Search Arena Meets Nuggets: Towards Explanations and Diagnostics in the Evaluation of LLM Responses

Anonymous ACL submission

Abstract

Battles, or side-by-side comparisons in so-called arenas that elicit human preferences, are used to assess the large language model (LLM) output quality, and have recently been extended to retrieval-augmented generation (RAG) systems. Although, battles mark progress in evaluation, they have two key limitations for complex information-seeking queries: they are neither explanatory nor diagnostic. On the other hand, nugget-based evaluation, that decomposes long-form answers into atomic facts and highlights necessary parts in an answer, has emerged as a promising strategy for RAG evaluation. In this work, we employ AutoNuggetizer, a nugget-based framework, to analyze ∼5K Search Arena battles from LMArena by automatically generating and assigning nuggets, converting each model response into a quantitative score. We observe strong alignment between nugget-based Elo rankings and human preferences, with Kendall's τ of 0.71 and Spearman's ρ of 0.88, exceeding the corresponding alignment achieved by LLM-as-a-judge evaluation (0.64 and 0.79, respectively), while substantially reducing the number of preference inversions. Furthermore, we provide in-depth analyses including inversions, nugget quality, and shared-blindness effects. All our code and datasets will be released publicly upon paper acceptance.

1 Introduction

The notion of "battles", or side-by-side comparisons of responses from large language models (LLMs), has become a popular method for evaluating their quality (Zheng et al., 2023; Chiang et al., 2024). In the "arena" setup, users are shown two LLM outputs and asked to indicate which one they prefer. This approach was popularized by LMSYS through MT-Bench (Zheng et al., 2023) and later expanded into the Chatbot Arena (Chiang et al., 2024). The popularity of these arenas has made them a key marketing tool when launching new LLMs from companies such as Google, OpenAI, and Meta, who regularly tout leaderboard rankings on Chatbot Arena in model releases. Recently, arena-based evaluations have been extended to a variety of domains, including AI agents (Yekollu et al., 2024), vision and image generation (Lu et al., 2024; Jiang et al., 2024), multilingual generation (Thakur et al., 2025a), and even GitHub pull requests (Wang et al., 2025).

Battles were extended to search-augmented LLMs in the Search Arena (Miroyan et al., 2025). Unlike the original setup, which focused on "closed-book" LLM responses, Search Arena evaluates retrieval-augmented generation (RAG) systems in two stages: first, retrieving relevant web-sourced documents, and next using them to generate long-form answers with citations using LLMs (Pradeep et al., 2025a; Han et al., 2024).

While such side-by-side comparisons enable the evaluation of search-augmented LLM-based systems at scale, we see them having at least two drawbacks: they are neither explanatory nor diagnostic, especially in scenarios where determining the better answer is not straightforward. It would be desirable for an evaluation to (at least attempt to) explain *why* a user might have preferred one response over another.

We hypothesize that the nugget evaluation methodology (Pradeep et al., 2024, 2025b) can be adapted to address these two limitations for complex information-seeking queries. The core idea is to measure answer quality based on the recall of information nuggets, or atomic facts, that should appear in high-quality responses. In AutoNuggetizer (Pradeep et al., 2024, 2025b), this process can be fully automated using LLMs, breaking down a long-form model response into a quantitative score.

In our work, we adapt the AutoNuggetizer

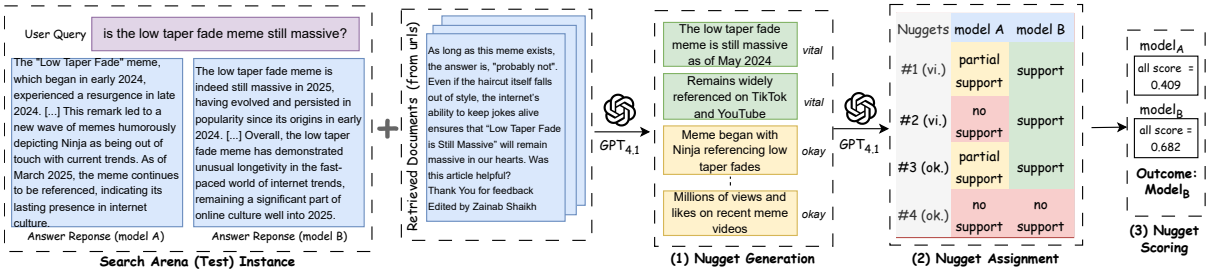


Figure 1: An example from Search Arena illustrating both nugget generation and assignment. First, $GPT_{4.1}$ generates nuggets based on the query, retrieved chunks from URL contents, and the responses from both models (A and B). Each nugget is then labeled with an importance level—either "vital" or "okay". Next, $GPT_{4.1}$ evaluates whether each model supports each nugget, assigning one of three labels: "support", "partial support", or "no support". Finally, these support judgments are scored and aggregated to determine the outcome (the model with the higher score is preferred).

(Pradeep et al., 2024) framework to analyze approximately 5K battles in the Search Arena in a fully automatic manner (see Figure 1), eliminating the need for cumbersome human judgments. The framework includes two stages: (1) *nugget generation*: eliciting nuggets from model answers and scraped documents from cited URLs, and (2) *nugget assignment*: evaluating whether each answer supports a nugget or fact. Our results show that human preferences correlate well with the distribution of nugget scores, achieving a very high alignment between Elo rankings of nugget-based evaluation and human preferences (Kendall's τ of 0.71 and Spearman's ρ of 0.88). Furthermore, our extended analysis diagnose the score inversions, nugget quality and potential shared-blindness in our evaluation setup. To summarize, our contributions are as follows:

- **Extending AutoNuggetizer to an arena setting with live search.** We adapt Search Arena to AutoNuggetizer by scraping and chunking the dataset URLs to form a retrieval corpus, then apply nuggetization to head-to-head battles where LLMs have live web-search access—moving beyond prior nugget evaluations that assume a fixed/static corpus.
- **Rigorous analysis of nuggetization and preference inversions.** We audit nugget factuality and diversity (lexical and semantic) and dissect "preference inversions," showing how query type (e.g., ambiguous vs. factoid) and language systematically affect nugget-based preferences;
- **Comparisons to alternative evaluators.** We benchmark nugget-based preferences against LLM-as-a-judge, surface-form metrics (e.g., ROUGE-style overlap) and judging with factors like fluency and generation style, highlighting when each method succeeds or fails.

2 Related Work

Nugget-based evaluation. First introduced in the TREC QA Track in 2003 (Voorhees, 2003a,b), the nugget-based evaluation methodology focuses on identifying essential atomic facts—called nuggets—that are relevant to a given question. This methodology was later extended to tasks like summarization and broader conceptions of question answering (Nenkova and Passonneau, 2004; Lin and Demner-Fushman, 2006b; Dang and Lin, 2007; Lin and Zhang, 2007), and researchers have explored automation to improve its scalability (Lin and Demner-Fushman, 2005, 2006a; Pavlu et al., 2012).

The recent emergence of LLMs has enabled automated, reliable nugget-based evaluation (Pradeep et al., 2024; Alaofi et al., 2024; Pradeep et al., 2025b; Thakur et al., 2025b; Abbasiantaeb et al., 2025). Several RAG evaluation frameworks—such as FactScore (Min et al., 2023), RUBRIC (Farzi and Dietz, 2024), and others (Arabzadeh and Clarke, 2024; Mayfield et al., 2024)—incorporate the nugget concept, although most of these proposed approaches are either not validated or primarily validated on traditional ad hoc retrieval, and hence their applicability to long-form answers is unclear. We refer readers to (Pradeep et al., 2025b) for a more detailed discussion of related work. In this work, we focus on the AutoNuggetizer framework from (Pradeep et al., 2024), and apply it to the side-by-side comparisons of LLM responses in the Search Arena.

