  
 1326        \* "Property 1"  
 1327        \* "Property 2"  
 1328        \* "Property 3"  
 1329  
 1330        Please order your response in terms of the most common differences be-  
 1331        tween the two groups. Your response:

1331     **LLM Prompt for Reducing Proposed Properties**

1332        PROPERTIES:  
 1333        #####  
 1334        {properties}  
 1335        #####  
 1336  
 1337        Above is a list of properties, several of which are similar.  
 1338        - Please reduce this to a list of the five most common distinct proper-  
 1339        ties, ordered by frequency of occurrence.  
 1340        - There should be no similar / overlapping properties in the final list.  
 1341        - Do not group multiple unrelated properties / features into a single  
 1342        property.  
 1343        - Do not use keywords, like "frequent", "rare", "common", "uncommon",  
 1344        etc.  
 1345        - Only respond with a numbered list of properties, no other text.

1350  
1351  
1352  **LLM Prompt for Ranking Properties**  
1353  
1354 PROMPT:  
1355 <START>  
1356 {text}  
1357 <END>  
1358  
1359 PROPERTY:  
1360 <START>  
1361 {hypothesis}  
1362 <END>  
1363  
1364 Check whether the PROMPT satisfies the PROPERTY. Respond with Yes or No.  
1365 If you are unsure, respond with No.  
1366 Output:  
1367  
1368

## 1364 C.2 RESPONSE DIFFERENCES

1365  
1366 To describe differences in responses between search and non-search models in cross-area deployments  
1367 (see analysis in [Subsection 3.3](#)), we use GPT-4.1 to propose properties (hypotheses) across 32 rounds,  
1368 with 8 samples per group in each round. We then use o3 to filter the top five properties and GPT-4.1-mini  
1369 to re-rank them. The prompts are shown below (for property reduction, we re-use the same prompt from  
1370 [Subsection C.1](#)). We use the default temperature parameter of 1.0.

1371 For each dataset pair, we extract the top properties and report the proportion of validation set examples  
1372 (100 samples) satisfying the properties, along with corresponding significance levels (p-values).

### 1373 **Text Arena Setting**

1374 Search vs non-search model responses to *Factual Lookup* prompts:

1375

- 1376 • Includes precise quantitative or technical specifics such as exact dates, numerical values,  
1377 specifications, or code snippets (53.8% vs 41%) (p = 0.26).

1378 Search vs non-search model responses to *Info Synthesis* prompts:

1379

- 1380 • Incorporates precise data, statistics, dates, names, and domain-specific terminology (80.1% vs  
1381 61.9%) (p = 0.18).

1382 Non-search vs search model responses to *Text Processing* prompts:

1383

- 1384 • Uses explicitly structured formatting (numbered or bulleted lists, headings) to organize  
1385 information (74.4% vs 56.4%) (p = 0.1).

### 1386 **Search Arena Setting**

1387 Search vs non-search model responses to *Factual Lookup* prompts:

1388

- 1389 • Provides extensive background context or explanatory preamble before addressing the specific  
1390 question (83.3% vs 72.2%) (p = 0.26).

1391 Search vs non-search model responses to *Info Synthesis* prompts:

1392

- 1393 • Explicitly states limitations, uncertainties, or caveats about the information provided (71.7% vs  
1394 52.2%) (p = 0.05).

1395  
1396  
1397  
1398  
1399  
1400  
1401  
1402  
1403

1404  
1405  
1406

1407  **LLM Prompt for Proposing Distinguishing Properties between Model A and Model B**  
 1408 **Responses**

1409  
1410  
1411

1412 The following is a list of Model A and Model B outputs to a set of user  
 1408 questions:  
 1409

1410 {text}  
 1411 <END>

1412  
1413  
1414

1415 I am a machine learning researcher trying to figure out the major differ-  
 1416 ences between the outputs of model A and model B. This is a very small  
 1417 portion of the data, so I want the differences to be general.

1418  
1419  
1420

1421 Please provide a list of three top differences (separated by bullet  
 1422 points "\*") between the outputs of model A and model B.  
 1423

1424 Follow the detailed instructions below:  
 1425

- 1426 - Identify properties that are more common in model A outputs compared to  
 1427 model B outputs.
- 1428 - The property descriptions should be simple and detailed. Do not pro-  
 1429 duce vague and generic descriptions.
- 1430 - Do not use keywords like "model A", "model B", "more common", "less  
 1431 common", "frequently", "occasionally" in your response.
- 1432 - Do not combine multiple properties / features in a single bullet point.
- 1433 - An example of the desired output format:  
 1434   \* "Property 1"  
 1435   \* "Property 2"  
 1436   \* "Property 3"

1437  
1438  
1439  
1440

1441 Please order your response in terms of the most common differences be-  
 1442 tween the two groups. Your response:  
 1443

1444

1445  **LLM Prompt for Ranking Properties**

1446  
1447  
1448

1449 **QUESTION:**  
 1450

1451 {question}  
 1452 <END>

1453  
1454  
1455

1456 **ANSWER:**  
 1457

1458 {answer}  
 1459 <END>

1460  
1461  
1462

1463 **PROPERTY:**  
 1464

1465 {hypothesis}  
 1466 <END>

1467  
1468  
1469

1470 Check whether the ANSWER satisfies the PROPERTY. Respond with Yes or No.  
 1471 If you are unsure, respond with No.  
 1472 Output:  
 1473

1474  
1475  
1476

1477 **D LEADERBOARD AND GENERAL FEATURE ANALYSIS**

1478  
1479  
1480

1481 **D.1 LEADERBOARD ANALYSIS**

1482  
1483  
1484

1485 Models supported on the Search Arena platform are shown in [Table A.1](#). Number of battles across models  
 1486 is shown in [Figure D.1](#); pairwise battle count is shown in [Figure D.2](#).

