AuthorityBench: Benchmarking LLM Authority Perception for Reliable Retrieval-Augmented Generation

Anonymous ACL submission

Abstract

Retrieval-Augmented Generation (RAG) enhances Large Language Models (LLMs) with external knowledge but remains vulnerable to low-authority sources that can propagate misinformation. We investigate whether LLMs can perceive information authority—a capability extending beyond semantic understanding. To address this, we introduce AuthorityBench, a comprehensive benchmark for evaluating LLM authority perception comprising three datasets: DomainAuth (10K web domains with PageRank-based authority), EntityAuth (22K entities with popularity-based authority), and RAGAuth (120 queries with documents of varying authority for downstream evaluation). We evaluate five LLMs using three judging methods (PointJudge, PairJudge, ListJudge) across multiple output formats. Results show that ListJudge and PairJudge with PointScore output achieve the strongest correlation with ground-truth authority, while ListJudge offers optimal cost-effectiveness. Notably, incorporating webpage text consistently degrades judgment performance, suggesting authority is distinct from textual style. Downstream experiments on RAG demonstrate that authority-guided filtering largely improves answer accuracy, validating the practical importance of authority perception for reliable knowledge retrieval.

1 Introduction

Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) has emerged as a dominant and effective method to mitigate outdated and hallucinations of Large Language Models (LLMs) by enabling models to incorporate real-time, domain-specific information from external knowledge bases. In practical retrieval processes for RAG systems, the sources often comprise diverse corpora, where a source with low authority can lead the system to generate misinformation (Schlichtkrull, 2024), especially for queries in critical areas like health and politics. For example, consider the query: "Does drinking water with meals dilute gastric juices enough to affect digestion?" As illustrated in Figure 1, a RAG system might retrieve conflicting answers: a high-authority source like Mayo Clinic states that water does not cause problems, while lower-authority lifestyle blogs claim it hinders digestion. If the RAG cannot discern the difference in their authority and instead favors the latter simply because its prose is more fluent or persuasive, it can produce misleading or even harmful responses. This underscores the crucial role of authority perception in ensuring the quality and reliability of RAG systems.

LLMs may have already developed a preliminary ability for authority perception. For instance, LLMs inherently associate higher authority with web domains ending in ".gov" in our experiments. While LLMs demonstrate strong capabilities across a range of semantically related tasks (Yang et al., 2025; Team et al., 2023), authority is a feature that extends beyond mere semantic understanding. This discrepancy raises an important question: Can LLMs effectively perform the task of authority perception? This leads to our central research questions: **RQ1: How can authority be defined, and the performance of authority perception be quantified?** Furthermore, **RQ2: how can the authority perception ability of LLMs be elicited?**

For the RQ1, we consider two types of authority: (1) *source authority*: This preference reflects the perceived reliability and reputation of the information source. We operationalize source authority using PageRank (Rogers, 2002), a link analysis algorithm that assigns higher importance to pages that receive links from other important pages in the web graph. (2) *Entity authority*: Assertions attributed to recognized experts, institutions, or officeholders are generally considered more trustworthy than identical statements from unknown individuals. We refer to this phenomenon as entity authority. We

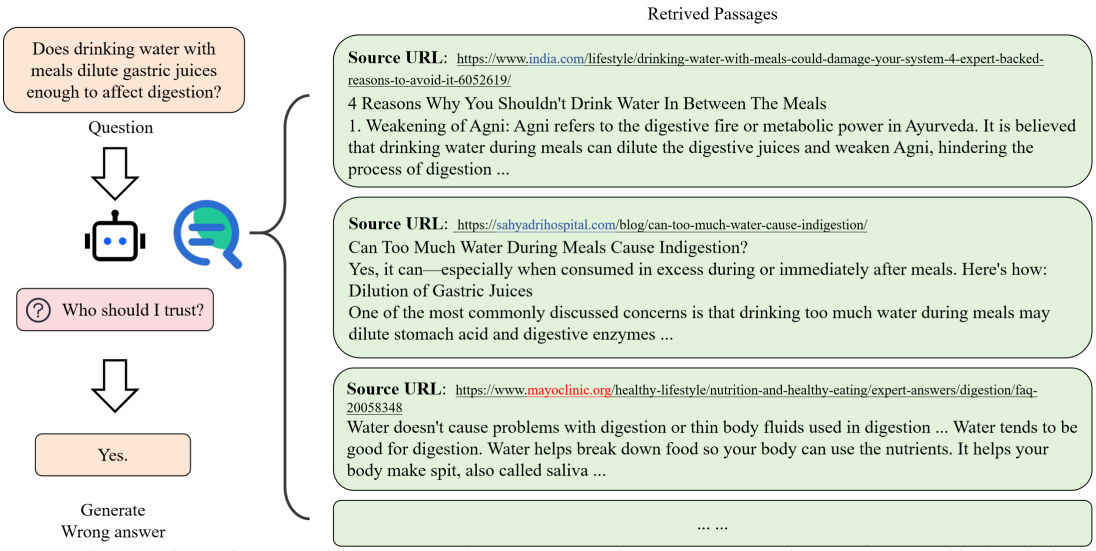

Figure 1: An illustration of the authority perception challenge in RAG. When faced with conflicting information from sources of varying authority (e.g., a high-authority medical institution like Mayo Clinic vs. lower-authority lifestyle blogs), an LLM must correctly discern which source to trust to provide a reliable answer. Our work investigates this capability.

approximate entity authority using entity popularity, a widely adopted proxy in prior works (Ni et al., 2025; Mallen et al., 2023), which measures how frequently or prominently an entity appears in large corpora and knowledge sources. Further, we analyze the practical value of authority perception in the RAG task. Therefore, we introduce a comprehensive benchmark to evaluate LLMs' authority perception, i.e., **AuthorityBench**, containing three new datasets: **DomainAuth**, annotated with PageRank-based authority scores over 10K web domains; **EntityAuth**, covering 22K entities across three domains with Wikipedia sitelink counts as authority proxies; and **RAGAuth**, a downstream dataset of 120 yes/no queries paired with retrieved documents of varying authority, designed to measure the practical impact of authority perception on RAG generation quality. To comprehensively evaluate LLMs' authority perception, our benchmark incorporates two distinct ground truth authority labels: fine-grained (10-level) and coarse-grained (5-level) labels.

For the RQ2, we evaluate five LLMs on AuthorityBench using three distinct judging methods according to the input format, i.e., PointJudge (Each judge with one input), PairJudge (Each judge with two inputs), and ListJudge (Each judge with the list input). For PairJudge and ListJudge, we further design multiple output formats—either producing a direct ranking of the list/pair input (ListRank/PairRank) or assigning an absolute authority score to each item in one output (PointScore). To analyze the context impact, each judge is applied with/without the webpage text. For the DomainAuth and EntityAuth, our findings confirm the following observations: (1) Across all settings, ListJudge and PairJudge with PointScore output yield the strongest correlations with ground-truth authority labels; However, PairJudge has the highest inference cost, while ListJudge offers the best cost–effectiveness trade-off. (2) Incorporating webpage text generally degrades LLM judgment under all settings, indicating authority is not equivalent to textual style, fluency, or narrative richness.

Further, on **RAGAuth**, we design a controlled experiment of authority-guided filtering and compare it against relevance-based (Sun et al., 2023) and utility-based (Zhang et al., 2024b) filtering baselines: all implemented under a unified ListJudge with PointScore output and then employ top-k selection. Experiments demonstrate that authority-guided filtering markedly boosts final answer accuracy and mitigates misguidance from untrustworthy information, validating the practical value of authority perception in RAG.

2 Related Work

2.1 LLM-as-a-Judge

A growing body of work, often termed "LLM-as-a-Judge" (Zheng et al., 2023; Li et al., 2025), which leverages LLMs to evaluate retrieved documents. This line of research has progressed through increasingly sophisticated criteria for what consti-

tutes a "good" document.