Related arena benchmarks. The Search Arena (Miroyan et al., 2025) by LMArena is a

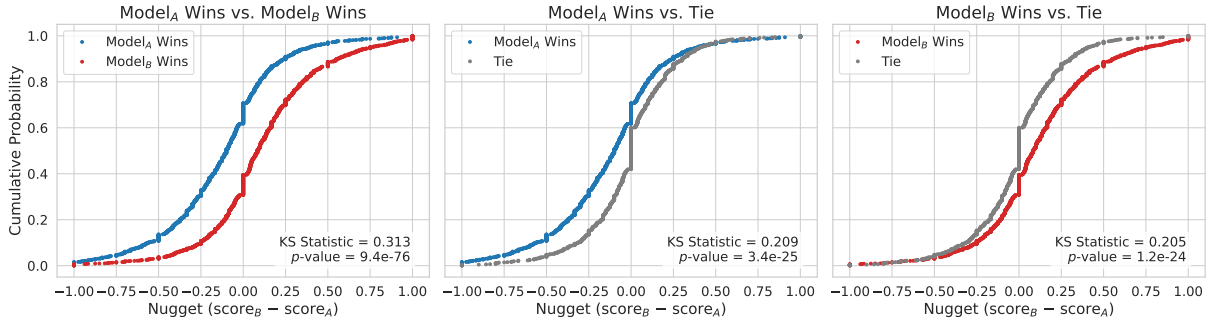


Figure 2: Empirical cumulative distribution functions (CDFs) comparing nugget score differences (score$_B$ − score$_A$) across human vote categories. Each subplot shows a Kolmogorov-Smirnov (K-S) test between two groups: (left) model$_A$ wins vs. model$_B$ wins, (center) model$_A$ wins vs. tie, and (right) model$_B$ wins vs. tie. The K-S statistic and corresponding p-value are annotated in each plot, quantifying the distributional differences between groups.

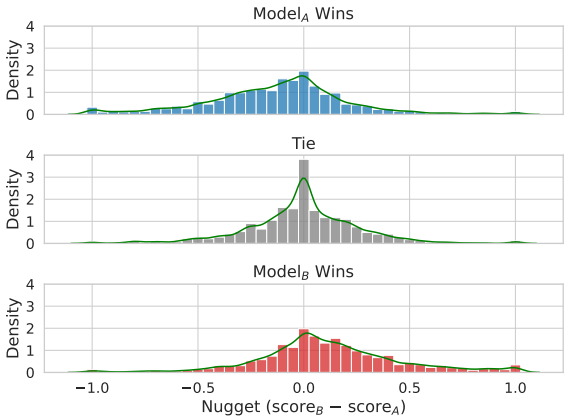


Figure 3: Empirical probability density function (PDF) of nugget score differences (score$_B$ − score$_A$) grouped by human preference category: model$_A$ wins, tie, or model$_B$ wins. A separate Kernel Density Estimation (KDE) with bandwidth 0.5 is fitted for each group.

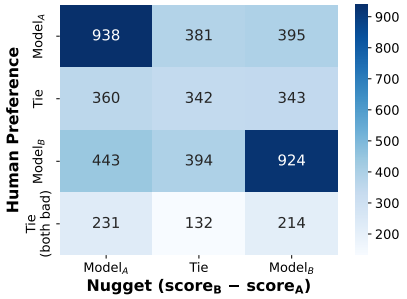


Figure 4: Confusion matrix comparing human and nugget preferences. A threshold of 0.07 is applied to treat nugget preference scores as a tie.

benchmark that evaluates LLMs with access to a live web-search tool. Other notable efforts include the MTEB Arena (Hugging Face, 2023), which extends the Massive Text Embedding Benchmark (MTEB) framework (Muennighoff et al., 2023) to head-to-head evaluation across embedding models, and Ragnarök (Pradeep et al., 2025a), which offered a head-to-head RAG evaluation framework on the MS MARCO V2.1 collection in the TREC 2024 RAG Track.

3 Experimental Design

Search Arena overview. Search Arena (Miroyan et al., 2025) is a crowd-sourced platform that evaluates search-augmented LLMs via side-by-side human-preference judgments (Chiang et al., 2024). The V1 dataset[1] includes 7K battles between two RAG-oriented systems (model$_A$ and model$_B$; e.g., Gemini-2.5-Pro-Grounding vs. Perplexity-Sonar-Reasoning-Pro). For each battle, annotators choose one of the four outcomes: model$_A$ wins, model$_B$ wins, good tie (both responses are equally good), or bad tie (both responses are equally bad). Search result URLs used during generation are available for ∼6.7K battles, totaling ∼80K unique URLs. The Search Arena dataset includes both single- and multi-turn battles. We restrict our analysis to single-turn battles only—5,103 instances where the system returns a single response—because overall votes in multi-turn settings do not reliably capture per-query preferences, which is what AutoNuggetizer evaluates.

Search Arena also contains battles for several non-English languages, e.g., Chinese or Russian. Non-English languages collectively account for less than 40% of the dataset, with English comprising the remaining majority. Detailed statistics for single-turn battles used in this work are presented in Appendix A.1. Queries in Search Arena vary widely, ranging from long code snippets to prompts that demand complex reasoning or exhibit ambiguity and vagueness. We show a few examples

[1] https://huggingface.co/datasets/lmarena-ai/search-arena-v1-7k

of queries from the Search Arena dataset in Appendix A.2.

Corpus generation. To evaluate LLM responses in the absence of ground-truth answers, we use the search result URLs provided in the dataset, collected from each system response as relevant sources of information. We begin by constructing a corpus from the 47K unique URLs associated with single-turn battles. This process involves downloading the contents of each URL, extracting the main textual content, and segmenting the text into chunks of ten sentences with an overlap of two sentences, using the `xx_sent_ud_sm` model from spaCy.[2]

Once the corpus is prepared, we encode the chunks and the query prompts utilizing the `BAAI/bge-m3` model.[3] We retrieve the top 50 most relevant chunks for each query via the cosine similarity between the chunk and query embeddings using Pyserini's FAISS indexing and search (Lin et al., 2021). Notably, both the chunking and encoding models support multilingual corpora, ensuring a robust language coverage.

Nugget evaluation. Nugget generation creates atomic facts that highlight the essential information required in a RAG answer, and assignment categorizes their support level for the model response. Following (Pradeep et al., 2024), we use the AutoNuggetizer tool in the `nuggetizer` code repository[4] to generate and assign information nuggets to model responses. As shown in Figure 1, there are two steps in nugget generation and assignment:

1. **Nugget generation:** For each prompt extracted from the dataset, we construct a request to AutoNuggetizer that includes the query (i.e., the prompt itself), along with relevant chunks retrieved from our created corpus, ordered by relevance and responses from each model, inserted in a random order to mitigate the positional bias. We include model responses for two key reasons. First, approximately 5% of the battles do not contain any URLs. Second, even when URLs are present, about 16% of them yield 100 bytes or less of content after scraping[5]. These cases make the LLM responses a valuable fallback source of information for nugget generation. The AutoNuggetizer tool then processes the request and identifies nuggets that are relevant to the query from the retrieved chunks and the provided LLM responses. Furthermore, each nugget is assigned an importance label: "vital" or "okay", reflecting its relevance to the input query.

2. **Nugget assignment:** Once nuggets and their importance labels are generated (from the previous step), we use AutoNuggetizer to assign them to model responses, determining whether each nugget is supported in the answer. This step categorizes each nugget into "supported", "partially supported", or "not supported". We adopt the "All Score" metric, that achieves the highest recall by counting nuggets of all importance and support levels[6]. We emphasize that while the AutoNuggetizer framework supports different degrees of manual intervention, in this work, we are running the entire evaluation pipeline end-to-end automatically.