1487 Average win rates across models are shown in [Figure D.3](#). Pairwise win rates are shown in [Figure D.4](#).

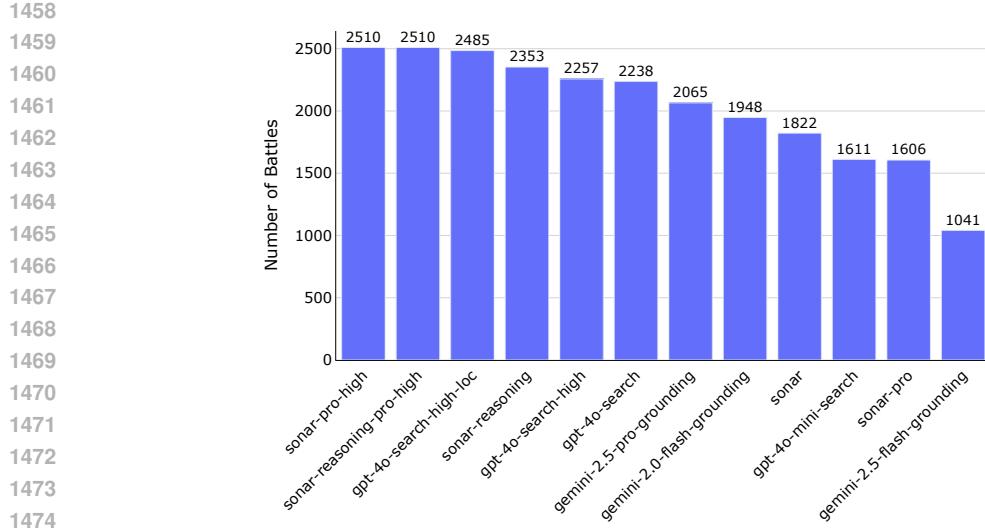


Figure D.1: **Battle Count Distribution.** Number of battles across Search Arena models. The distribution is not even due to (1) different sampling weights and (2) models were not added to the platforms at the same time.

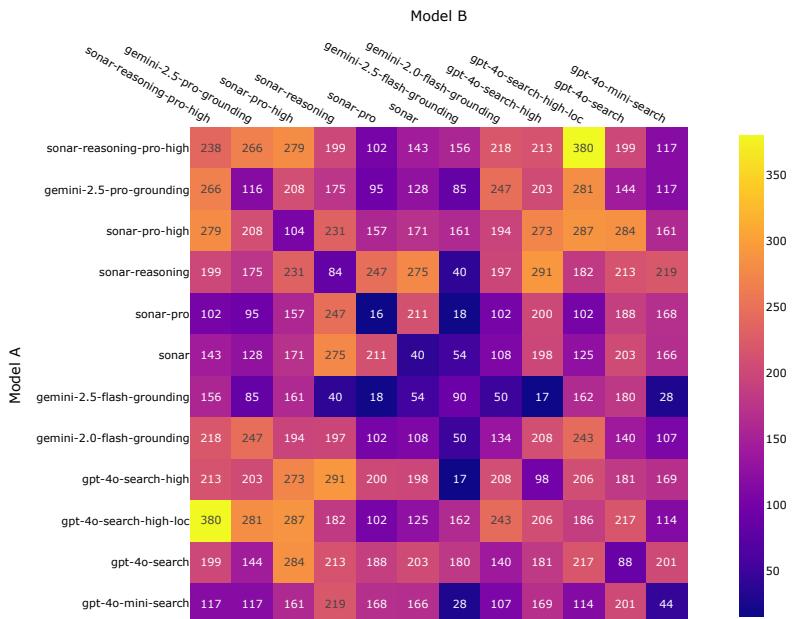
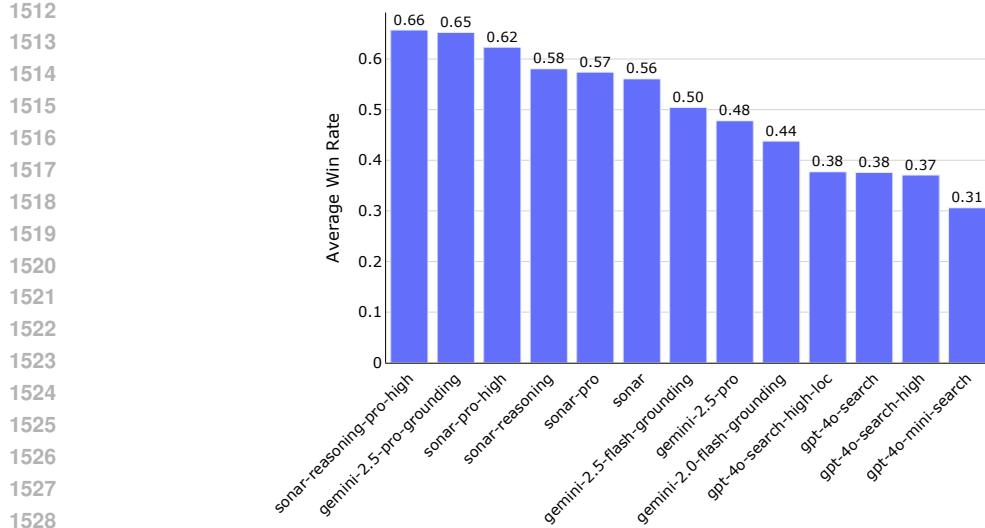
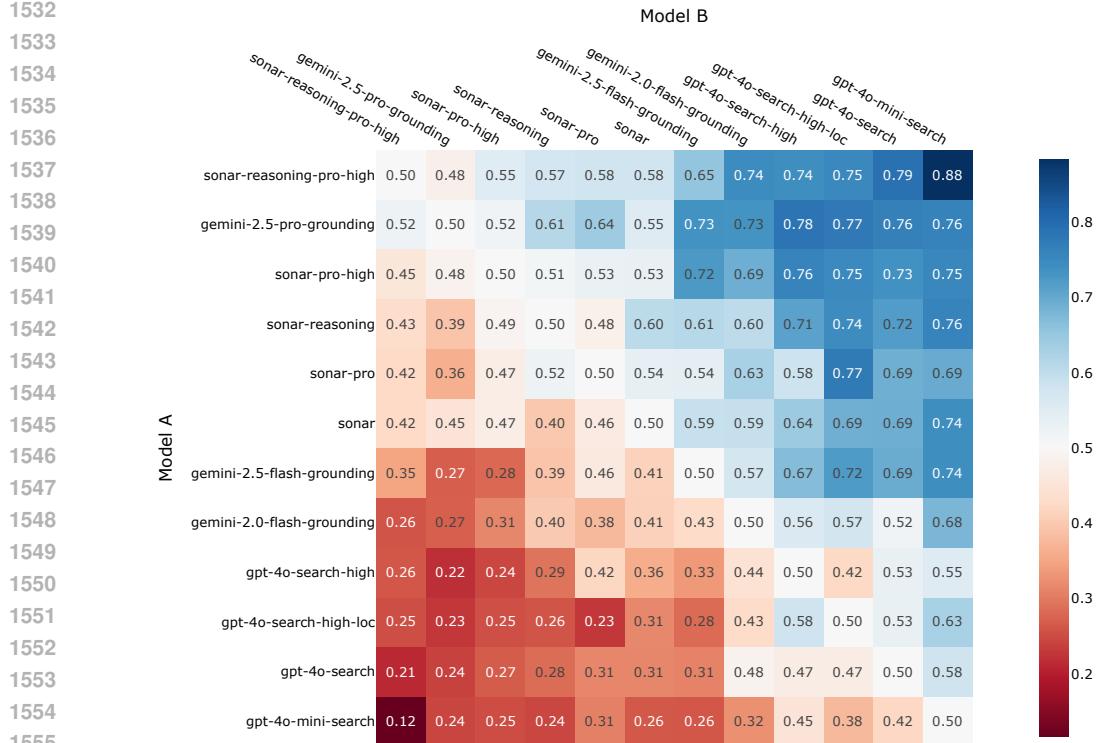