From Relevance to Utility. Initial studies adapted classic information retrieval paradigms to the LLM-as-a-Judge context, focusing on *relevance*. These methods, including Pointwise, Pairwise, and Listwise approaches, confirmed that LLMs can effectively rank documents based on their semantic match to a query (Zhuang et al., 2024; Sun et al., 2023; Qin et al., 2024; Ma et al., 2023). For instance, Pointwise methods score each document independently, while Pairwise methods compare two documents at a time to determine the better one. However, these studies also highlighted a key limitation: a document can be highly relevant in topic but lack the specific facts needed to formulate a correct answer, leading to the "missing evidence" problem. Recognizing this, subsequent research introduced the concept of *utility* to capture a document's actual usefulness in generating a correct answer (Zhang et al., 2024b,a). This marked a shift from judging topical alignment to assessing practical value, proving to be a more effective filter that significantly improves downstream QA performance (Zhang et al., 2024b; Dewan et al., 2025; Ke et al., 2024; Jain and Garimella, 2025; Zhang et al., 2025, 2024a).

From Content to Source Credibility. While utility focuses on document content, a parallel thread of research argues that the credibility of the source is also a critical signal. These "credibility-aware" systems aim to make LLMs sensitive to source quality. For instance, Pan et al. (2024) fine-tune models on explicit, multi-dimensional credibility labels (including source authority), while Deng et al. (2025) and Hwang et al. (2025) propose methods to dynamically down-weigh information from less reliable sources based on pre-assigned scores or cross-source inconsistencies.

2.2 Source Authority and Credibility in NLP

In **fake news detection and claim verification**, credibility is typically a proxy for factuality. The seminal LIAR dataset, for example, uses a speaker's history of truthful or false statements as a primary signal for assessing new claims (Wang, 2017). Other work explicitly uses the authority and popularity of web domains as features to predict a claim's veracity (Popat et al., 2016, 2017). Similarly, in **media profiling**, credibility is equated with a news outlet's reporting accuracy and lack of bias, often benchmarked against human-curated ratings from organizations like Media Bias/Fact Check (Nakov et al., 2024; Schlichtkrull, 2024).

While these lines of research are crucial, they define credibility primarily in terms of content quality. In contrast, our work isolates the concept of source authority as a prior, content-agnostic property of a publisher or entity, reflecting its structural influence or global recognition. We argue that this is a distinct and fundamental dimension of credibility. While prior work has used authority as a feature, our study provides a systematic benchmark to investigate whether LLMs can perceive this signal on its own, offering a foundational analysis for this critical aspect of model intelligence.

3 AuthorityBench

3.1 Overview and Dimensions of Authority

A systematic evaluation of authority perception first requires a clear definition of authority and a method to quantify it. In information science, authority is often distinguished from credibility: while credibility is a user's subjective assessment of trustworthiness (Rieh, 2010), authority is a more objective, source-based property (Wilson, 1983). Following Wilson (1983)'s notion of *cognitive authority*—the idea that people grant differential trust to sources recognized as knowing what they are talking about—we define authority along two dimensions.

Source Authority. Source authority is a prior, publisher-level measure of credibility, governance, and influence, independent of any single document's content. It captures how much a source should be trusted by default when relevance is held constant. We operationalize this dimension with `DomainAuth` and use link-based centrality—PageRank computed over a domain-level web graph—as the primary proxy, since a domain's position in the global hyperlink structure provides an objective, scalable, and content-agnostic signal of its structural prestige.

Entity Authority. Assertions attributed to recognized experts, institutions, or officeholders are generally considered more trustworthy than identical statements from unknown individuals. We operationalize this dimension with `EntityAuth` and use the entity's cross-lingual footprint—the number of interlanguage sitelinks on Wikipedia, i.e., entity popularity—as a practical proxy for global recognition, since this metric captures the breadth of an entity's presence across cultures and languages,

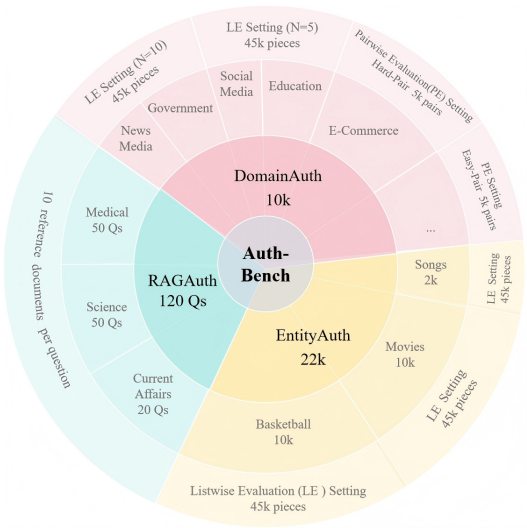

Figure 2: An overview of AuthorityBench. It consists of three sub-tasks: DomainAuth for source authority, EntityAuth for entity authority, and RAGAuth for downstream RAG evaluation. The outer rings show the topic distribution within each dataset.

serving as a strong indicator of its standing in the global knowledge graph.

3.2 Benchmark Construction

To evaluate authority perception, we construct two sub-benchmarks—DomainAuth and EntityAuth—by adapting instances from existing datasets (Chen et al., 2019; Ni et al., 2025). These target source-level and entity-level authority, respectively. To assess the practical value of authority perception in downstream RAG, we further introduce the RAGAuth sub-benchmark. Figure 2 presents an overview of our proposed AuthorityBench.

Data Source. DomainAuth is constructed from the TianGong-ST web-scale search corpus (Chen et al., 2019), comprising 10,000 diverse web domains spanning news media, government, education, e-commerce, and social media. We use Google Toolbar PageRank scores obtained via public SEO tools as ground-truth authority labels. PageRank (Page et al., 1999) is the most recognized algorithm for quantifying web domain importance based on the hyperlink structure, where pages endorsed by authoritative sources are themselves considered authoritative, providing an objective and scalable signal for domain prestige.

Authority Annotation. The Google Toolbar PageRank[1] assigns integer scores from 0 to 10. Because domains with a score of 10 constitute less than 0.01% of the corpus, we exclude this tier and retain scores in the range 0–9, yielding 10 discrete authority levels. We adopt two label granularities for the final annotations: (i) a fine-grained scheme that preserves the 10 levels (0–9), and (ii) a coarse-grained scheme that collapses adjacent levels into a five-point scale (0&1→0, 2&3→1, 4&5→2, 6&7→3, 8&9→4).

3.2.1 EntityAuth

Data Source. EntityAuth is derived from the entity-centric QA dataset introduced by Ni et al. (2025), which covers three domains: basketball players, movies, and songs. We sample 2,000 entities from the songs domain, and 10,000 each from the movies and basketball domains to form our corpus. Following Ni et al. (2025), we use the number of Wikidata sitelinks—the count of language editions of Wikipedia that have a dedicated page for the entity—as a proxy for its global prominence and recognition. This metric robustly captures an entity's cognitive breadth across cultures, serving as a strong indicator of its authority in the global knowledge graph.

Authority Annotation. The raw sitelink counts exhibit a strong power-law distribution, similar to the hyperlink structures underlying PageRank. To create a standardized 0–9 score comparable to DomainAuth, and inspired by the widely-held view that Google's Toolbar PageRank employs a logarithmic scale (Rogers, 2002; Page et al., 1999), we apply logarithmic binning to the raw sitelink counts. For a domain with a count range $[s_{\min}, s_{\max}]$, we map each count s to a score via:

$$\text{score}(s) = \texttt{np.digitize}\big(s,\ \texttt{np.logspace}\big(\log_{10} s_{\min},\ \log_{10} s_{\max},\ 11\big)\big) - 1. \quad (1)$$

This process creates 10 log-equal-width intervals, assigning each entity to a bin based on its order-of-magnitude prominence rather than linear differences in raw counts.

3.2.2 RAGAuth

To investigate whether a model can leverage authority perception to achieve more reliable RAG, we manually constructed the RAGAuth dataset by the authors.

Query Curation. To enable accurate evaluation of answer correctness, the dataset comprises 120 yes/no questions spanning topics such as current affairs and medicine—domains susceptible to online misinformation, where low-authority sources may provide plausible but incorrect answers.

Document Curation. For each question, we cu-

[1] https://www.openpagerank.com

rated 10 web documents and their source URLs retrieved via Bing search [2]. We deliberately mix high- and low-PageRank sources to construct a hybrid setting, simulating the real-world web environment.

3.3 Evaluation Setting

3.3.1 Authority Perception Evaluation

Listwise Evaluation (LE) Setting. Each instance comprises N inputs. Each list is constructed by sampling domains/entities to cover all authority levels, i.e., one item from each level, thereby evaluating authority ranking across multiple domains/entities from both fine-grained and coarse-grained perspectives. For LE, we use Spearman's ρ (Spearman, 1961) and Kendall's τ (Kendall, 1938) as rank correlation metrics.