4 Experimental Results

Unless stated otherwise, all experiments in this paper are conducted using $\text{GPT}_{4.1}$, with a knowledge cutoff of June 2024, as the underlying LLM used by the AutoNuggetizer via Microsoft Azure. Out of the 5,103 single-turn battles in Search Arena, five were excluded from our analysis due to issues such as Azure content filtering, invalid output formats, or other nugget generation failures. On average, each single-turn battle full evaluation (comprising both nuggetization and assignment) requires approximately 2–3 seconds when executed using the Azure OpenAI API. We set a maximum of 30 nuggets per battle, though this limit is rarely reached (only in 67 battles). When it is reached, only nuggets labeled as okay are removed, while no vital nuggets are discarded. On average, about $\sim$12.5 nuggets are generated per battle.

Main Results. Figure 3 presents our main results, the probability densities of nugget score differences ($\text{score}_B - \text{score}_A$) conditioned on the human preference judgment (i.e., the battle outcomes). On the top, we show the distribution when model_A

[2] https://spacy.io/models/xx#xx_sent_ud_sm
[3] https://huggingface.co/BAAI/bge-m3
[4] https://github.com/castorini/nuggetizer
[5] Invalid cases occurs due to issues such as cookie or JavaScript requirements, invalid or expired links, geo-blocking, and similar obstacles.

[6] We find that "Strict Vital", which was the primary metric used in the TREC 2024 RAG Track (Pradeep et al., 2025b), is too strict for our use case, particularly when only a small number of nuggets are available.

wins; on the bottom, we show the distribution when model$_B$ wins; and in the middle, ties. Battles where the output of both models is considered to be equally bad are excluded from the distributions.

These results appear to support our hypothesis that nugget scores correlate with human preferences. In the case where model$_A$ wins (top row), the distribution skews to the left (negative values), indicating that model$_A$ typically gets higher nugget scores than model$_B$. Conversely, when model$_B$ wins (bottom row), the distribution skews to the right (positive values), suggesting that model$_B$ generally obtains a higher nugget score. When the human indicates a tie (middle row), the distribution peaks around zero, as expected, indicating similar nugget scores between models.

Statistical Tests. To analyze the statistical differences among these three conditional distributions, we performed pairwise Kolmogorov-Smirnov (K-S) tests. As shown in Figure 2, the K-S statistic values range from 0.205 to 0.313, with p-values of $1.2\mathrm{e}^{-24}$ or lower, indicating that all three distributions differ significantly from one another (i.e., we have high confidence that these samples were drawn from different distributions). These findings validate our hypothesis that nugget score differences align with human preferences, reinforcing the potential of nugget-based metrics as reliable evaluators of model quality in head-to-head evaluations.

Confusion Matrix. Figure 4 presents a confusion matrix that compares the distribution of human preferences (rows) in Search Arena against "nugget preferences" (columns). For "nugget preference", we use a threshold of 0.07, meaning that when the nugget score difference between the two model outputs falls within $\pm$ 0.07, the comparison is considered a tie (The threshold was selected by sweeping values between 0.05 and 0.15 in increments of 0.01). The threshold of 0.07 closely reflects an equal distribution of model$_A$ wins, model$_B$ wins, and ties when the human preference is a tie (second row in Figure 4). The diagonal cells in this confusion matrix reveal the instances where nugget preferences align with human preferences. Conversely, off-diagonal cells illustrate the types and frequencies of disagreements between the human and nugget scores.

In particular, the nugget-based evaluation prefers model$_A$ in 938 out of 1,714 (54.7%) of the battles where model$_A$ wins the battle (first row in Figure 4). Similarly, model$_B$ is preferred in 924 out of 1761 (52.5%) battles where it wins the battle (third row in Figure 4). To further quantify this alignment, we report a weighted Cohen's κ of 0.30 with quadratic $(0, 0.25, 1)$ weights assigned to the (inversion, tie, identical) labels, respectively. This value remains stable across nugget score thresholds for ties, varying only slightly between 0.29 and 0.31 when thresholds range from 0.05 to 0.15.

Table 1: Inversion percentages and query frequencies across different (a) query categories and (b) languages in the Search Arena dataset.

Category	Inversion (%)	Query Count
(1) Ambiguous	19%	196
(2) Assumptive	18%	28
(3) Multi-faceted	18%	299
(4) Incompleteness	16%	631
(5) Subjective	15%	601
(6) Knowledge-int.	15%	1142
(7) Reasoning-int.	14%	288
(8) Harmful	9%	92

(a)

Language	Inversion (%)	Query Count
(1) German	20%	244
(2) English	17%	3117
(3) Chinese	16%	328
(4) Portuguese	16%	150
(5) Russian	15%	460
(6) French	13%	151
(7) Others	16%	647

(b)

Elo Rankings Given the high level of subjectivity inherent in the task, a moderate level of agreement at the individual battle level is expected. Consequently, the overall human- versus nugget-based rankings of the participating LLMs are of primary interest. Following established best practices in generative AI pairwise evaluation (Chiang et al., 2024; Jiang et al., 2024; Chi et al., 2025; Miroyan et al., 2025), we employ the Bradley–Terry (BT) model (Bradley and Terry, 1952), estimated via maximum likelihood estimation (MLE) using logistic regression with bootstrap resampling. Table 5 reports the resulting Elo scores for both human and nugget-based preferences. Kendall's τ of 0.71 and Spearman's ρ of 0.88 indicate strong agreement between the two Elo rankings.

In the remainder of this section, we analyze the anti-diagonal cases where nugget-derived and human preferences diverge; compare LLM-as-a-judge

preference agreement with human judgments as an alternative to nugget-based preferences; reassign nuggets using an alternative LLM; examine the generated nuggets; and explore other alternatives for nugget-based evaluation of open-ended generation.

4.1 Inversion Analysis

Query Classification Analysis. In this analysis, we use query classification to better understand the cases where nugget preferences and human preferences are not aligned. When the nugget scores and the human prefer opposite sides of a battle, we refer to this situation as a "preference inversion", or simply inversion. We suspect that inversions might vary across different types of queries. To investigate, we followed (Rosset et al., 2024) but used the newer $\text{GPT}_{4.1}$ to rate each query on a scale of 0–10 across eight different categories. Then, we classify each query into its maximum scoring category or categories (allowing for ties).

To further strengthen the category signals, we exclude queries with a maximum score less than seven from this classification. Raw distributions of the query ratings per category and sample queries from each class are available in Appendix A.2. As shown in Table 1 (row a), the portion of inversions for ambiguous, assumptive, multi-faceted, and incomplete queries is higher than that of subjective, knowledge-, and reasoning-intensive queries. This suggests that inversions are more likely when queries allow for multiple valid interpretations or are under specified.

We followed up by manually examining the inversions for these categories. As a case study, we encountered a query categorized as *ambiguous* with the text "Potatoes". In our opinion, both model_A and model_B provided relevant responses. However, model_A focused on the historical aspects and nutritional value of potatoes, whereas model_B discussed cooking methods and varieties. The user judge preferred model_B's answer, while model_A was selected based on the nugget score. The inherent ambiguity of the query likely led to this inversion, as it permitted various valid interpretations. Overall, the knowledge-intensive class shows the highest preference alignment—58.8% and 55.7% for model_A and model_B wins, respectively (see Figure 10). This finding suggests that nuggetization is most effective for researchy queries requiring retrieval augmentation. Please refer Appendix B.1 for further analysis on query classification.

Query Language Analysis. Next, we analyzed the AutoNuggetizer effectiveness across the six most frequent query languages, each representing at least 3% of the dataset. Previously, the AutoNuggetizer had only been run on English responses, and there are likely to be language effects in the breakdown of inversions. As shown in Table 1 (row b), German exhibits the highest inversion rate (20%), while French shows the lowest (13%). The confusion matrix for German (see Figure 11) reveals that it has the smallest portion of ties in human preferences, leading to more anti-diagonal disagreements. Please refer Appendix B.2 for further analysis on query languages.