Figure D.2: **Pairwise Battle Count Distribution.** Number of battles between Search Arena models.

Consistent with Chatbot Arena’s ranking system (Chiang et al., 2024), we use Bradley-Terry regression (Bradley & Terry, 1952) to calculate model coefficients and then re-scale to match the scale of the Elo rating system. The resulting model scores are shown in Figure D.5.

Search Arena model ratings and ranks are shown in Table D.1. We also compute model scores and the leaderboard for two subsets of the full dataset:

- English and non-English prompts: The gap between the two top models increases, with `sonar-reasoning-pro-high` performing better on non-English prompts. Additionally, `gemini-2.0-flash-grounding` and `gemini-2.5-flash-grounding` perform better on non-English prompts.

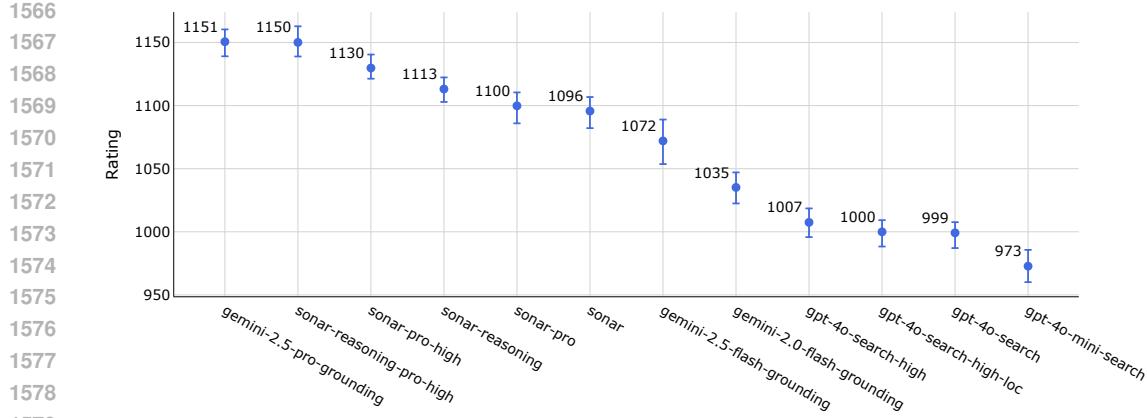
Figure D.3: **Win Rate Distribution.** Average win rates of Search Arena models.Figure D.4: **Pairwise Win Rates.** Pairwise win rates between Search Arena models.

1557  
1558  
1559  
1560  
1561  
1562  
1563  
1564  
1565

- Factual and non-Factual prompts: We observe a clear split in the leaderboard on the Factual subset. Additionally, the gap between the three top models (including sonar-pro-high) “zeros out” on the non-factual subset.

## D.2 RESPONSE LENGTH ANALYSIS

Average response length distribution across Search Arena models is shown in Figure D.6 (Left). Reasoning models tend to be more verbose except for sonar-reasoning. sonar-pro's version with a higher

Figure D.5: **Search Arena Leaderboard.** Model scores based on Elo-scaled Bradley-Terry coefficients.Table D.1: **Search Arena Leaderboard.** Elo-scaled Bradley-Terry ratings along with corresponding ranks (Rating (Rank)). Model ratings based on two subsets of the full dataset: (1) English vs non-English prompts, (2) Factual (*Factual Lookup* and *Info Synthesis*) vs non-Factual prompts.

Model	Rating	Rating (English)	Rating (non-English)	Rating (Factual)	Rating (non-Factual)
gemini-2.5-pro-grounding	<b>1150.6 (1)</b>	<b>1146.8 (1)</b>	1158.3 (1)	1155.8 (1)	1143.9 (1)
sonar-reasoning-pro-high	1150.1 (1)	1137.5 (1)	<b>1174.2 (1)</b>	<b>1161.4 (1)</b>	1142.5 (1)
sonar-pro-high	1129.8 (1)	1125.7 (1)	1136.2 (2)	1110.1 (3)	<b>1144.1 (1)</b>
sonar-reasoning	1113.2 (3)	1112.0 (2)	1117.9 (3)	1105.9 (3)	1121.0 (2)
sonar-pro	1099.9 (4)	1101.5 (3)	1093.1 (4)	1092.0 (3)	1106.8 (4)
sonar	1095.7 (4)	1092.8 (4)	1100.6 (3)	1085.6 (3)	1106.0 (4)
gemini-2.5-flash-grounding	1072.0 (5)	1061.4 (6)	1098.4 (3)	1095.0 (3)	1059.2 (7)
gemini-2.0-flash-grounding	1035.1 (8)	1017.8 (8)	1060.7 (5)	1035.5 (8)	1035.4 (7)
gpt-4o-search-high	1007.5 (9)	1000.9 (8)	1018.7 (8)	1007.7 (8)	1005.6 (8)
gpt-4o-search-high-loc	999.9 (9)	989.9 (8)	1019.7 (9)	987.1 (9)	1009.4 (8)
gpt-4o-search	999.2 (9)	999.4 (8)	1002.1 (9)	1000.5 (8)	999.5 (9)
gpt-4o-mini-search	972.7 (12)	970.7 (9)	973.6 (11)	973.1 (9)	969.6 (11)

search context generates longer responses compared to the version with a medium context. Response length distribution is uniform across OpenAI’s models.