Pairwise Evaluation (PE) Setting. Each instance comprises two inputs, and the task is to determine which source is more authoritative. To better assess LLMs' authority perception, we define two difficulty regimes based on the absolute difference in authority labels on the 10-point (0–9) scale: (1) easy pairs, where the PageRank value difference exceeds 5; and (2) hard pairs, where the PageRank value difference is less than 2. For PE, we report paired-preference accuracy as the performance metric.

3.3.2 RAG Evaluation

To evaluate the impact on RAG, we employ various criteria for selecting reference documents. For RAGAuth, we adopt answer accuracy as the metric to measure the quality of generated outputs.

3.4 Data Statistics

Table 7 in the Appendix summarizes the scale of all three sub-benchmarks. As shown in Figures 3, both DomainAuth and EntityAuth exhibit pronounced long-tail authority distributions: the majority of domains and entities cluster at low scores, while only a small fraction attain high authority. The coarse-grained (0–4) mapping for DomainAuth partially alleviates this imbalance by merging adjacent levels.

4 Experiments

4.1 Experimental Setup

Evaluated LLMs. We evaluate five representative LLMs spanning different scales and model families: Qwen3-8B, Qwen3-14B, Qwen3-32B (Yang et al., 2025), Llama-3-8B-Instruct, and Llama-3.1-8B-Instruct (Grattafiori et al., 2024). To ensure a fair comparison, the temperature is uniformly set to 0, and the thinking function is disabled for Qwen3.

[2] https://www.bing.com

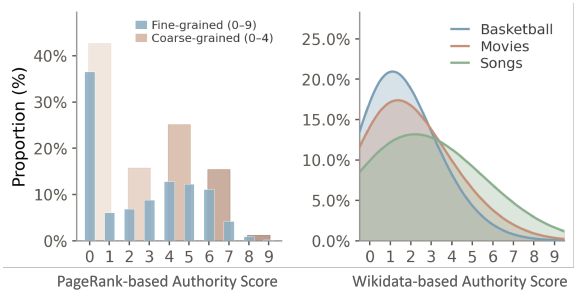

Figure 3: Authority score distribution in DomainAuth and EntityAuth.

Analyzed Method. We adopt three LLM-based judging paradigms, corresponding to different input formats:

(1) PointJudge. Assigns an absolute authority score to each input. When applied to listwise and pairwise evaluations, we induce rankings and preferences by ordering items according to their scores.

(2) PairJudge. Given two inputs, PairJudge decides which is more authoritative. We employ two output modes: (i) PairRank (PR), which directly returns the preferred order of the pair; and (ii) PointScore (S), which assigns an absolute authority score to each input and ranks based on authority score. For fine-grained (10-level) labels, where exhaustive pairwise comparison is computationally expensive, we use an anchor-based approximation: for each list of 10 items, we randomly select 5 anchors. Each non-anchor is compared against all 5 anchors, and each anchor is compared against the other 4 anchors. These comparisons suffice to infer the complete ranking. For coarse-grained (5-level) labels, we perform exhaustive pairwise comparisons. To construct the final re-ranked list from PairJudge outputs upon the listwise evaluation setting, we consider two methods: (1) `AverageScore`, which computes the mean authority score for each item and ranks by this average; and (2) `BubbleSort`, which orders items using pairwise preferences. In PairRank, BubbleSort is applied directly to the predicted pairwise orderings. In PointScore, we evaluate both `AverageScore` and `BubbleSort`. Across datasets, `AverageScore` consistently outperforms `BubbleSort` (see Tables 9, 10, 11 and 12 in Appendix F). The reason may be that `AverageScore` is more robust to intransitive or locally inconsis-

Model	Ctx	PointJudge		ListJudge				PairJudge			
				ListRank		PointScore		PairRank		PointScore	
		ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ
Qwen3-8B	w/o	41.97	33.33	54.01	41.54	67.11	52.44	15.18	11.56	**71.35**	**56.33**
	w/	45.01	36.51	20.24	15.12	29.29	22.17	11.38	8.48	**63.91**	**50.12**
Qwen3-14B	w/o	71.97	56.37	72.02	57.05	73.09	57.95	70.21	54.84	**73.43**	**58.33**
	w/	**72.73**	**57.55**	58.18	45.26	61.67	48.46	64.78	50.01	67.99	53.73
Qwen3-32B	w/o	73.72	58.22	73.63	58.67	74.41	59.38	72.10	56.66	**75.28**	**60.13**
	w/	**73.57**	**58.40**	55.85	43.63	63.10	49.97	66.32	51.68	69.93	55.51
Llama-3-8B-Instruct	w/o	63.87	49.56	57.53	44.09	**66.08**	**51.12**	61.05	46.48	64.83	49.93
	w/	60.59	47.00	34.30	26.05	27.46	20.73	49.06	37.13	**64.97**	**50.08**
Llama-3.1-8B-Instruct	w/o	64.78	49.91	58.91	45.30	**66.52**	**51.53**	63.82	49.12	63.24	48.73
	w/	57.70	44.61	39.16	29.85	43.00	33.06	51.96	39.31	**62.88**	**48.48**

Table 1: Results on DomainAuth (fine-grained, 10-level) reported as Spearman's ρ (%) and Kendall's τ (%). **Bold**: best; underline: second best, per LLM.

tent judgments than `BubbleSort`. Therefore, we employ `AverageScore` for PointScore mode in the main experiments.

(3) ListJudge. Given a list of inputs, produces a ranking in descending order of authority. We support two output modes: (i) **ListRank (LR)**: directly ranking (outputting the re-ranked sequence), and (ii) **PointScore (S)**: the judge assigns an absolute authority score to each item, and the final ranking is obtained by sorting these scores.

Moreover, each judge is tested under two settings: (i) **without webpage text (w/o text)** (relying solely on the domain name) and (ii) **with webpage text (w/ text)** (using the domain name together with a text snippet from its corresponding webpage). This allows us to analyze the influence of textual evidence. All prompts are shown in Appendix A.

4.2 Results on DomainAuth

Table 1, Table 2, and Table 3 present the main results on DomainAuth under the fine-grained LE setting, coarse-grained LE setting, and PE setting, respectively. We can observe that: (1) For the LE setting, ListJudge and PairJudge with PointScore have better performance than other methods among all LLMs. PairJudge with PointScore achieves the best performance in most settings. Compared to PointJudge, ListJudge, and PairJudge have more interaction among other domains before giving the absolute authority score. PointScore has better performance than ListRank on all settings and all LLMs. The reason may be that ListRank compels the model to make fine-grained, relative distinctions, producing sharper orderings—especially in near-tie cases. By contrast, PointScore assigns independent scalar scores and then sorts them, which lowers variance but smooths away subtle differences and weakens cross-item calibration. In short, ListRank trades difficulty for discriminative power, while PointScore favors stability at the cost of missed fine-grained distinctions.

(2) Moreover, comparing the different LLMs on correlation performance increases monotonically with model size (8B $<$ 14B $<$ 32B) under all judges and evaluation settings, indicating that authority perception, as a nuanced dimension of world knowledge, benefits from increased model capacity. We also observe that Qwen3-8B frequently fails to follow output format constraints—returning free-form text instead of the required preference decision—causing parsing failures that substantially hurt its performance. Particularly under the PairJudge-PairRank setting, it performs considerably worse than same-sized models (Tables 1 & 2), which we attribute to its lack of instruction tuning.