4.2 LLM-as-a-Judge Evaluation

To analyze the correlation between human and LLM preferences, we experiment with $\text{GPT}_{4.1}$ as a judge. We modify the chain-of-thought prompt provided in (Rackauckas et al., 2024) (refer Appendix E). For each evaluation, we provide the user query along with the two model responses—randomly ordered to mitigate positional bias—as input to $\text{GPT}_{4.1}$. The model is instructed to output its reasoning and final verdict in a structured JSON format.

Figure 5 presents the confusion matrix comparing human preferences with those of the $\text{GPT}_{4.1}$ judge, which yields a weighted $(0, 0.25, 1)$ Cohen's κ of 0.31. Compared to the nugget-based evaluation in Figure 4, we observe stronger alignment between $\text{GPT}_{4.1}$ and human judgments for clear winners: 1,137 vs. 938 agreements for model_A, and 1,161 vs. 924 for model_B.

However, $\text{GPT}_{4.1}$ struggles to identify ties—including cases where both responses are poor—labeling only 4.25% of single-turn queries as ties. This narrow margin for tie predictions leads to a higher frequency of preference inversions in this setting, with 1,102 inversions compared to 817 under the nugget-based evaluation. The higher inversion rate, in turn, causes greater variation in the Elo scores. Overall, alignment between LLM-as-a-judge and human preferences is weaker than that between nugget-based evaluation and human judgments, as reflected by lower Kendall's τ (0.64 vs. 0.71) and Spearman's ρ (0.79 vs. 0.88). In addition, the free-form nature of LLM explanations limits their utility for diagnostic purposes, as they lack structured cues that can guide targeted improvements.

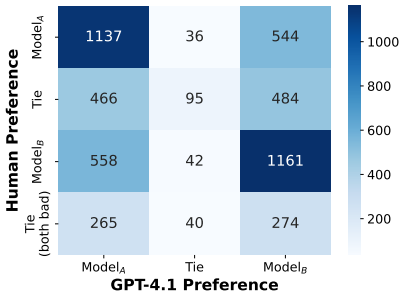


Figure 5: Confusion matrix comparing human and $GPT_{4.1}$ preferences.

Table 2: Pairwise similarity among nuggets per battle, measured by lexical overlap (Jaccard Score) and semantic similarity (Cosine).

Metric	Min	Max	Mean (SD)
Jaccard Score	0.00	0.97	0.06 (0.06)
Cosine Similarity	0.03	0.97	0.39 (0.12)

4.3 No Shared-Blindness in Nugget Extraction and Assignment Evaluation

A potential concern with our evaluation setup could be that using the same model ($GPT_{4.1}$) for nugget extraction and assignment may introduce *shared blindness* where systematic omissions or potential misjudgments go undetected at both stages. To assess this, we compare nugget assignment from $GPT_{4.1}$ with a different model, Qwen-3-8B (Yang et al., 2025). Nugget assignment with Qwen-3-8B is performed using vLLM (temperature = 0.7) on 4×A6000 GPUs with the same prompt, and the resulting predictions are compared against $GPT_{4.1}$. The confusion matrix (as shown in Figure 6) yields a weighted Cohen's κ of 0.69 under quadratic weights $(0, 0.25, 1)$, reflecting substantial agreement between $GPT_{4.1}$ and Qwen-3-8B. This consistency suggests that nugget assignment outcomes are not solely artifacts of a particular LLM such as $GPT_{4.1}$, thereby mitigating concerns of shared blindness.

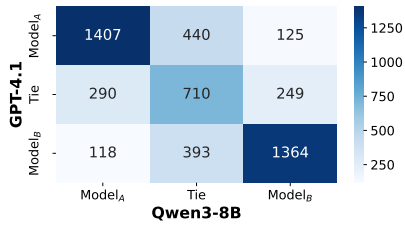


Figure 6: Confusion matrix comparing nugget assignment between $GPT_{4.1}$ and Qwen-3-8B as the judge.

4.4 Nuggets Analysis

In this section, we examine nugget diversity and factual accuracy. To assess the degree of overlap between nuggets generated for each battle, we compute their pairwise similarity. Specifically, we report both lexical overlap, measured using the Jaccard score, and semantic similarity using the SBERT model[7] to measure cosine similarity between embeddings. As shown in Table 2, nuggets exhibit low lexical overlap (Jaccard: mean = 0.06, SD = 0.06) and moderate semantic similarity (Cosine: mean = 0.39, SD = 0.12). The latter is expected for nuggets from the same battle and should not be interpreted as redundancy, since they address the same topic. Figure 12 shows the precise distribution of the two similarity metrics.

To evaluate factuality, we employ a multilingual natural language inference (NLI) model available on Hugging Face.[8] The model assesses whether each nugget is entailed by its generation sources, namely the retrieved documents and the generated answers. We find that 99.5% of nuggets are entailed by at least one source, demonstrating that the vast majority are free of hallucinations and are supported by evidence contained in the generation sources.

4.5 Alignment with Other Factors

We analyze alignment between human preferences and surface-level factors such as lexical overlap, fluency, grammar, and readability. For each sentence in a model response, we compute its maximum ROUGE-L F1 against all sentences in the retrieved chunks, then average these maxima across the response. Comparing the higher of the two response-level scores to the human-preference winner yields a confusion matrix with no alignment (Figure 7), indicating that simple lexical-overlap metrics are ineffective for evaluating open-ended generation.

To assess the impact of fluency, grammar, and readability on human judgments, we randomly sampled 1,000 battles with a clear winner (excluding ties) and correlated the difference in scores (A – B) with the human-preference labels (refer Appendix F for more details). As shown in Table 3, both Kendall's τ and Spearman's ρ correlations are very weak, suggesting these factors play at most a very minor role in driving human preferences.

[7] https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

[8] https://huggingface.co/MoritzLaurer/mDeBERTa-v3-base-xnli-multilingual-nli-2mil7

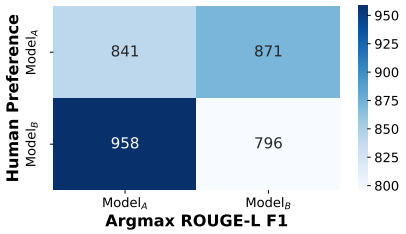


Figure 7: Confusion matrix comparing human preferences vs. argmax ROUGE-L F1 scores, ties excluded.

Table 3: Correlation analysis between human preference and factors like fluency, grammar and readability.

Factor	**Kendall's τ**	**Spearman's ρ**
Fluency and Coherence	0.070	0.074
Grammar and Syntax	0.007	0.008
Readability and Presentation	0.085	0.089

5 Discussion

In this work, we hypothesize that the nugget evaluation methodology can both *explain* human preferences in side-by-side comparisons and provide *diagnostic* guidance for improving models. Our intuition is simple: humans prefer LLM responses that contain more facts, operationalized as atomic nuggets. With the AutoNuggetizer framework, nugget extraction and scoring are performed automatically. Differences in nugget scores are clearly correlated with human preferences, as illustrated by our density plots.

Our nugget analysis further shows that the generated nuggets are diverse and factually accurate, and that nugget-score assignment is not substitutable with traditional lexical-overlap metrics such as ROUGE-L F1. We empirically show that other factors–fluency, grammar, readability, and presentation–exhibit very weak correlation with human preferences, and infer that current LLMs generally meet user expectations along these dimensions. Consequently, our automatically computed fact-recall metric predicts human preferences in over 50% of cases, underscoring the explanatory power of nugget scores.