We control for response length difference in the Bradley-Terry model and compute the corresponding coefficient across different subsets of the full data, broken down by intent category (see Figure D.7). For all intent categories, the coefficient (effect) is statistically significant; however, as expected, the effect is smallest for *Factual Lookup* prompts.

## E CITATION ANALYSIS

### E.1 CITATION COUNT

Citation count distribution is shown in Figure D.6 (Right). As expected, models with higher search context size cite more sources (e.g., sonar-pro-high vs sonar-pro). Furthermore, reasoning models tend to cite less sources compared to non-reasoning variants (e.g., sonar-reasoning-pro-high vs sonar-pro-high, sonar-reasoning vs sonar). We hypothesize that reasoning models synthesize and filter irrelevant sources before final response generation, resulting in less cited sources in the final response (see Figure 3). Interestingly, gemini-2.5-pro-grounding cites fewer sources compared to gemini-2.5-flash-grounding, suggesting that even though both are reasoning models, gemini-2.5-pro-grounding filters out more sources from the final response.

We then control for citation count difference in the Bradley-Terry model and compute the corresponding coefficient across different subsets of the dataset broken down by intent category (see Figure E.1). The effect of citation count on *Guidance* (e.g., debugging, problem-solving) prompts is not significant. The effect is largest on *Analysis* prompts.

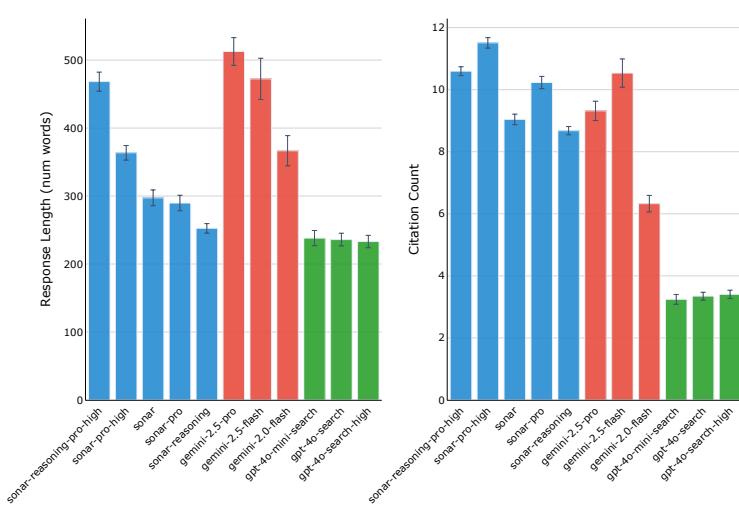


Figure D.6: **(Left)** Response length distribution. Reasoning models and models with high search context size tend to be more verbose. **(Right)** Citation count distribution. As expected, models with higher search context size cite more sources. Reasoning models cite fewer sources compared to non-reasoning variants (e.g., sonar-reasoning-pro-high vs sonar-pro-high).

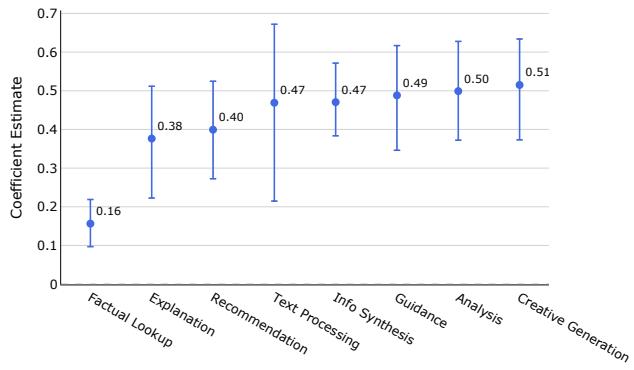


Figure D.7: **Response Length Control across Intents.** Bradley-Terry coefficients corresponding to response length across different intent categories. Length has less effect on *Factual Lookup* prompts.

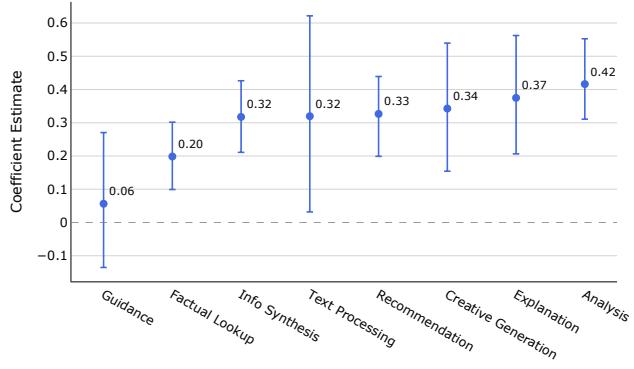


Figure E.1: **Citation Count Control across Intents.** Bradley-Terry coefficients corresponding to citation count across different intent categories. Citation count does not have significant effect on *Guidance* (e.g., debugging, problem-solving) prompts. The effect is largest for *Analysis* queries.