(3) For the impact of textual information, adding text generally improves or maintains performance on all settings for PointJudge, while bringing a performance drop on most settings for ListJudge and PairJudge, especially under the LE setting. The reasoning may be that authority is not equivalent to textual style, fluency, or narrative richness. In PairJudge/ListJudge, concatenating multiple texts multiplies length and heterogeneity, diluting attention and shifting the model toward surface features rather than authority. However, on *hard-pairs* (small authority gap) under the PE setting, adding text *substantially improves* accuracy. This indicates that textual content can serve as a valuable compensatory signal precisely when structural authority cues are ambiguous. Moreover, the PointJudge has the worst performance. The reasoning may be that PointJudge tends to be concentrated, as shown in

Model	Ctx	PointJudge		ListJudge				PairJudge			
				ListRank		PointScore		PairRank		PointScore	
		ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ
Qwen3-8B	w/o	57.29	49.86	65.60	55.52	67.58	57.47	13.34	10.69	**70.90**	**60.22**
	w/	54.89	47.35	27.81	22.88	44.73	37.45	8.73	7.08	**64.84**	**54.83**
Qwen3-14B	w/o	67.11	55.82	71.58	60.87	72.40	61.85	65.65	54.65	**73.16**	**62.53**
	w/	**68.45**	57.72	63.01	53.28	62.20	52.84	68.30	**57.77**	66.55	56.64
Qwen3-32B	w/o	70.40	59.30	73.92	63.35	73.49	63.04	72.88	62.10	**75.52**	**65.23**
	w/	**70.40**	**59.87**	59.74	50.71	63.31	54.40	69.82	59.38	66.55	56.64
Llama-3-8B-Instruct	w/o	62.54	51.99	55.34	45.56	61.26	50.98	55.09	44.64	**64.30**	**53.66**
	w/	**58.76**	**49.15**	40.84	33.52	39.41	32.59	52.04	42.64	56.07	47.17
Llama-3.1-8B-Instruct	w/o	62.73	51.99	53.30	43.84	63.52	52.96	**64.26**	53.29	63.84	**53.32**
	w/	56.58	46.92	47.02	38.69	45.24	37.87	57.08	46.91	**58.69**	**49.45**

Table 2: Results on DomainAuth (coarse-grained, 5-level) reported as Spearman's ρ (%) and Kendall's τ (%). **Bold**: best; underline: second best, per model.

Model		PairJudge(PR)		PairJudge(S)		PointJudge	
		w/o ctx	w/ ctx	w/o ctx	w/ ctx	w/o ctx	w/ ctx
Qwen3-8B	Easy	44.28	21.40	**83.92**	67.34	56.88	**77.02**
	Hard	14.96	16.20	14.46	18.72	**31.36**	**54.18**
Qwen3-14B	Easy	**99.76**	95.62	97.02	91.94	98.52	**99.38**
	Hard	**51.94**	**84.18**	40.24	76.64	39.68	75.50
Qwen3-32B	Easy	**99.42**	95.20	98.68	94.74	98.54	**99.54**
	Hard	**37.32**	72.98	35.64	**75.12**	36.38	72.84
Llama-3-8B	Easy	**98.72**	81.52	96.32	**90.98**	92.76	90.60
	Hard	32.00	63.90	**36.68**	**74.50**	23.82	55.96
Llama-3.1-8B	Easy	**97.92**	85.26	95.66	**92.94**	96.68	86.80
	Hard	23.56	65.02	**34.72**	**81.56**	22.86	51.74

Table 3: Pairwise accuracy (%) on easy vs. hard pairs on DomainAuth. **Bold**: best; underline: second best, per LLM per difficulty level.

Appendix E.

4.3 Results on EntityAuth

Table 4 and Table 6 (in the Appendix B) show the performance of different methods on EntityAuth benchmark and using fine- and coarse-labels, respectively. From the tables, we can find that: (1) The ListJudge and PairJudge methods, when combined with PointScore, perform better than other methods and exhibit similar performance trends. Moreover, the cross-LLMs patterns on EntityAuth largely mirror those observed on DomainAuth. (2) Performance varies considerably across the three domains. LLMs achieve notably stronger and more consistent results on Basketball and Movies compared to Songs. The Basketball and Movies in Wikipedia links have roughly five times as many links as Songs. So, the performance difference may stem from differences in the volume of domains in the pre-training data of LLMs. (3) Compared to the performance on DomainAuth benchmark in Table 1& 2, LLMs generally exhibit stronger and more consistent authority perception on EntityAuth compared to DomainAuth, particularly in the Basketball and Movies domains. For instance, Qwen3-32B achieves a Spearman's ρ of 0.8594 on Basketball, higher than its best score of 0.7528 on DomainAuth. This performance difference can be attributed to the nature of the authority signal: entity authority is often more concrete, objective, and well-documented in structured formats (e.g., player statistics, filmographies) within the training data, making it an easier concept for models to grasp compared to the more abstract notion of website authority.

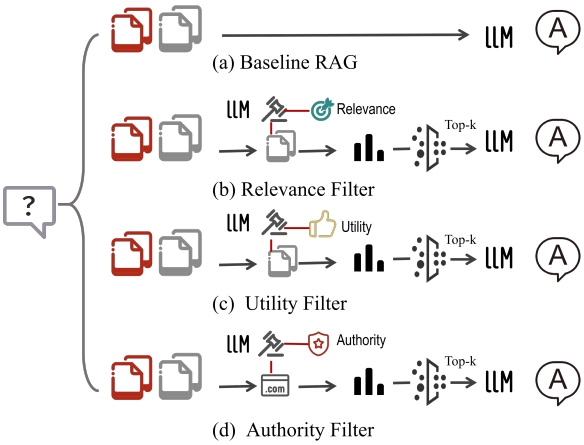

Figure 4: Authority-aware RAG pipeline: retrieved documents are LLM-scored; top-k go to the generator. Filtering signals: (a) **w/o Filter**; (b) **Relevance Filter**; (c) **Utility Filter**; (d) **Authority Filter** (source URL only; no document content). All prompts in this section are provided in Appendix A.

4.4 Authority-Aware RAG

To evaluate the impact of authority-awareness on RAG, we compare several document filtering strategies on our proposed RAGAuth. All strategies follow a unified pipeline: first, LLMs re-rank N documents based on different criteria; then, the top-k documents are passed to the final answer generation. While our findings indicate that PairJudge-PointScore generally achieves the best performance

Domain	Model	PointJudge		ListJudge				PairJudge			
				ListRank		PointScore		PairRank		PointScore	
		ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ
Basketball	Qwen3-8B	74.93	61.26	65.38	54.93	72.53	59.32	81.29	68.07	**86.70**	**73.90**
	Qwen3-14B	80.26	66.10	69.37	60.94	76.42	65.29	85.09	71.97	**85.89**	**73.07**
	Qwen3-32B	85.17	70.98	64.22	57.53	75.33	64.96	82.25	69.22	**85.94**	**73.66**
	Llama-3-8B-Instruct	76.00	67.33	59.91	49.13	75.54	65.94	83.54	70.66	**88.90**	**77.07**
	Llama-3.1-8B-Instruct	82.55	70.94	59.38	50.22	76.00	64.90	79.90	66.10	**87.43**	**76.45**
Movies	Qwen3-8B	81.25	66.02	78.30	63.91	82.40	67.22	61.33	47.18	**87.87**	**75.20**
	Qwen3-14B	78.62	62.77	87.70	74.62	85.12	70.74	86.19	72.75	**88.60**	**75.94**
	Qwen3-32B	85.30	70.69	87.76	74.77	87.67	74.60	83.27	69.16	**89.56**	**77.40**
	Llama-3-8B-Instruct	78.24	62.77	75.06	60.60	**84.91**	**71.50**	79.04	64.74	83.24	69.66
	Llama-3.1-8B-Instruct	81.33	66.37	80.69	66.28	**84.38**	**71.11**	70.00	55.42	82.72	69.32
Songs	Qwen3-8B	**46.56**	**34.96**	34.78	25.87	42.71	31.84	10.57	7.78	44.58	33.52
	Qwen3-14B	47.19	35.31	**51.75**	**38.64**	49.02	36.64	47.62	35.52	50.54	37.75
	Qwen3-32B	49.72	37.10	51.83	38.64	**53.24**	**39.84**	39.92	29.68	51.66	38.78
	Llama-3-8B-Instruct	49.92	37.46	45.98	34.28	**52.38**	**39.11**	25.62	18.90	48.33	35.95
	Llama-3.1-8B-Instruct	42.63	32.03	**49.69**	**37.16**	49.59	37.15	40.27	29.86	46.76	34.84

Table 4: Results on EntityAuth (fine-grained, 10-level, w/o text) across three domains reported as Spearman's ρ (%) and Kendall's τ (%). **Bold**: best; underline: second best, per LLM per domain.

in discerning authority, its high computational cost makes it impractical for real-time RAG applications. ListJudge, which scores all documents in a single pass, offers a more cost-effective trade-off. For a fair comparison across criteria, all strategies use ListJudge inputs and the PointScore protocol to rank all documents, as shown in Figure 4:

- **Baseline RAG:** No filtering is applied.
- **Relevance Filter:** The document is scored based on its relevance to the query.
- **Utility Filter:** Following Zhang et al. (2024b), the document is scored based on its utility for answering the question. The model is first prompted to generate a pseudo-answer from the document list, and then scores each document based on its utility in answering the question.
- **Authority Filter:** Score each document based on the perceived authority of its source using the ListJudge-PointScore method, without considering the query content.