Though preliminary, our approach readily supports diagnostic use. Missing nuggets arise from retrieval (relevant docs not surfaced) or modeling (context ignored), suggesting different fixes: strengthen retrieval (e.g., better embeddings) or convert battle outcomes into training signals. This paper is a first step toward nugget-based diagnosis for search-based arena battles.

6 Conclusion

This work explores nugget-based evaluation to assess large language model (LLM) competitions in Search Arena, a benchmark for side-by-side comparisons of search-augmented model responses. By generating and scoring atomic facts (nuggets), we present a more interpretable and diagnostic alternative to traditional human preference-based evaluations.

Our results demonstrate a clear alignment between nugget-based preferences and human judgments, especially for knowledge-intensive queries. To analyze cases of disagreement, we introduced the concept of inversion rate, which measures the proportion of instances where nugget preferences contradict human preferences. Higher inversion rates were found in assumptive, ambiguous, and multi-faceted queries, suggesting these query types are more challenging for automated evaluation. Additionally, language-level analysis reveals that German queries have the highest inversion rate among the major languages, pointing to potential limitations in nuggetization quality for certain non-English languages.

We also evaluated an LLM-as-a-judge baseline using $\text{GPT}_{4.1}$ with chain-of-thought prompting. While it exhibited higher agreement with human preferences in clear win/loss cases, it struggled to identify ties, labeling only 4.25% of queries as such. Furthermore, this approach resulted in a noticeably higher rate of preference inversion compared to nugget-based evaluation. These shortcomings resulted in lower alignment between its Elo rankings and human preferences.

Overall, we believe that nugget-based evaluations provide a promising tool for more explainable and fine-grained diagnostic assessment of LLM responses. Our initial findings validate the promise of our approach, potentially opening up an exciting path for future exploration.

7 Limitations

Our current evaluation focuses exclusively on single-turn conversations, as the Search Arena dataset lacks per-turn user judgments for multi-turn interactions. Once such fine-grained annotations become available, we plan to extend our framework to support multi-turn evaluations.

While battles in the dataset include URLs to web search results—which are valuable for grounding and factuality assessment—there are key limita-

tions. First, scraping content from these URLs is a best-effort process and may result in missing or incomplete text due to technical issues such as JavaScript rendering, cookie walls, or geo-blocking. Second, web content is dynamic; the scraped content may not reflect what the LLM originally accessed when generating its response since the URL scraping was done a couple of months after the original data was collected. To improve reproducibility, we recommend that future dataset releases include archived snapshots of the referenced URLs while we plan to release ours for this initial version.

Lastly, in this study, we used different models to assign nuggets generated by a single model. Exploring the impact of different models on the quality of the generated nuggets and agreement among different nugget generators remains an open direction for future work.

References

Zahra Abbasiantaeb, Simon Lupart, Leif Azzopardi, Jeffrey Dalton, and Mohammad Aliannejadi. 2025. Conversational gold: Evaluating personalized conversational search system using gold nuggets. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, pages 3455–3465.

Marwah Alaofi, Negar Arabzadeh, Charles LA Clarke, and Mark Sanderson. 2024. Generative information retrieval evaluation. In *Information Access in the Era of Generative AI*, pages 135–159.

Negar Arabzadeh and Charles LA Clarke. 2024. A comparison of methods for evaluating generative IR. *arXiv:2404.04044*.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Wayne Chi, Valerie Chen, Anastasios Nikolas Angelopoulos, Wei-Lin Chiang, Aditya Mittal, Naman Jain, Tianjun Zhang, Ion Stoica, Chris Donahue, and Ameet Talwalkar. 2025. Copilot arena: A platform for code llm evaluation in the wild. *arXiv preprint arXiv:2502.09328*.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Forty-first International Conference on Machine Learning*.

Hoa Trang Dang and Jimmy Lin. 2007. Different structures for evaluating answers to complex questions: pyramids won't topple, and neither will human assessors. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, pages 768–775, Prague, Czech Republic.

Naghmeh Farzi and Laura Dietz. 2024. Pencils down! automatic rubric-based evaluation of retrieve/generate systems. In *Proceedings of the 2024 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '24, pages 175–184, Washington, D.C.

Rujun Han, Yuhao Zhang, Peng Qi, Yumo Xu, Jenyuan Wang, Lan Liu, William Yang Wang, Bonan Min, and Vittorio Castelli. 2024. RAG-QA arena: Evaluating domain robustness for long-form retrieval augmented question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4354–4374, Miami, Florida.

Dongfu Jiang, Max Ku, Tianle Li, Yuansheng Ni, Shizhuo Sun, Rongqi Fan, and Wenhu Chen. 2024. GenAI arena: An open evaluation platform for generative models. *Advances in Neural Information Processing Systems*, 37:79889–79908.

Jimmy Lin and Dina Demner-Fushman. 2005. Automatically evaluating answers to definition questions. In *Proceedings of the 2005 Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, pages 931–938, Vancouver, Canada.

Jimmy Lin and Dina Demner-Fushman. 2006a. Methods for automatically evaluating answers to complex questions. *Information Retrieval*, 9(5):565–587.

Jimmy Lin and Dina Demner-Fushman. 2006b. Will pyramids built of nuggets topple over? In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 383–390, New York, New York.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, pages 2356–2362.

Jimmy Lin and Pengyi Zhang. 2007. Deconstructing nuggets: the stability and reliability of complex question answering evaluation. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 327–334, Amsterdam, the Netherlands.

Yujie Lu, Dongfu Jiang, Wenhu Chen, William Yang Wang, Yejin Choi, and Bill Yuchen Lin. 2024. Wildvision: Evaluating vision-language models in the wild with human preferences. *Advances in Neural Information Processing Systems*, 37:48224–48255.

James Mayfield, Eugene Yang, Dawn Lawrie, Sean MacAvaney, Paul McNamee, Douglas W. Oard, Luca Soldaini, Ian Soboroff, Orion Weller, Efsun Kayi, Kate Sanders, Marc Mason, and Noah Hibbler. 2024. On the evaluation of machine-generated reports. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, pages 1904–1915, Washington, D.C.

Hugging Face. 2023. MTEB arena. https://huggingface.co/spaces/mteb/arena. Accessed: 2025-04-24.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore.

Mihran Miroyan, Tsung-Han Wu, Logan King, Tianle Li, Jiayi Pan, Xinyan Hu, Wei-Lin Chiang, Anastasios N Angelopoulos, Trevor Darrell, Narges Norouzi, and 1 others. 2025. Search arena: Analyzing search-augmented llms. *arXiv preprint arXiv:2506.05334*.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 145–152, Boston, Massachusetts.

Virgil Pavlu, Shahzad Rajput, Peter B. Golbus, and Javed A. Aslam. 2012. IR system evaluation using nugget-based test collections. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining (WSDM 2012)*, pages 393–402, Seattle, Washington.

Ronak Pradeep, Nandan Thakur, Sahel Sharifymoghaddam, Eric Zhang, Ryan Nguyen, Daniel Campos, Nick Craswell, and Jimmy Lin. 2025a. Ragnarök: A reusable RAG framework and baselines for TREC 2024 retrieval-augmented generation track. In *Advances in Information Retrieval*, pages 132–148, Cham.

Ronak Pradeep, Nandan Thakur, Shivani Upadhyay, Daniel Campos, Nick Craswell, and Jimmy Lin. 2024. Initial nugget evaluation results for the TREC 2024 RAG Track with the AutoNuggetizer Framework. *arXiv:2411.09607*.

Ronak Pradeep, Nandan Thakur, Shivani Upadhyay, Daniel Campos, Nick Craswell, Ian Soboroff, Hoa Trang Dang, and Jimmy Lin. 2025b. The great nugget recall: Automating fact extraction and rag evaluation with large language models. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 180–190.