1674

## E.2 CITATION SOURCES

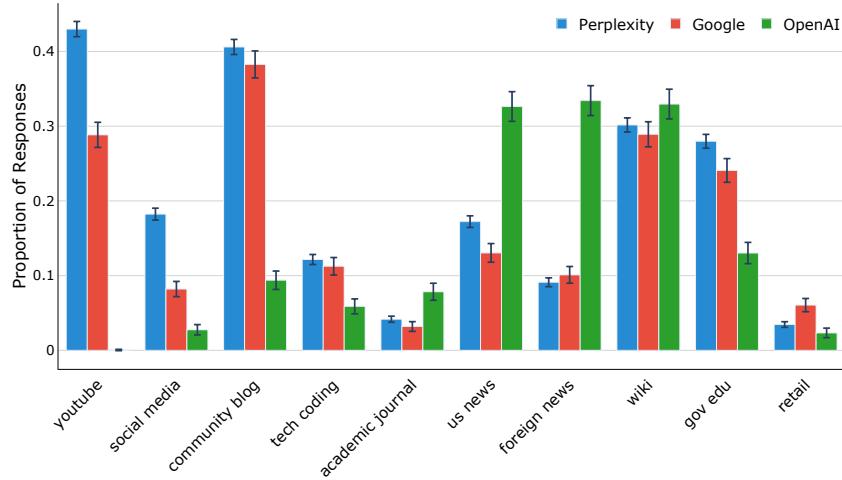
1675

1676 To categorize citation sources, we use the following mapping from source domains to categories:

1677

- **youtube**: “youtube.com”.
- **gov\_edu**: “.gov”, “.edu”, “.mil”.
- **wiki**: “wikipedia”, “wikihow”, “wikimedia”.
- **us\_news**: “cnn.com”, “apnews.com”, “cnbc.com”, “bloomberg.com”, “economist.com”, “nytimes.com”, “washingtonpost.com”, “wsj.com”, “nbcnews.com”, “abcnews.go.com”, “usatoday.com”, “npr.org”, “latimes.com”, “vox.com”, “huffpost.com”, “ft.com”, “foxnews.com”, “axios.com”, “time.com”, “buzzfeed.com”, “cbsnews.com”, “politico.co”, “newsweek.com”, “fortune.com”, “theatlantic.com”, “whattowatch.com”, “scrippsnews.com”, “investopedia.com”, “yahoo.com”, “breitbart.com”, “washingtontimes.com”, “dailycaller.com”, “thefederalist.com”, “townhall.com”, “pjmedia.com”, “westernjournal.com”, “forbes.com”.
- **foreign\_news**: “reuters.com”, “bbc.com”, “aljazeera.com”, “dw.com”, “france24.com”, “as.com”, “elpais.com”, “cbc.ca”, “theglobeandmail.com”, “smh.com.au”, “abc.net.au”, “japantimes.co.jp”, “straitstimes.com”, “hindustantimes.com”, “thehindu.com”, “economictimes.indiatimes.com”, “indianexpress.com”, “independent.co.uk”, “theguardian.com”, “cadenaser.com”, “lemonde.fr”, “vnexpress.net”, “ndtv.com”.
- **social\_media**: “tiktok.com”, “facebook.com”, “instagram.com”, “x.com”, “twitter.com”, “linkedin.com”, “snapchat.com”, “pinterest.com”.
- **community\_blog**: “reddit.com”, “quora.com”, “blog”, “medium.com”, “wordpress.com”, “substack.com”, “tumblr.com”.
- **tech\_coding**: “github.com”, “gitlab.com”, “stackexchange.com”, “microsoft.com”, “dev.to”, “codecademy.com”, “stackoverflow.com”.
- **academic\_journal**: “jstor.org”, “springer.com”, “sciencedirect.com”, “nature.com”, “arxiv.org”, “researchgate.com”, “biorxiv.org”.
- **retail**: “amazon.com”, “ebay.com”, “walmart.com”, “target.com”, “bestbuy.com”, “costco.com”.
- **other**: non-matched domains.

1704



1720

Figure E.2: **Citation Source Distribution across Model Families.** Models from different providers are biased towards different types of sources: (1) Perplexity’s models prefer citing YouTube, social media, and community blogs, (2) OpenAI’s GPT-4o-based models are more biased towards mainstream news outlets. Google’s Gemini models are in between. To further understand the effect of reasoning, we analyze the distributional differences of the cited sources by *sonar-pro-high* (non-reasoning) and *sonar-reasoning-pro-high* (reasoning), as reported

1721

1722

1723

1724

1725

1726

1727

The distribution of citation domain categories is shown in Figure E.2. Perplexity’s models tend to cite social media (including YouTube) and community platforms (e.g., Reddit, Quora). OpenAI’s GPT-4o-based models are more biased towards mainstream news outlets. Google’s Gemini models are in between. To further understand the effect of reasoning, we analyze the distributional differences of the cited sources by *sonar-pro-high* (non-reasoning) and *sonar-reasoning-pro-high* (reasoning), as reported

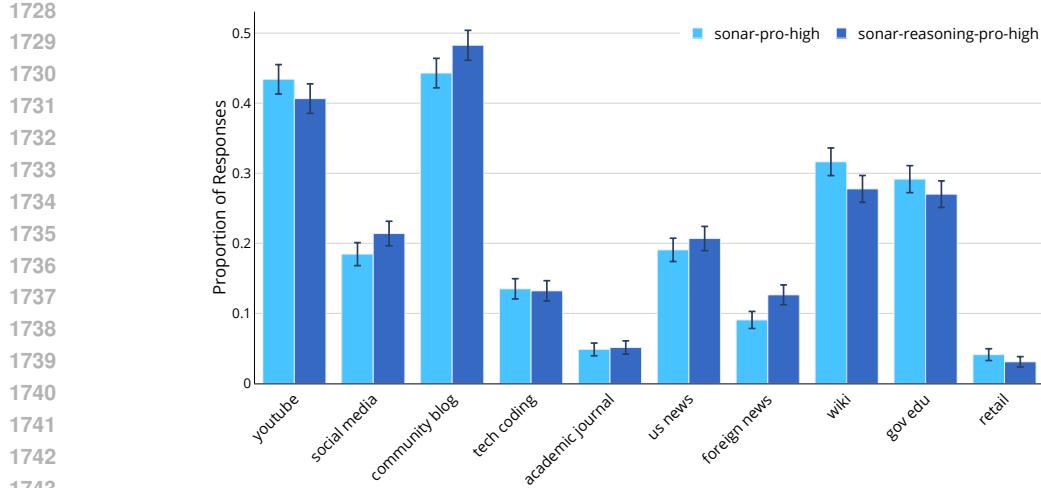


Figure E.3: **Citation Source Distribution across Non-reasoning (sonar-pro-high) and Reasoning (sonar-reasoning-pro-high) Models.** Although both systems are based on the same search module, the reasoning model (dark blue) cites more community and blog sources and less Wikipedia articles compared to the non-reasoning variant (light blue).

We find that even though both systems are based on the same search module, the reasoning model cites more community and blog sources and less Wikipedia articles, compared to the non-reasoning model. These results are consistent with our correlational analysis of user preferences and cited domains (e.g., users rejecting Wikipedia sources in specific settings).

We can also control for all features simultaneously (response length, number of citations, and citation sources) to compute adjusted model scores. The scores and rankings before and after applying these controls are shown in Figure E.4. We observe that the model scores and rankings tend to converge after the controls are applied. This convergence is particularly evident within model families: the top three models from Perplexity show significant convergence. This suggests that much of the variation between models in the same family is accounted for by the control features. We note that features such as factuality and coverage may be correlated with the number of citations or specific citation sources; we leave this analysis to future work.