RAG Performance. As shown in Table 5, we can find that: (1) All models demonstrate performance improvement when incorporating filtering methods based on different criteria. (2) Incorporating authority signals during filtering has the best RAG performance improvement across all strategies. This underscores the central role of source authority in enhancing answer accuracy and reducing susceptibility to low-credibility evidence.

Model	k	Relevance	Utility	Authority
Qwen3-8B	–	w/o Filter: 53.33		
	1	48.33	51.67	**66.67**
	3	51.67	58.33	**69.16**
	5	51.67	58.33	**69.16**
Qwen3-14B	–	w/o Filter: 58.33		
	1	51.67	60.00	**76.67**
	3	45.00	66.67	**75.00**
	5	45.00	66.67	**73.33**
Qwen3-32B	–	w/o Filter: 55.00		
	1	63.33	65.00	**70.00**
	3	58.33	63.33	**68.33**
	5	60.00	60.00	**70.00**
Llama-3-8B	–	w/o Filter: 50.83		
	1	37.50	49.17	**64.17**
	3	41.67	52.50	**64.17**
	5	41.67	52.50	**64.17**
Llama-3.1-8B	–	w/o Filter: 57.50		
	1	42.50	51.67	**63.34**
	3	55.00	48.33	**71.76**
	5	55.00	48.33	**71.76**

Table 5: Answer accuracy (%) on RAGAuth across four models under different filtering strategies and top-k settings ($k \in \{1, 3, 5\}$). **Bold**: best per LLM per k.

5 Conclusion

In this work, we introduced AuthorityBench, a comprehensive benchmark designed to evaluate the authority perception capabilities of LLMs. Through extensive experiments on our three datasets, DomainAuth, EntityAuth, and RAGAuth, our key findings are as follows: (1) ListJudge and PairJudge have the strongest performance compared to other methods. Moreover, the optimal method for eliciting authority judgments is context-dependent. (2) Finally, we demonstrated the practical value of authority perception in a downstream RAG task, where authority-aware filtering achieved the best answer accuracy compared to other methods. Our work provides a comprehensive benchmark for understanding and harnessing the authority perception of LLMs, a crucial step towards building more reliable and trustworthy AI systems.

6 Limitations

Our study has several limitations that offer avenues for future research. First is our reliance on PageRank as the primary ground truth for domain authority. While PageRank is a well-established, objective, and scalable proxy for authority based on the web's link structure, it is not without its flaws. It can be slow to adapt to the rise of new authoritative sources and may not fully capture the nuanced, topic-specific authority that a human expert might perceive. For instance, a niche but highly respected academic blog might have a lower PageRank than a mainstream news aggregator, yet be more authoritative on its specific topic. Moreover, certain highly authoritative official websites may receive disproportionately low PageRank scores due to sparse inbound links. We analyze this issue in detail in Appendix E. Future work can address this gap by integrating more dynamic, context-aware, or topic-specific authority metrics into the evaluation framework. Second, the scale of our RAGAuth dataset is modest due to the manual cost. While the 120 questions are sufficient to demonstrate the potential of authority-aware filtering, a larger and more diverse set of questions and documents would be needed to draw more robust conclusions about its general applicability and to analyze the performance on different types of questions.

7 Ethics Statement

In our work, the data and models we use are publicly available. We have transformed the original datasets to construct our AuthorityBench. Both the queries and documents for the RAGAuth are obtained from public web pages collected by authors and do not contain personal or sensitive information. To advance the community, we will release the benchmark we constructed along with the experimental code.

References

Jia Chen, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2019. Tiangong-st: A new dataset with large-scale refined real-world web search sessions. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 2485–2488.

Boyi Deng, Wenjie Wang, Fengbin Zhu, Qifan Wang, and Fuli Feng. 2025. Cram: Credibility-aware attention modification in llms for combating misinformation in rag. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23760–23768.

Mouly Dewan, Jiqun Liu, Aditya Gautam, and Chirag Shah. 2025. Llm-driven usefulness judgment for web search evaluation. *arXiv preprint arXiv:2504.14401*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jeongyeon Hwang, Junyoung Park, Hyejin Park, Dongwoo Kim, Sangdon Park, and Jungseul Ok. 2025. Retrieval-augmented generation with estimation of source reliability. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34267–34291.

Akriti Jain and Aparna Garimella. 2025. Modeling contextual passage utility for multihop question answering. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 464–471.

Zixuan Ke, Weize Kong, Cheng Li, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. 2024. Bridging the preference gap between retrievers and llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10438–10451.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, and 1 others. 2025. From generation to judgment: Opportunities and challenges of llm-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791.

Xueguang Ma, Xinyu Zhang, Ronak Pradeep, and Jimmy Lin. 2023. Zero-shot listwise document reranking with a large language model. *arXiv preprint arXiv:2305.02156*.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 9802–9822.

Preslav Nakov, Jisun An, Haewoon Kwak, Muhammad Arslan Manzoor, Zain Muhammad Mujahid, and Husrev Taha Sencar. 2024. A survey on predicting the factuality and the bias of news media. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15947–15962.

Shiyu Ni, Keping Bi, Jiafeng Guo, and Xueqi Cheng. 2025. How knowledge popularity influences and enhances llm knowledge boundary perception. *arXiv preprint arXiv:2505.17537*.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford infolab.

Ruotong Pan, Boxi Cao, Hongyu Lin, Xianpei Han, Jia Zheng, Sirui Wang, Xunliang Cai, and Le Sun. 2024. Not all contexts are equal: Teaching llms credibility-aware generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19844–19863.

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2016. Credibility assessment of textual claims on the web. In *Proceedings of the 25th ACM international on conference on information and knowledge management*, pages 2173–2178.

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2017. Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In *Proceedings of the 26th international conference on world wide web companion*, pages 1003–1012.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, and 1 others. 2024. Large language models are effective text rankers with pairwise ranking prompting. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1504–1518.

Soo Young Rieh. 2010. Credibility and cognitive authority of information. *Encyclopedia of library and information sciences*, 1(1):1337–1344.

Ian Rogers. 2002. The google pagerank algorithm and how it works.

Michael Schlichtkrull. 2024. Generating media background checks for automated source critical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4927–4947.

Charles Spearman. 1961. The proof and measurement of association between two things.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is chatgpt good at search? investigating large language models as re-ranking agents. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 14918–14937.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 2: short papers)*, pages 422–426.

Patrick Wilson. 1983. Second-hand knowledge: An inquiry into cognitive authority.

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 preprint arXiv:2505.09388*.

Hengran Zhang, Keping Bi, Jiafeng Guo, and Xueqi Cheng. 2024a. Iterative utility judgment framework via llms inspired by relevance in philosophy. *arXiv preprint arXiv:2406.11290*.

Hengran Zhang, Minghao Tang, Keping Bi, Jiafeng Guo, Shihao Liu, Daiting Shi, Dawei Yin, and Xueqi Cheng. 2025. Utility-focused llm annotation for retrieval and retrieval-augmented generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1683–1702.

Hengran Zhang, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. 2024b. Are large language models good at utility judgments? In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1941–1951.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Shengyao Zhuang, Honglei Zhuang, Bevan Koopman, and Guido Zuccon. 2024. A setwise approach for effective and highly efficient zero-shot ranking with large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 38–47.

A Prompts

Prompts for DomainAuth

A. DomainAuth / PointJudge (w/o text)

You are a web analysis expert. Please score the source authority of the given domain based on the following criteria, with a rating range of 0 to 9.

Source authority refers to the credibility, professionalism, official certification, or public recognition of the domain itself.

DOMAIN: [Domain name]

Based on the above criteria, rate the source authority of this DOMAIN (0–9 points).

Output only: {"score": X} where X is your evaluated score (0, 1, 2, 3, 4, 5, 6, 7, 8, or 9). Do not add any explanation.

Your answer:

B. DomainAuth / PointJudge (w/ text)

You are a web analysis expert. Please score the source authority of the given domain based on the following criteria, with a rating range of 0 to 9.