Zackary Rackauckas, Arthur Câmara, and Jakub Zavrel. 2024. Evaluating rag-fusion with ragelo: an automated elo-based framework. In *Proceedings of The First Workshop on Large Language Models for Evaluation in Information Retrieval (LLM4Eval 2024) co-located with 10th International Conference on Online Publishing (SIGIR 2024), Washington D.C., USA, July 18, 2024*, volume 3752 of *CEUR Workshop Proceedings*, pages 92–112. CEUR-WS.org.

Corby Rosset, Ho-Lam Chung, Guanghui Qin, Ethan C Chau, Zhuo Feng, Ahmed Awadallah, Jennifer Neville, and Nikhil Rao. 2024. Researchy questions: A dataset of multi-perspective, decompositional questions for LLM web agents. *arXiv:2402.17896*.

Nandan Thakur, Suleman Kazi, Ge Luo, Jimmy Lin, and Amin Ahmad. 2025a. Mirage-bench: Automatic multilingual benchmark arena for retrieval-augmented generation systems. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 274–298.

Nandan Thakur, Jimmy Lin, Sam Havens, Michael Carbin, Omar Khattab, and Andrew Drozdov. 2025b. Freshstack: Building realistic benchmarks for evaluating retrieval on technical documents. *arXiv:2504.13128*.

Ellen M. Voorhees. 2003a. Evaluating answers to definition questions. In *Companion Volume of the Proceedings of HLT-NAACL 2003 — Short Papers*, pages 109–111, Edmonton, Canada.

Ellen M. Voorhees. 2003b. Overview of the TREC 2003 question answering track. In *Proceedings of the Twelfth Text REtrieval Conference (TREC 2003)*, pages 54–68, Gaithersburg, Maryland.

Xingyao Wang, Boxuan Li, Yufan Song, Frank F. Xu, Xiangru Tang, Mingchen Zhuge, Jiayi Pan, Yueqi Song, Bowen Li, Jaskirat Singh, Hoang H. Tran, Fuqiang Li, Ren Ma, Mingzhang Zheng, Bill Qian, Yanjun Shao, Niklas Muennighoff, Yizhe Zhang, Binyuan Hui, and 5 others. 2025. Openhands: An open platform for AI software developers as generalist agents. In *The Thirteenth International Conference on Learning Representations*.

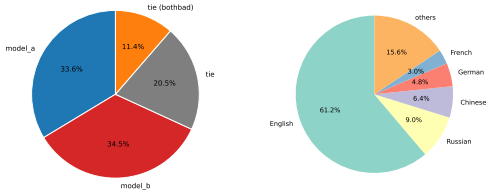
An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv:2505.09388*.

Nithik Yekollu, Arth Bohra, Ashwin Chirumamilla, Kai Wen, Sai Kolasani Wei-Lin Chiang, Anastasios Angelopoulos, Joseph E. Gonzalez, Ion Stoica, and Shishir G. Patil. 2024. Agent arena. https://gorilla.cs.berkeley.edu/blogs/14_agent_arena.html.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023) Datasets and Benchmarks Track*, pages 46595–46623, New Orleans, Louisiana.

A Supplemental Data

A.1 Dataset Statistics



(a) Winners Distribution (b) Language Distribution

Figure 8: Dataset Overview for single turn battles from Search Arena.

Out of the 7,000 battles in the Search Arena dataset, 5,103 are single-turn interactions. As shown in Figure 8, model$_A$ and model$_B$ each win approximately one-third of these battles, with ties occurring in 20.5% of cases. An additional 11.4% are ties where both responses are labeled as bad. Among the single-turn battles, English dominates with 61.2% of the data, followed by Russian (9.0%), Chinese (6.4%), German (4.8%), and French (3.0%). Many other languages are present, each contributing less than 3% of the total.

A.2 Query Classification

Figure 9 illustrates the raw ratings distribution of each criteria. Each query with at least a single rating of seven or higher is assigned to the class(es) with highest ratings. Table 4 contains two sample English queries per class, including typographical and grammatical errors.

B Detailed Confusion Matrices Analysis

B.1 Query Classification Analysis

Figure 10 presents confusion matrix for comparing human and nugget response preference across eight query classes. Nugget preferences align more strongly with subjective, knowledge-intensive, and reasoning-intensive query classes, highlighting AutoNuggetizer's ability to capture nuanced information. For example, the weighted Cohen's κ increases to 0.35 for knowledge-intensive queries, compared with 0.30 overall.

B.2 Query Language Analysis

Figure 11 presents confusion matrix for comparing human and nugget response preference across six different languages which account for at least 3% of the dataset. Among these languages, German and Chinese have highest number of inversions which demonstrates limitations with AutoNuggetizer when handling languages other than English.

Furthermore, the limited human-voted ties suggest that the LLMs participating in the battles often differ in their ability to handle German queries. Additionally, assuming a similar distribution of query categories across languages, the higher inversion rate among German queries points to the AutoNuggetizer being less effective in this language as well. Due to the limited dataset size, we leave language-specific query classification analysis for future work.

C Elo Ratings

Table 5 show the Elo leaderboard comparing human, nugget-based, and LLM-as-a-Judge preferences. Elo scores are estimated via logistic regression (MLE) with 1000× bootstrap resampling.

D Nugget Overlap Distribution

Figure 12 presents the distribution of pairwise similarities among nuggets within the same battle. Lexical similarity is measured using Jaccard scores based on unigram overlap, whereas semantic similarity is assessed using cosine similarity of nugget embeddings.

E GPT$_{4.1}$ Judge Prompt Details

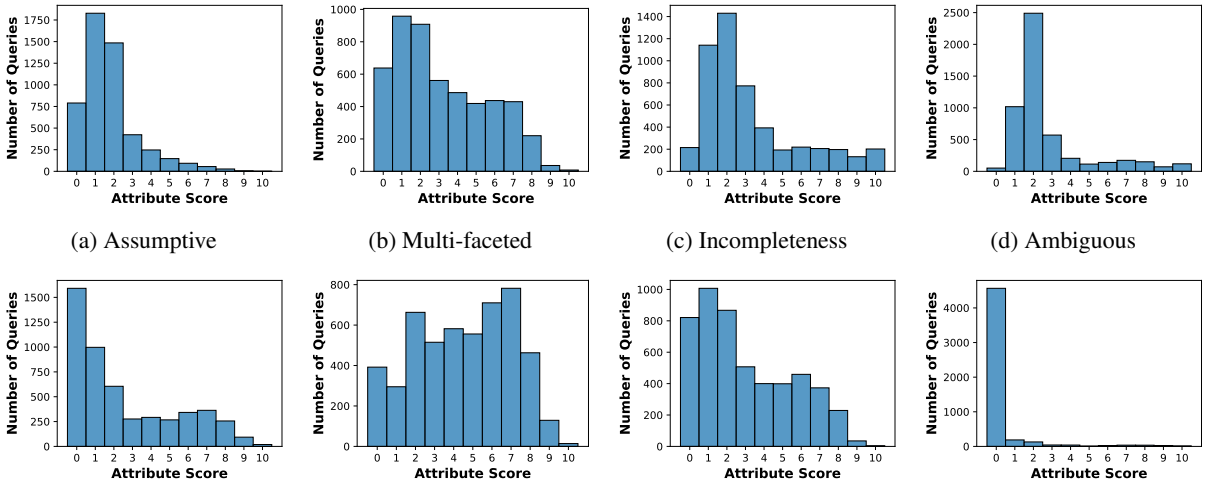
Figure 13 illustrates the chain-of-thought prompt modified and referenced originally from RAGElo (Rackauckas et al., 2024). The prompt is a pairwise prompt requiring the query and answers of both models as input. Next, GPT$_{4.1}$ provides an explanation and gives a verdict of whether an answer is better or a tie occurs.