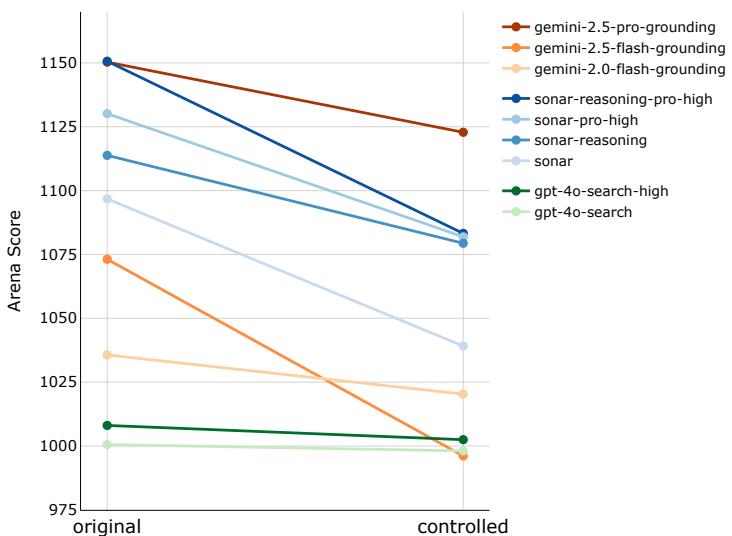


Figure E.4: **Model Scores Before and After Control.** Model scores and rankings converge after the controls are applied.

1782  
1783

## E.3 CITATION ATTRIBUTION ANALYSIS

1784  
1785

As described in [Subsection 3.2](#) and illustrated in [Figure 7](#), we analyzed citation attribution on approximately 100 conversations per intent category. This section outlines the full pipeline.

1786  
1787  
1788  
1789  
1790  
1791  
1792  
1793

We first sampled around 100 conversations from each intent category, excluding the “Other” class. To retrieve the cited content, we used Firecrawl<sup>1</sup> as our default batch scraping tool. For social media platforms or other domains where Firecrawl could not be applied, such as Reddit and YouTube, we leveraged official APIs or extracted video transcripts when available. For the remaining URLs that failed due to technical limitations, access restrictions, or licensing concerns, we excluded the corresponding samples. Overall, we took care to ensure ethical data usage, relying only on publicly accessible content and following best practices regarding rate limits and terms of service. In total, we collected over 20,000 web documents across 780 conversations.

1794  
1795  
1796  
1797  
1798  
1799

Next, we used `gemini-2.5-flash-preview-0417` to analyze the data, selected for its strong performance on long-context reasoning, reliable structured (JSON) outputs, and reasonable cost. For each conversation, we first used the model to parse user messages into `(claim, URL)` pairs, following the structure shown in [Figure 7\(a\)](#). Then, for each pair, we provided the scraped markdown content and the associated claim to infer attribution judgments. Each attribution was labeled as `Support`, `Irrelevant`, or `Contradict`. The prompt used for LLM-based tagging is shown on the next page.

1800  
1801  
1802  
1803

As illustrated in [Figure 7\(b\)](#), we aggregated the number of supporting, irrelevant, and contradicting claims per turn and used these counts as features in the Bradley-Terry model, following the same setup as the other control experiments. Human experts validated a subset of the outputs. Future work can further scale this pipeline or add onto analysis methodology.

1804  
1805
 **LLM Prompt for Claim–Citation Attribution Assessment**
1806  
1807  
1808  
1809  
1810  
1811  
1812  
1813

Task: Tell me if the claim can be verified or supported by the web content.  
 Specifically:  

- If the claim is supported by the web content, return "support"
- If the claim is contradict to the web content, return "contradict"
- If the claim is completely irrelevant to the web content, return "irrelevant"

1814  
1815  
1816  
1817  
1818  
1819  
1820  
1821  
1822

Claim: {claim}  
 Web content: {web content}  
 Return in the json format:  
 {  
     "reasoning": "<explain the reasoning>",  
     "answer": "<support" | "contradict" | "irrelevant">  
 }

1823  
1824**F CROSS-SETTING ANALYSIS**1825  
1826  
1827

In this section, we extend our analysis of [Subsection 3.3](#) and study model performance changes across different benchmarks. Specifically, we explore the following research question: *How does model performance and ranking change across different search and non-search benchmarks?*

1828  
1829  
1830  
1831  
1832  
1833

For offline evaluation of models in search settings, we use `BrowseComp` ([Wei et al., 2025](#)) and `SimpleQA` ([Wei et al., 2024](#))<sup>2</sup>. For testing the general performance of the models in non-search settings, we use `ArenaHard-v2` ([Li et al., 2024](#)). `BrowseComp` and `SimpleQA` assess factuality based on well-specified ground truths, while `ArenaHard-v2` utilizes an LLM-as-a-judge framework on a set of challenging prompts filtered from Chatbot Arena’s Text Arena dataset. Although these benchmarks test different aspects of

1834  
1835

<sup>1</sup><https://www.firecrawl.dev>

<sup>2</sup>Due to the high cost of running search-augmented LLMs, we evaluated each model on the same randomly sampled subset of 500 questions. The subset was sampled once and shared across all models.

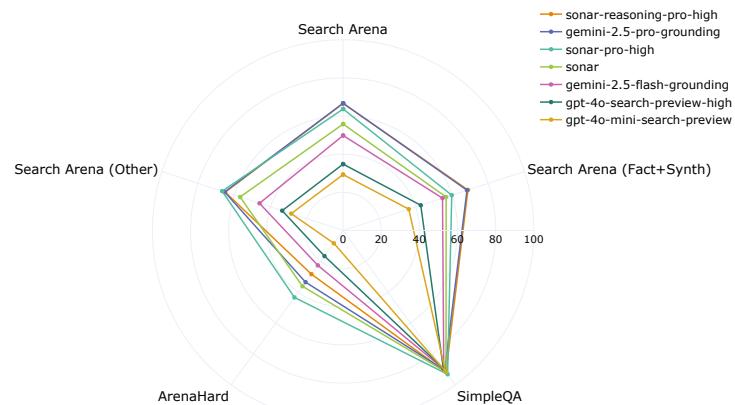
standard and search-augmented LLMs, all are partially captured by the Search Arena prompt distribution shown in [Figure 1](#).