Source authority refers to the credibility, professionalism, official certification, or public recognition of the domain itself.

DOMAIN: [Domain name]

Webpage Text: [Webpage Text]

Based on the domain name, their web content and the above criteria, rate the source authority of this DOMAIN (0–9 points).

Output only: {"score": X} where X is your evaluated score (0, 1, 2, 3, 4, 5, 6, 7, 8, or 9). Do not add any explanation.

Your answer:

C. DomainAuth / PairJudge-PairRank (w/o text)

You are a web analysis expert. Please determine which of the two given domains is more authoritative as an information source based on the following criteria.

Source authority refers to the credibility, professionalism, official certification, or public recognition of the domain itself.

Consider both domains together when making the comparison.

DOMAINS:

Domain A: [Domain_A]

Domain B: [Domain_B]

Which domain is more authoritative? Only return 'A' or 'B'. Do not add any explanation.

Your answer:

D. DomainAuth / PairJudge-PairRank (w/ text)

You are a web analysis expert. Please determine which of the two given domains is more authoritative as an information source based on the following criteria.

Source authority refers to the credibility, professionalism, official certification, or public recognition of the domain itself and its content.

Consider both domains and their contents together when making the comparison.

DOMAINS & CONTENTS:
Domain A: [Domain A]
Content A: [Text A]
Domain B: [Domain B]
Content B: [Text B]

Which domain is more authoritative? Only return 'A' or 'B'. Do not add any explanation.

Your answer:

E. DomainAuth / PairJudge-PointScore

You are a web analysis expert. Please score the source authority of each of the two given domains based on the following criteria, with a rating range of 0 to 9.

Source authority refers to the credibility, professionalism, official certification, or public recognition of the domain itself.

Consider both domains together when assigning scores to ensure consistency.

DOMAINS:
Domain A: [Domain A]
Domain B: [Domain B]

Your output MUST be a single, valid JSON object in the format: {"A": X, "B": X}, where each X is an integer score from 0 to 9.

Example format: {"A": 7, "B": 4}. Do not add any explanation.

Your answer:

F. DomainAuth / ListJudge-ListRank (w/ text)

You are a web analysis expert that can rank domains based on their authority.

I will provide you with 10 domains and their web contents, each indicated by a numerical identifier []. Please rank the domains in the provided list based on the following criteria.

Source authority refers to the credibility, professionalism, official certification, or public recognition of the domain itself.

DOMAINS:
[0] [Domains_0]: [Text_0]
[1] [Domains_1]: [Text_1]
...

Rank the 10 domains above based on their authority (highest to lowest). All the domains should be included and listed using identifiers, in descending order of authority.

The output format should be [] > [], e.g., [4] > [0].

Only respond with the ranking results. Do not add any explanation.

Your answer:

G. DomainAuth / ListJudge-PointScore(w/o text)

You are a web analysis expert. Please score the source authority of each domain in the provided list based on the following criteria, with a rating range of 0 to 9.

Source authority refers to the credibility, professionalism, official certification, or public recognition of the domain itself.

Consider all domains together when assigning scores to ensure consistency and relative calibration.

DOMAINS:
[0] [Domains_0]: [Text_0]
[1] [Domains_1]: [Text_1]
...

Your output MUST be a single, valid JSON object. The JSON object must map each domain's numerical identifier (as a string) to its integer score (0–9).

Example format: {"0": 7, "1": 4, ..., "9": 8}. Do not add any explanation.

Your answer:

Prompts for EntityAuth

H. EntityAuth / PointJudge

You are a Wikipedia notability expert and a music industry expert. Please score the authority of the given song entity based on the following criteria, with a rating range of 0 to 9.

Entity authority refers to the public recognition, fame, historical significance, or widespread knowledge of the entity (person, place, organization, etc.).

QUESTION CONTEXT: [Question]

TARGET ENTITY: [Question_entity]

Based on the question context and the above criteria, rate the authority of this ENTITY (0–9 points).

Output only: {"score": X} where X is your evaluated score (0, 1, 2, 3, 4, 5, 6, 7, 8, or 9). Do not add any explanation.

Your answer:

I. EntityAuth / PairJudge-PairRank

You are a Wikipedia notability expert and a music industry expert. Please determine which of the two given entities is more authoritative based on the following criteria.

Entity authority refers to the public recognition, fame, historical significance, or widespread knowledge of the entity (person, place, organization, etc.).

Consider both entities together when assigning scores to ensure consistency.

QUESTION CONTEXT: These entities are the subjects of questions like 'Who is the performer of the song ...'

TARGET ENTITIES:
Entity A: "[Entity_A]"
Entity B: "[Entity_B]"

Which entity is more authoritative? Only return 'A' or 'B'. Do not add any explanation.

Your answer:

J. EntityAuth / PairJudge-PointScore

You are a Wikipedia notability expert and a music industry expert. Please score the entity authority of each of the two given song entities based on the following criteria, with a rating range of 0 to 9.

Entity authority refers to the public recognition, fame, historical significance, or widespread knowledge of the entity (person, place, organization, etc.).

Consider both entities together when assigning scores to ensure consistency.

QUESTION CONTEXT: These entities are the subjects of questions like 'Who is the performer of the song...'

TARGET ENTITIES:
Entity A: "[Entity_A]"
Entity B: "[Entity_B]"

Your output MUST be a single, valid JSON object in the format: {"A": X, "B": Y}, where X and Y are integer scores from 0 to 9.

Example format: {"A": 7, "B": 4}. Do not add any explanation.

Your answer:

K. EntityAuth / ListJudge-ListRank

You are a Wikipedia notability expert and a music industry expert. I will provide you with 10 song entities, each indicated by a numerical identifier []. Please rank the song entities in the provided list based on the following criteria.

Entity authority refers to the public recognition, fame, historical significance, or widespread knowledge of the entity (person, place, organization, etc.).

Consider all entities together when assigning scores to ensure consistency and relative calibration.

QUESTION CONTEXT: These entities are the subjects of questions like 'Who is the performer of the song...'

TARGET ENTITIES:
[0] [Question_entity_0]
[1] [Question_entity_1]
...

Rank the 10 song entities above based on their authority (highest to lowest). All the song entities should be included and listed using identifiers, in descending order of authority.

The output format should be [] > [], e.g., [4] > [0].

Only respond with the ranking results, do not add any explanation.

Your answer:

L. EntityAuth / ListJudge-PointScore

You are a Wikipedia notability expert and a music industry expert. Please score the entity authority of each song entity in the provided list based on the following criteria, with a rating range of 0 to 9.

Entity authority refers to the public recognition, fame, historical significance, or widespread knowledge of the entity (person, place, organization, etc.).

Consider all entities together when assigning scores to ensure consistency and relative calibration.

QUESTION CONTEXT: These entities are the subjects of questions like 'Who is the performer of the song...'

TARGET ENTITIES:
[0] [Question_entity_0]
[1] [Question_entity_1]
...

Your output MUST be a single, valid JSON object. The JSON object must map each entity's numerical identifier (as a string) to its integer score (0–9).

Example format: {"0": 7, "1": 4, ..., "9": 8}. Do not add any explanation.

Your answer:

Prompts for RAGAuth

Strategy a: w/o Filter / Generate Final Answer

You are given a question and several pieces of documents. Your task is to analyze the documents and provide a concise answer to the question based solely on the provided reference documents.

Question: [Question]

Reference Documents:
Document [0]
Source URL: [URL]
Content: [Doc_text]
...

Please answer with ONLY 'yes' or 'no' based on the documents above. Do not add any explanation.

Your answer:

Strategy b: Relevance Filter — Scoring Prompt

You are a web analysis expert. Please score the relevance of each document to the question with a rating range of 0 to 9, where 0 is the lowest and 9 is the highest.
Consider all documents together when assigning scores to ensure consistency and relative calibration.

QUESTION: [Question]

DOCUMENTS:
[0] [URL]: [Doc_text]
[1] [URL]: [Doc_text]
...

Your output MUST be a single, valid JSON object. The JSON object must map each document's numerical identifier (as a string) to its integer score (0–9). Example format: {"0": 7, "1": 4, ..., "9": 8}. Do not add any explanation.