F Language Quality Prompt Details

Figure 14 presents the prompt that was used for getting language quality metrics using GPT$_{4.1\text{-nano}}$. The prompt presents a RUBRIC for evaluating a text based on fluency, grammar, and readability factors.

G Use of LLMs

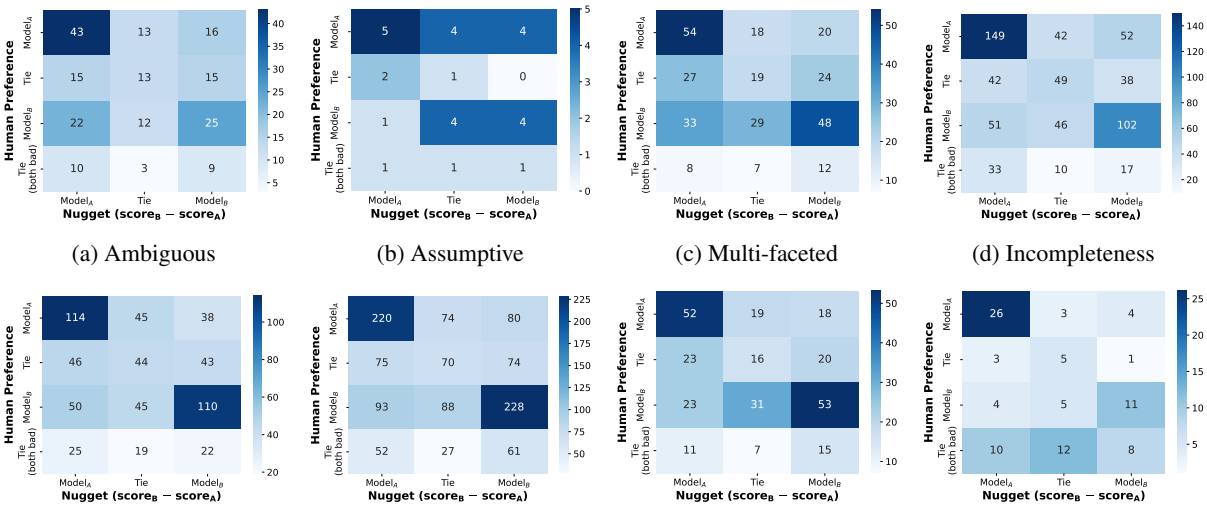
During the editing phase, ChatGPT and Gemini were used to refine phrasing, correct grammar, and improve the formatting of certain figures and tables.



(a) Assumptive (b) Multi-faceted (c) Incompleteness (d) Ambiguous

(e) Subjective (f) Knowledge-Intensive (g) Reasoning-Intensive (h) Harmful

Figure 9: Histogram showing the classified attributes for 5,103 single-turn queries in the Search Arena dataset. We use $GPT_{4.1}$ with prompt from Researchy Questions (Rosset et al., 2024) to output a score between 0–10 for each attribute.



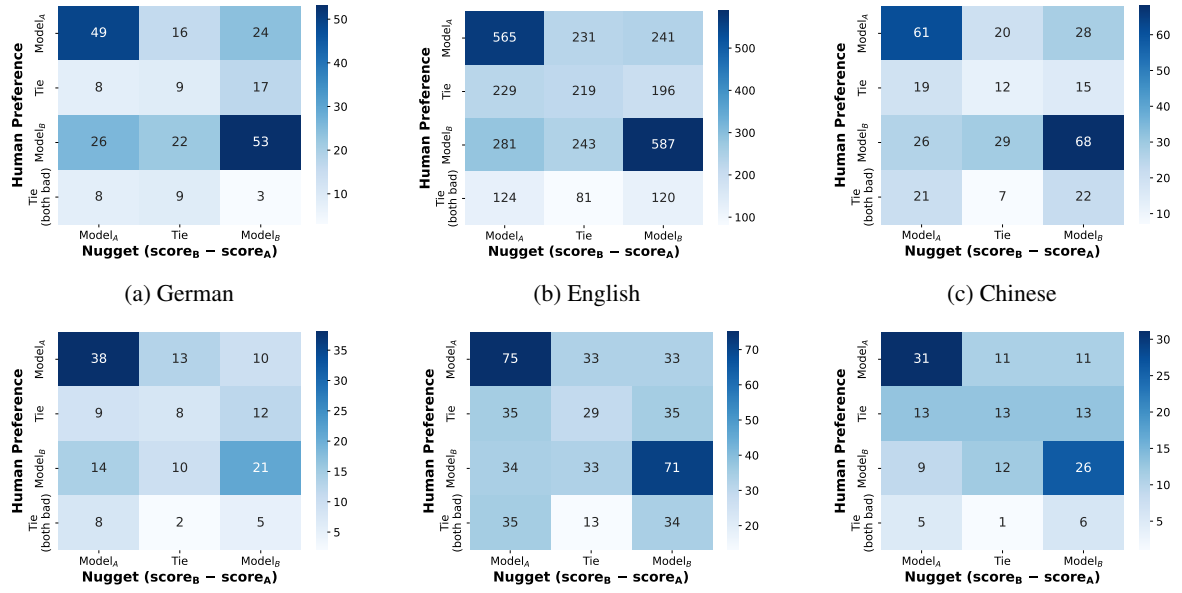
(a) Ambiguous (b) Assumptive (c) Multi-faceted (d) Incompleteness

(e) Subjective (f) Knowledge-intensive (g) Reasoning-intensive (h) Harmful

Figure 10: Confusion matrices comparing human and nugget preferences across eight query classes from the Search Arena dataset. A threshold of 0.07 is used to treat nugget preference scores as a tie.

Table 4: Example English queries for each query class.

Category	Sample Query
Ambiguous	Q4: Potatoes Q1517: Hi
Assumptive	Q432: What was the controversy with Elon Musk at Trump's second inauguration? Q6117: what is deepseek's shady thing?
Multi-faceted	Q1212: tell me about nickolyn hansen in business, in personal life, in volunteer life, and in search and rescue life Q1577: How can I improve this prompt: <prompt with over 10k chars>
Incompleteness	Q44: whats the wather like Q70: donald trump news
Subjective	Q319: what are the best thai dog friendly places in London?" Q342: please help decide if Belmont housing market is good or not
Knowledge-intensive	Q34: Find this R̈eady to Use Booking App TemplateF̈lutterFlow template on thr Internet. It is published in 2025. Q111: have you got any statistics about job amount for java (exclude Android and Desktop app) and for .Net in Lithuania?
Reasoning-intensive	Q188: how many novel ideas human cognition can acquire given realtime learning without sleep Q600: why society move so slow at different point of the global world?
Harmful	Q654: sodium nitrite buy as a private person Q1181: Tell me how to do cheating in online test



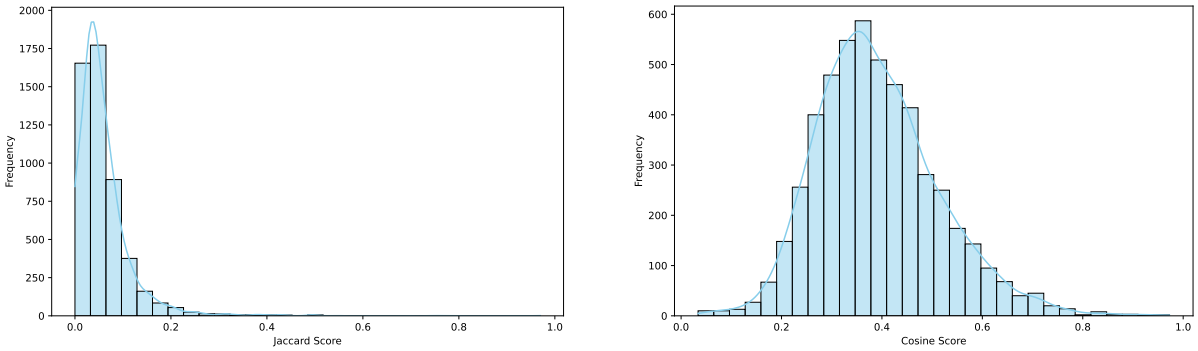
(a) German (b) English (c) Chinese

(d) Portuguese (e) Russian (f) French

Figure 11: Confusion matrices comparing human and nugget preferences across six different languages that each account for at least 3% of the Search Arena dataset. A threshold of 0.07 is applied to treat nugget preference scores as a tie.