The selected models have near-zero accuracy on the BrowseComp benchmark; this finding is not surprising, as the questions are specifically designed to evaluate and challenge deep research pipelines. The model scores for the rest of the benchmarks—Search Arena, SimpleQA, and ArenaHard-v2—are provided in [Table F.1](#). Additionally, we calculate model win rates on two subsets of the Search Arena dataset based on the annotated intent classes (see [Section 2](#))—Search Arena (Fact+Synth) includes only *Factual Lookup* and *Info Synthesis* queries, while Search Arena (Other) contains the remaining subset. We use Kendall’s tau for comparing ranks and Pearson correlation for comparing raw scores across benchmarks. On SimpleQA, model accuracy is saturated (ranging from 89% to 93%), with minimal separability between the models and low agreement with Search Arena ( $\tau = 0.422$ ,  $r = 0.582$ ). In contrast, ArenaHard-v2 shows greater performance variance and separability, and has a higher agreement ( $\tau = 0.556$ ,  $r = 0.844$ ) with Search Arena compared to SimpleQA. However, the rankings differ; notably, the three reasoning models rank lower in ArenaHard-v2, suggesting that, while combining web search with reasoning improves performance in Search Arena, it may degrade performance on ArenaHard-v2 prompts. Additionally, when comparing the two subsets of the Search Arena with ArenaHard, Search Arena (Other) has higher agreement ( $\tau = 0.644$ ,  $r = 0.882$ ) compared to Search Arena (Fact+Synth) ( $\tau = 0.511$ ,  $r = 0.675$ ). This finding is expected as the prompt distribution in Search Arena (Other) is closer to that of ArenaHard (e.g., creative writing, problem-solving).

Furthermore, we compare model performance with and without search on the SimpleQA and ArenaHard-v2 benchmarks. We used Gemini models in this case study, as search is implemented as a tool (GoogleSearch tool) and can be easily turned on and off. The performance change is shown in [Figure F.2](#). Search significantly improves model performance on SimpleQA (models’ performance saturates at around 90%). However, performance degrades on ArenaHard-v2, with non-search variants achieving higher scores.

[Table F.1](#): Model scores across three benchmarks—Search Arena (win rate), SimpleQA (accuracy), and ArenaHard-v2 (win rate). On SimpleQA, models’ performance is saturated and lead to minimal separability in model scores. Search Arena and ArenaHard-v2 provide more separability, but the rankings are different.

Model	Search Arena	Search Arena (Fact.+Synth.)	Search Arena (Other)	SimpleQA	ArenaHard-v2
sonar-reasoning-pro-high	<b>66.6</b> (-2.1 / +2.2)	<b>68.8</b> (-3.3 / +3.4)	65.0 (-3.1 / +2.7)	91.8 (-2.4 / +2.4)	28.3 (-1.7 / +2.1)
gemini-2.5-pro-grounding	<b>66.7</b> (-2.5 / +2.4)	68.3 (-3.8 / +3.7)	65.3 (-3.2 / +3.2)	90.6 (-2.6 / +2.4)	33.5 (-2.1 / +2.4)
sonar-pro-high	63.7 (-2.2 / +2.4)	59.9 (-3.5 / +3.3)	<b>66.7</b> (-2.8 / +2.9)	<b>93.2</b> (-2.2 / +2.2)	<b>43.4</b> (-2.1 / +2.2)
sonar-reasoning	60.1 (-2.5 / +2.3)	58.8 (-3.4 / +3.3)	61.3 (-3.2 / +3.4)	92.6 (-2.4 / +2.2)	29.1 (-2.4 / +2.3)
sonar-pro	57.5 (-3.1 / +2.8)	56.8 (-4.3 / +4.2)	58.4 (-3.9 / +4.0)	92.6 (-2.4 / +2.2)	39.9 (-2.2 / +2.2)
sonar	55.8 (-2.7 / +2.8)	54.8 (-4.1 / +3.8)	56.8 (-3.9 / +3.8)	91.8 (-2.4 / +2.4)	36.2 (-1.9 / +2.1)
gemini-2.5-flash-grounding	49.8 (-3.7 / +3.5)	54.5 (-5.9 / +5.8)	46.1 (-4.8 / +4.7)	89.8 (-2.6 / +2.6)	22.6 (-1.4 / +1.8)
gemini-2.0-flash-grounding	41.7 (-2.8 / +2.9)	42.8 (-4.2 / +4.2)	41.0 (-3.5 / +3.4)	89.2 (-2.8 / +2.6)	13.4 (-1.0 / +1.1)
gpt-40-search-preview-high	34.8 (-2.5 / +2.5)	36.2 (-3.6 / +3.8)	33.6 (-3.5 / +3.2)	89.4 (-2.8 / +2.6)	16.6 (-1.2 / +1.2)
gpt-40-mini-search-preview	29.3 (-2.7 / +2.6)	30.3 (-3.7 / +4.2)	28.6 (-3.6 / +3.6)	91.2 (-2.6 / +2.4)	8.3 (-0.6 / +0.7)



[Figure F.1: Model Scores Across Benchmark.](#) (1) Models’ performance on SimpleQA is saturated with minimal separability. (2) Performance variance in ArenaHard-v2 is comparable to that of Search Arena, but the rankings are different, with reasoning models having higher performance in search settings. (3) Model scores and ordering in Search Arena (Other) is closer to that of ArenaHard compared to Search Arena (Fact+Synth).

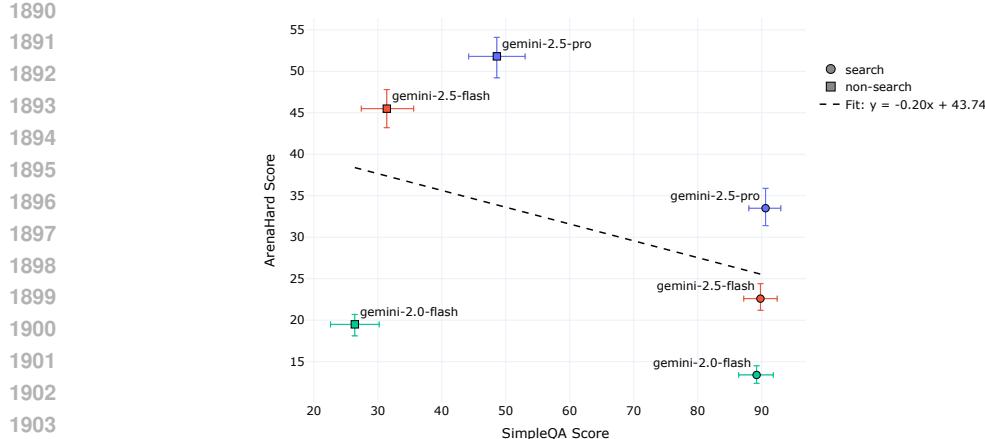


Figure F.2: **SimpleQA vs ArenaHard-v2 Performance.** Search and non-search performance of Gemini models on SimpleQA and ArenaHard-v2 benchmarks. Search improves performance on SimpleQA, while degrades the score on ArenaHard-v2.