Your answer:

Strategy c: Utility Filter — Scoring Prompt

You are a web analysis expert. Please score the utility of each document in answering the question based on the following criteria, with a rating range of 0 to 9, where 0 is the lowest and 9 is the highest.
The passage has utility in answering the question, meaning that the passage not only is relevant to the question, but also is useful in generating a correct, reasonable, and perfect answer to the question.
Consider all documents together when assigning scores to ensure consistency and relative calibration.

QUESTION: [Question]

DOCUMENTS:
[0]: [URL]: [Doc_text]
[1]: [URL]: [Doc_text]
...

Step 1: Generate a pseudo-answer to the question based on all the documents.

Step 2: Based on the pseudo-answers, rate each document's utility in answering the question on a scale of 0 to 9.

Your output MUST be a single, valid JSON object with two keys:

- `"answers"`: maps each document's numerical identifier (as a string) to its pseudo-answer
- `"scores"`: maps each document's numerical identifier (as a string) to its integer score (0–9)

Example format: {"answers": {"0": "...", "1": "...", ...}, "scores": {"0": 7, "1": 4, ...}}. Do not add any explanation.

Your answer:


```
{
  "qid": "MED_EN_001",
  "question": "Can vitamin D effectively prevent heart disease and cancer?",
  "question_type": "yes_no",
  "language": "en",
  "ground_truth": "No",
  "documents": [
    {
      "docid": "doc_001",
      "url": "https://www.cncbd.org.cn/News/Detail/8157",
      "domain": "cncbd.org.cn",
      "PageRank": 0,
      "doc_text": "...",
      "authority": "low"
    },
    ...
  ]
}
```


Figure 5: An example from the RAGAuth dataset. Each instance includes a yes/no question, the ground-truth answer, and a list of 10 retrieved documents with their source URL, domain, PageRank score, and text snippet. The task is to generate a correct answer based on this information.

Strategy d: Authority Filter — Scoring Prompt

You are a web analysis expert. Please score the source authority of each source URL in the provided list based on the following criteria, with a rating range of 0 to 9, where 0 is the lowest and 9 is the highest. Source authority refers to the credibility, professionalism, official certification, or public recognition of the domain itself and its content. Consider all URLs together when assigning scores to ensure consistency and relative calibration.

SOURCES:
[0]: [URL]
[1]: [URL]
...

Your output MUST be a single, valid JSON object. The JSON object must map each domain's numerical identifier (as a string) to its integer score (0–9). Example format: {"0": 7, "1": 4, ..., "9": 8}. Do not add any explanation.

Your answer:

B Coarse-Grained Results

This section shows the results of EntityAuth under coarse-grained, as shown in Table 6.

C An Example from RAGAuth Dataset

Figure 5 shows an example from RAGAuth dataset.

D Benchmark Statistics

Table 7 shows the statistics details of Authority-Bench.

E Analysis of Scoring Bias

In this section, we analyze the overall scoring behavior of the PointJudge method and its discrepan-

Domain	Model	PointJudge		ListJudge				PairJudge			
				ListRank		PointScore		PairRank		PointScore	
		ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ
Basketball	Qwen3-8B	70.97	60.04	50.60	44.77	66.64	56.60	29.52	25.73	**84.19**	**74.97**
	Qwen3-14B	78.08	67.70	70.10	64.93	71.52	63.60	**86.50**	**78.39**	83.80	75.04
	Qwen3-32B	79.67	68.93	62.53	60.06	73.39	65.96	**84.76**	**76.81**	83.94	76.22
	Llama-3-8B-Instruct	73.76	67.14	65.38	58.05	68.86	62.50	83.91	76.56	**87.46**	**80.23**
	Llama-3.1-8B-Instruct	78.28	69.31	57.09	51.54	76.34	68.21	78.53	70.60	**86.52**	**77.91**
Movies	Qwen3-8B	66.25	55.08	51.61	45.30	71.94	60.96	48.32	40.90	**86.75**	**78.75**
	Qwen3-14B	75.17	63.49	88.56	80.64	82.07	71.12	**89.03**	**81.72**	88.77	80.89
	Qwen3-32B	76.49	65.26	88.42	80.66	84.31	74.11	85.90	77.85	**90.27**	**83.40**
	Llama-3-8B-Instruct	76.49	65.26	75.77	65.11	76.22	65.80	80.84	71.60	**87.42**	**79.68**
	Llama-3.1-8B-Instruct	58.45	48.15	78.30	68.03	80.09	69.60	67.35	57.21	**80.84**	**71.07**
Songs	Qwen3-8B	**42.51**	**34.34**	23.80	19.45	30.11	24.32	24.58	19.93	41.01	33.40
	Qwen3-14B	45.67	37.09	**51.76**	**42.34**	47.28	38.39	48.89	39.92	50.77	41.40
	Qwen3-32B	47.71	38.80	**52.25**	**42.79**	50.50	41.07	39.60	32.23	51.42	42.16
	Llama-3-8B-Instruct	48.41	39.85	45.27	36.88	43.74	35.51	40.20	32.71	**49.65**	**40.55**
	Llama-3.1-8B-Instruct	40.46	32.83	45.36	36.97	46.09	37.75	19.50	15.64	**48.41**	**39.54**

Table 6: Results on EntityAuth (coarse-grained, 5-level, w/o text) across three domains reported as Spearman's ρ and Kendall's τ (%). **Bold**: best; underline: second best, per model per domain.

Sub-task	Domain	Authority Proxy	Score Scale	Corpus	Listwise (N=5/10)†	Pairwise (Hard/Easy)‡
DomainAuth	Web Domains	PageRank	0–9 / 0–4	10,000	4.5k / 4.5k	5k / 5k
EntityAuth	Basketball	Sitelinks	0–9	10,000	– / 4.5k	–
	Movies			10,000	– / 4.5k	–
	Songs			2,000	– / 4.5k	–
RAGAuth	Medical Science Current Affairs	–	N/A	120 Queries	10 docs/query	

Table 7: Overview of the AuthorityBench. **Corpus** is the number of unique items. †**Listwise** instances are split between N=5 and N=10 formats. ‡**Pairwise** instances are split between hard ($0 < \Delta$PageRank≤ 2) and easy (ΔPageRank≥ 5) pairs. EntityAuth uses only N=10 lists and does not include a Pairwise setting. For RAGAuth, each query is paired with 10 retrieved webpage documents.

cies against the ground-truth authority proxy. Figure 6 illustrates the bias in the average predicted score for each true authority score (0-9) on DomainAuth. A consistent trend is observed across all models: they tend to overestimate the authority of low-authority sources while underestimating that of high-authority sources to a milder degree. Such a "regression to the mean" phenomenon reveals that, although PointJudge achieves competitive ranking performance, there still exist noticeable discrepancies between its predicted scores and the true authority proxy.

To further investigate this bias, we perform a fine-grained analysis on three typical domain categories: government (`.gov`), education (`.edu`), and social media. Table 8 reports the mean scoring bias for each category

Effective Identification of .gov/.edu Authority. Across all models, government and education domains consistently receive high authority scores

Table 8: Mean predicted score and scoring bias across three domain categories. *Bias* = avg. (predicted − true PageRank). True means: Gov 5.68 (N=378); Edu 5.64 (N=160); Social 6.00 (N=34).

Model	Ctx.	.gov		.edu		Social	
		Pred	Bias	Pred	Bias	Pred	Bias
Qwen3-14B	w/o	8.97	+3.28	8.84	+3.20	8.44	+2.44
	w/	8.39	+2.70	8.25	+2.61	7.59	+1.59
Qwen3-32B	w/o	8.92	+3.24	8.02	+2.38	7.68	+1.68
	w/	8.25	+2.57	7.26	+1.62	6.79	+0.79
Llama-3-8B	w/o	8.20	+2.52	7.66	+2.02	6.74	+0.74
	w/	7.93	+2.25	7.39	+1.75	5.94	−0.06
Llama-3.1-8B	w/o	8.71	+3.03	8.03	+2.39	7.53	+1.53
	w/	8.09	+2.41	7.74	+2.10	5.79	−0.21

(often in the 8-9 range), leading to notable positive divergence from their PageRank values (e.g., Qwen3-14B yields a mean divergence of +3.28 for `.gov` and +3.20 for `.edu`). As registration of `.gov` and `.edu` domains is strictly restricted to official government and educational entities, `.gov` and

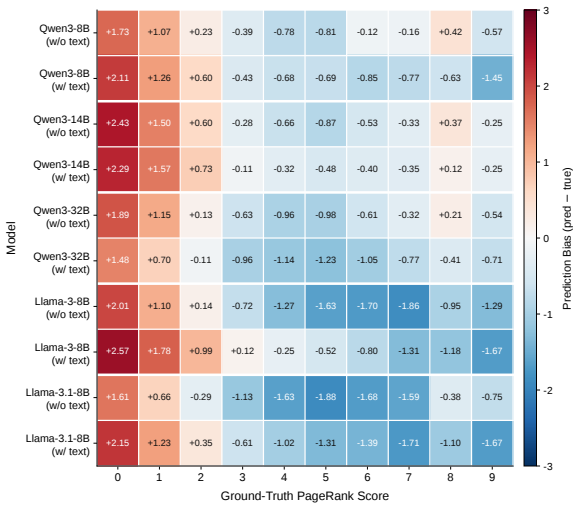

Figure 6: Heatmap of prediction bias (predicted - true score) on DomainAuth for the Pointwise-Judge. Red cells indicate overestimation, while blue cells indicate underestimation. Most models tend to overestimate low-authority sources and underestimate high-authority ones.