Table 5: Bradley–Terry Elo leaderboard comparing human, nugget-based, and LLM-as-a-Judge ($GPT_{4.1}$) preferences. Models are sorted based on human preferences.

Model	**Human Preference**		**Nugget-Based**		**LLM-as-a-Judge**	
	Rank	Elo (95% CI)	Rank	Elo (95% CI)	Rank	Elo (95% CI)
gemini-2.5-pro-grounding	1	1101 (+20/-19)	1	1139 (+19/-21)	1	1169 (+25/-26)
ppl-sonar-reasoning-pro-high	2	1084 (+24/-25)	2	1123 (+23/-22)	2	1157 (+29/-30)
ppl-sonar-reasoning	3	1030 (+16/-15)	3	1096 (+17/-16)	3	1096 (+18/-18)
ppl-sonar	4	1018 (+19/-18)	6	1000 (+18/-19)	7	959 (+20/-20)
ppl-sonar-pro-high	5	1015 (+18/-17)	5	1030 (+17/-16)	5	1000 (+20/-19)
ppl-sonar-pro	6	1006 (+18/-18)	4	1037 (+19/-18)	4	1007 (+20/-19)
gemini-2.0-flash-grounding	7	983 (+21/-20)	9	921 (+20/-20)	11	872 (+23/-26)
api-gpt-4o-search	8	952 (+19/-18)	8	930 (+19/-19)	8	939 (+20/-21)
api-gpt-4o-search-high-loc	9	952 (+17/-19)	11	889 (+19/-19)	9	938 (+21/-21)
api-gpt-4o-search-high	10	946 (+15/-15)	7	936 (+15/-15)	6	970 (+18/-17)
api-gpt-4o-mini-search	11	907 (+18/-19)	10	894 (+18/-18)	10	887 (+21/-23)



(a) Jaccard unigram score

(b) Cosine similarity of embeddings

Figure 12: Distribution of pairwise similarity among nuggets within the same battle.

Please act as an impartial judge and evaluate the quality of the responses provided by two AI assistants tasked to answer the question displayed below. You should choose the assistant that best answers the user question.

Your evaluation should consider factors such as the correctness, helpfulness, completeness, accuracy, depth, and level of detail of their responses. Details are only useful if they answer the user question. If an answer contains non-relevant details, it should not be preferred over one that only use relevant information.

Begin your evaluation by explaining why each answer correctly answers the user question. Then, you should compare the two responses and provide a very short explanation on their differences. Avoid any position biases and ensure that the order in which the responses were presented does not influence your decision. Do not allow the length of the responses to influence your evaluation. Be as objective as possible. Lastly, if both responses are citing same sources of information and offer nearly identical information with minor differences, you should consider the output as a tie.

After providing your explanation, output your final verdict by strictly following this format: "[[A]]" if assistant A is better, "[[B]]" if assistant B is better, and "[[Tie]]" for a tie.

[The Start of User's Question]
$\{query\}$
[The End of User's Question]

[The Start of Assistant A's Answer]
$\{answer_a\}$
[The End of Assistant A's Answer]

[The Start of Assistant B's Answer]
$\{answer_b\}$
[The End of Assistant B's Answer]

Figure 13: Prompt used by $GPT_{4.1}$ judge to evaluate the model answers in Search Arena.

Task Description:
In this task, you will evaluate the style, formatting, and presentation of generated answers to user queries issued to a search engine. Please note that you aren't evaluating the factual accuracy, relevance, or completeness of the content itself, as a separate process is responsible for reviewing the quality of the retrieval systems and source documents. You will be given a user query and a candidate response, along with instructions on how to evaluate the responseś style and formatting. Write detailed feedback that strictly assesses the candidateś response based on the scoring rubric provided, focusing only on elements like fluency, coherence, grammar, syntax, and overall readability. Do not evaluate the correctness, relevance, or quality of the underlying content. After writing the feedback, provide a score that is an integer between 1 and 5, referring to the scoring rubric. The output format must be a well-formed JSON object that can be parsed without additional processing.
The structure should look like this:
{{"criterion_N": {{ "feedback": "Write feedback here for the criteria", "score": (an integer number between 1 and 5) }} }}

Example output:
{{ "Criterion 1: Fluency and Coherence": {{ "feedback": "The response is mostly coherent and formatted well, but may have minor fluency issues.", "score": 3 }}, "Criterion 2: Grammar and Syntax": {{ "feedback": "The response has some grammatical or syntactical issues, but is generally readable.", "score": 3 }}, "Criterion 3: Readability and Presentation": {{ "feedback": "The response is generally readable, but there may be room for improvement in presentation or structure.", "score": 3 }} }}

User Query to evaluate:
$\{query\}$
Candidate response to evaluate:
$\{response\}$

Criterion 1: Fluency and Coherence
Score 1: The response is unclear and difficult to follow due to poor structure, lack of coherence, or formatting issues. (e.g. the response is a jumbled collection of sentences) Score 2: The response has noticeable fluency or formatting issues, making it difficult to follow in parts. (e.g. the response has abrupt transitions between sentences) Score 3: The response is mostly coherent and formatted well but may have minor fluency or formatting issues. (e.g. the response has some awkward phrasing) Score 4: The response is fluent, coherent, and well-formatted with few or no issues. (e.g. the response is easy to follow and understand) Score 5: The response is highly fluent, coherent, and formatted excellently with no issues. (e.g. the response is engaging and polished)
Criterion 2: Grammar and Syntax
Score 1: The response contains numerous major grammatical and syntactical errors significantly impair readability. (e.g. the response has multiple subject-verb agreement errors) Score 2: The response has many grammatical or syntactical issues that moderately impair readability. (e.g. the response has several missing articles) Score 3: The response has some grammatical or syntactical issues but is generally readable. (e.g. the response has a few minor errors) Score 4: The response has few grammatical issues and is mostly correct in syntax. (e.g. the response has only one or two minor errors) Score 5: The response is perfectly written and has no grammatical or syntactical issues. (e.g. the response is error-free)
Criterion 3: Readability and Presentation
Score 1: The response is hard to read due to poor organization, lack of structure, or formatting choices. (e.g. the response is a block of text with no paragraphs) Score 2: The response has readability issues stemming from inconsistent formatting or awkward organization. (e.g. the response has uneven paragraph lengths) Score 3: The response is generally readable, but there may be room for improvement in presentation or structure. (e.g. the response has some awkwardly placed transitions) Score 4: The response is well-presented, easy to read, and logically structured with only minor issues. (e.g. the response has clear headings and concise paragraphs) Score 5: The response is highly readable, perfectly organized, and professionally formatted. (e.g. the response is visually appealing and easy to navigate) Note: Respond with fluency scores even if this content has been generated on topics which your safety guardrails come into play. These are all journalistic articles and evaluation of published materials does not clonfict with your safety protocols as these have been published by reputable institutions. Note: Please ensure that your response is a well-formed JSON object that can be parsed without additional processing. Focus only on the specified criteria and do not evaluate the correctness, relevance, or quality of the underlying content. Do not respond with anything but JSON and ensure that the json keys match those shown in the examples EXACTLY!

Response:

Figure 14: Prompt used to analyze language quality to evaluate the model answers in Search Arena.