## G ADDITIONAL COMPARATIVE ANALYSES

### G.1 CORAL

As our setup is similar to prior work in conversational search, we conduct additional comparative analysis with CORAL (Cheng et al., 2025a), a benchmark designed to assess RAG systems in multi-turn conversational settings. Compared to CORAL, Search Arena conversations are collected from natural open-ended interactions between humans and search-augmented LLMs and contain user-provided preference labels.

Similar to the datasets studied in the main paper, CORAL focuses on English-only conversations, whereas the Search Arena prompts cover a diverse set of 71 languages. Search Arena (24k) also is larger in scale compared to CORAL (8k). Furthermore, in contrast to SimpleQA (Wei et al., 2024) and BrowseComp (Wei et al., 2025) and similar to Search Arena, CORAL includes multi-turn conversations. Compared to Search Arena (Figure B.3), CORAL conversations are longer on average (Figure G.1, Left). In terms of prompts, CORAL focuses on information-seeking queries only, whereas Search Arena covers a broader range of intents (e.g., factual lookup, creative generation). The prompt length distribution of Search Arena prompts is also more varied compared to CORAL, with the latter containing more concise questions (see Figure G.1, Right).

Additionally, in contrast to Search Arena, CORAL explicitly constructs ground truth answers, assuming a bounded set of documents (including golden). Search Arena focuses on human preferences, which, as observed in Section 3, do not need to correlate with factuality and quality of the retrieval. Finally, Search Arena data is collected from real-world natural interactions, enabling us to study real-world usage of the models and analyze how model response features correlate with the users’ preferences. Thus, the differences between the two datasets are due to the differences in the corresponding intents, use cases, and motivations.

### G.2 WILDCHAT

Similar to the Text Arena and Search Arena datasets studied in the main paper, WildChat (Zhao et al., 2024) is a crowd-sourced human-LLM conversation dataset. Similar to the Text Arena and in contrast to the Search Arena, the interactions in WildChat are with regular LLMs without access to additional tools, such as web search.

Despite the differences in the settings, we compare WildChat with Search Arena across multiple axes; we use a random sample (100k) of the non-toxic filtered version (3.2M) of the original dataset for the comparative analysis below. First, the full WildChat dataset is much larger in scale (4.8M), compared to Search Arena (24k). Both datasets are multilingual, containing interactions with over 70 languages. Turn and prompt length distributions of the WildChat dataset are provided in Figure G.2. WildChat turn count distribution is similar to that of Search Arena, with Search Arena containing slightly more multi-turn conversations. Additionally, WildChat prompts are longer on average (594.9) compared to those of Search Arena (571). This difference arises from the difference in the user intents across the two datasets: as reported in the Text Arena comparative analysis, search-enabled settings significantly impact the intent of the queries (e.g., users more likely to ask real-time information-seeking questions).

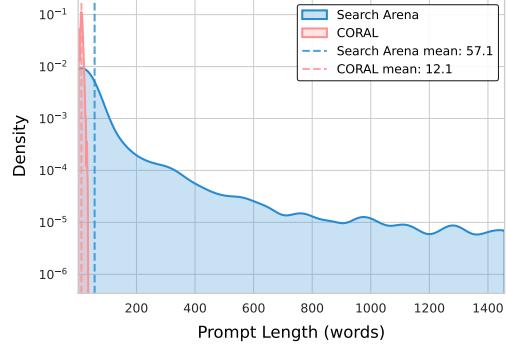
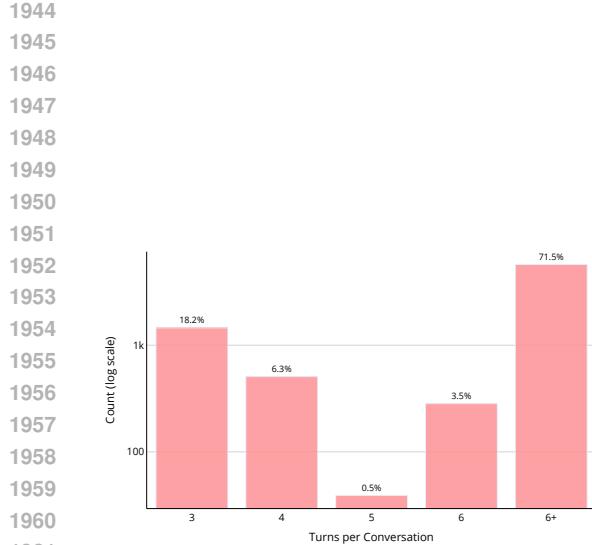


Figure G.1: **(Left)** CORAL conversation length (number of turns) distribution; CORAL conversations are longer than those in Search Arena, with 8.3 turns on average. **(Right)** CORAL prompt length distribution; CORAL prompts (red) are more concise and less diverse (lower variance) compared to Search Arena prompts (blue).

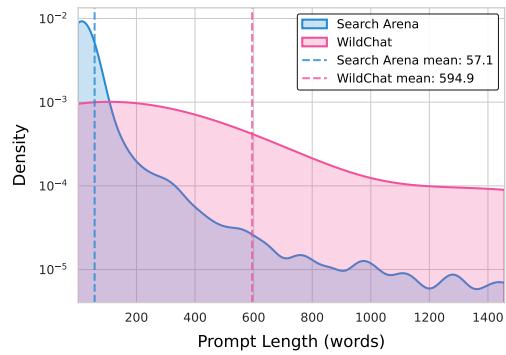
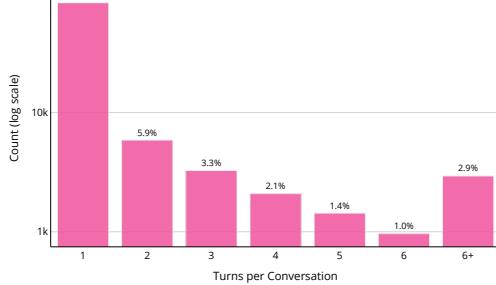


Figure G.2: **(Left)** WildChat conversation length (number of turns) distribution; WildChat conversations are similar in distribution to Search Arena Figure B.3, with slightly fewer multi-turn samples. **(Right)** WildChat prompt length distribution; WildChat prompts (pink) are longer compared to those of Search Arena (blue).