`.edu` can be regarded as inherently high-authority TLDs. The models' high scoring for such domains thus reflects accurate authority judgment. This divergence further reflects a critical limitation of PageRank as an authority proxy: it fails to capture intrinsic authority unaccompanied by a dense hyperlink structure. For instance, a local government website may have low PageRank due to insufficient external links, yet remains the definitive authority within its domain.

Overestimation Bias on Social Media Domains. For social media platforms, the degree of overestimation varies substantially across models, ranging from +0.74 (Llama-3-8B, w/o text) to +2.44 (Qwen3-14B, w/o text). Notably, incorporating webpage text mitigates the overestimation of domain authority for most models: Llama-3.1-8B's bias drops from +1.53 to -0.21, and Llama-3-8B achieves near-zero bias (-0.06) with webpage text, likely because the content itself is not authoritative.

Despite basic security measures offered by mature commercial operators, social media domains are inherently low-authority sources. As open platforms hosting user-generated content (UGC) from heterogeneous and unvetted publishers without mandatory fact-checking or official verification mechanisms, they lack institutional authority by nature. The overestimation without webpage text appears driven by brand recognition—well-known platforms such as `jianshu.com` and `spotify.com` (both with PageRank=0 in our dataset) receive high scores because the models associate these brand names with popularity, overriding the structural authority signal. This suggests a practical implication: when deploying PointJudge on social media content without access to page text, applying a slightly higher filtering threshold can help counteract this systematic overestimation.

Table 9: PairJudge ranking derivation methods on DomainAuth (fine-grained). Spearman's ρ (%). **Bold**: best; underline: second best, per row.

Model	Ctx.	PairRank	$\text{PointScore}_{\text{Bbl}}$	$\text{PointScore}_{\text{Pt}}$
Qwen3-8B	w/o	44.19	64.61	**71.35**
	w/	53.05	61.83	**63.91**
Qwen3-14B	w/o	70.21	69.36	**73.43**
	w/	64.78	64.65	**67.99**
Qwen3-32B	w/o	72.10	72.52	**75.28**
	w/	66.32	67.26	**69.93**
Llama-3-8B	w/o	61.05	59.70	**64.83**
	w/	49.06	59.75	**64.97**
Llama-3.1-8B	w/o	**63.82**	60.17	63.24
	w/	51.96	60.01	**62.88**

Table 10: PairJudge ranking derivation methods on DomainAuth (coarse-grained). Spearman's ρ (%). **Bold**: best; underline: second best, per row.

Model	Ctx.	PairRank	$\text{PointScore}_{\text{Bbl}}$	$\text{PointScore}_{\text{Pt}}$
Qwen3-8B	w/o	13.34	64.71	**70.90**
	w/	8.73	63.12	**64.84**
Qwen3-14B	w/o	65.65	71.73	**73.16**
	w/	**68.30**	65.21	66.55
Qwen3-32B	w/o	72.88	74.72	**75.52**
	w/	**69.82**	65.21	66.55
Llama-3-8B	w/o	55.09	61.25	**64.30**
	w/	52.04	54.92	**56.07**
Llama-3.1-8B	w/o	63.84	61.23	**64.26**
	w/	57.08	58.12	**58.69**

Case Study: Qwen3-32B PointJudge Scoring (w/o text)

Limitation of PageRank for Official Domains
Local government and educational sites receive low PageRank due to sparse inbound links, yet the model correctly recognizes their institutional authority:

- `xjedu.gov.cn` (Xinjiang Edu. Dept.): PageRank 2 → Pred **9** (+7)
- `bnuz.edu.cn` (BNU Zhuhai Campus): PageRank 3 → Pred **8** (+5)

Overestimation on Social Media
Well-known social media platforms are overestimated due to brand recognition, despite being inherently low-authority open platforms:

- `zhihu.com` (Zhihu): PageRank 6 → Pred **8** (+2)
- `tieba.com` (Baidu Tieba): PageRank 0 → Pred **6** (+6)

Table 11: PairJudge ranking derivation methods on EntityAuth (fine-grained). Spearman's ρ (%). **Bold**: best; underline: second best, per row.

Domain	Model	PairRank	$\text{PointScore}_{\text{Bbl}}$	$\text{PointScore}_{\text{Pt}}$
Basketball	Qwen3-8B	81.29	83.28	**86.67**
	Qwen3-14B	85.09	82.16	**85.89**
	Qwen3-32B	82.25	82.77	**85.94**
	Llama-3-8B	83.54	87.14	**88.90**
	Llama-3.1-8B	79.29	84.59	**87.43**
Movies	Qwen3-8B	61.33	81.51	**87.87**
	Qwen3-14B	86.19	86.45	**88.60**
	Qwen3-32B	83.27	87.02	**89.56**
	Llama-3-8B	79.04	82.19	**83.24**
	Llama-3.1-8B	70.00	81.00	**82.72**
Songs	Qwen3-8B	10.57	37.06	**44.58**
	Qwen3-14B	47.62	48.04	**50.54**
	Qwen3-32B	39.92	49.88	**51.66**
	Llama-3-8B	25.62	46.49	**48.33**
	Llama-3.1-8B	40.27	45.87	**46.76**

Table 12: PairJudge ranking derivation methods on EntityAuth (coarse-grained). Spearman's ρ (%). **Bold**: best; underline: second best, per row.

Domain	Model	PairRank	$\text{PointScore}_{\text{Bbl}}$	$\text{PointScore}_{\text{Pt}}$
Basketball	Qwen3-8B	29.52	83.50	**84.19**
	Qwen3-14B	**86.50**	82.84	83.80
	Qwen3-32B	**84.76**	84.21	83.94
	Llama-3-8B	83.91	86.92	**87.46**
	Llama-3.1-8B	78.53	85.83	**86.52**
Movies	Qwen3-8B	48.32	65.22	**86.75**
	Qwen3-14B	**89.03**	88.66	88.77
	Qwen3-32B	85.90	90.15	**90.27**
	Llama-3-8B	80.84	87.04	**87.42**
	Llama-3.1-8B	67.35	77.78	**80.84**
Songs	Qwen3-8B	24.58	**41.01**	39.27
	Qwen3-14B	48.89	49.14	**50.77**
	Qwen3-32B	39.60	51.22	**51.42**
	Llama-3-8B	40.20	49.58	**49.65**
	Llama-3.1-8B	19.50	46.32	**48.41**

F Comparison of PairJudge Ranking Derivation Methods

To validate our choice of ranking derivation method for `PairJudge`, we compare three distinct approaches for converting pairwise judgments into a listwise ranking:

- **PairRank (Bubble Sort)**: Applies bubble sort directly to the binary outcomes (A > B or B > A) of pairwise comparisons.
- **PointScore (Bubble Sort)**: First elicits a score for each item in a pair (e.g., A:7, B:4), then applies bubble sort based on which item received the higher score in each comparison.
- **PointScore (Point-based)**: Ranks items based on their average score received across all comparisons. This is the method we refer to as **PointScore** in the main paper.

Tables 9, 10, 11 and 12 present the results across all datasets. The point-based **PointScore** method consistently outperforms the other two, especially the **PairRank** bubble sort. This is because averaging scores is more robust to the transitivity errors that can arise from isolated, inconsistent pairwise judgments, thus providing a more stable and reliable final ranking.