From Relevance to Authority: Authority-aware Generative Retrieval in Web Search Engines

Sunkyung Lee[1*], **Jihye Back**[2*], **Donghyeon Jeon**[2], **Soonhwan Kwon**[2], **Moonkwon Kim**[2], **Inho Kang**[2], **Jongwuk Lee**[1†]
[1]Sungkyunkwan University, Republic of Korea, [2]Naver Corporation, Republic of Korea
[1]{sk1027, jongwuklee}@skku.edu, [2]{1oojihye, donghyeon.jeon, soonhwan2.kwon, moonkwon.kim, once.ihkang}@navercorp.com

Abstract

Generative information retrieval (GenIR) formulates the retrieval process as a text-to-text generation task, leveraging the vast knowledge of large language models. However, existing works primarily optimize for relevance while often overlooking document trustworthiness. This is critical in high-stakes domains like healthcare and finance, where relying solely on semantic relevance risks retrieving unreliable information. To address this, we propose an ***Auth**ority-aware **G**enerative **R**etriever (**AuthGR**)*, the first framework that incorporates authority into GenIR. AuthGR consists of three key components: (i) *Multimodal Authority Scoring*, which employs a vision-language model to quantify authority from textual and visual cues; (ii) a *Three-stage Training Pipeline* to progressively instill authority awareness into the retriever; and (iii) a *Hybrid Ensemble Pipeline* for robust deployment. Offline evaluations demonstrate that AuthGR successfully enhances both authority and accuracy, with our 3B model matching a 14B baseline. Crucially, large-scale online A/B tests and human evaluations conducted on the commercial web search platform confirm significant improvements in real-world user engagement and reliability.

1 Introduction

Generative Information Retrieval (GenIR) has emerged as a promising paradigm for the retrieval task, driven by recent advances in large language models (LLMs) (Metzler et al., 2021). Unlike traditional methods that encode queries and documents into vector representations (Karpukhin et al., 2020), GenIR reformulates retrieval as a text generation task (Tay et al., 2022; Zeng et al., 2024b). It aims to directly generate *document identifiers (DocIDs)* that satisfy users' information needs. Recently, GenIR has also gained much attention in industrial applications beyond academic research, *e.g.*, e-commerce search (Pang et al., 2025; Wu et al., 2024), food delivery (Zhang et al., 2025b), and financial services (Shen et al., 2025), boosting both accuracy and user satisfaction.

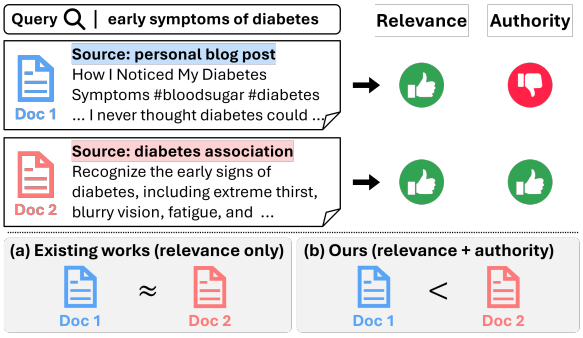

Figure 1: Illustration of our motivation. (a) Models relying solely on relevance fail to distinguish an unreliable blog from authoritative ones. (b) By integrating authority, our model can prioritize trustworthy documents.

Despite these advancements, existing GenIR methods primarily optimize semantic relevance, overlooking *document authority*. This poses risks in high-stakes domains like healthcare and finance for everyday users on commercial search engines, where trustworthiness is essential. As illustrated in Figure 1, a relevance-only model may rank an unverified personal health blog (Doc 1) as highly as an official medical association (Doc 2) solely due to topical similarity (Figure 1a). To prevent exposing users to potentially inaccurate or unverified information, it is essential to move beyond relevance and incorporate authority, ensuring the model to prioritize authoritative sources (Figure 1b).

However, explicitly integrating document authority into GenIR faces three challenges. (i) *Defining authority*: Textual cues alone often fail to distinguish trustworthy sources from sophisticated promotional websites, making it difficult to define authority at scale. (ii) *Learning authority*: Instilling the subtle and complex concept of authority

[*] Equal contribution
[†] Corresponding author

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track), pages 796–811
July 2-7, 2026

without compromising semantic relevance is non-trivial, requiring training methods beyond standard fine-tuning. (iii) *Deploying the authority-aware model*: Finally, simply replacing existing production rankers is impractical and risky. Instead, the model must be seamlessly integrated into current pipelines of a large-scale search platform, enhancing trustworthiness maintaining relevance.

To this end, we propose an ***Auth**ority-aware **G**enerative **R**etriever (**AuthGR**)*, which explicitly incorporates the document authority into the generative retrieval model. AuthGR consists of three key components: (i) We introduce *Multimodal Authority Scoring* to establish a scalable and quantifiable definition of authority. This component employs a vision-language model to assess site trustworthiness using textual and visual signals, effectively automating human-like judgments. (ii) To effectively learn the abstract concept of authority, our model is trained via a progressive *Three-stage Training Pipeline*. The process begins with domain-continued pre-training for foundational knowledge of the search domain, followed by supervised fine-tuning to learn the core task of DocID generation. The *group relative policy optimization (GRPO)* (Shao et al., 2024) stage lastly refines the model, explicitly training it to prefer high-authority documents. (iii) Finally, AuthGR is integrated using a *Hybrid Ensemble Pipeline* with existing rankers, enabling seamless real-world deployment without compromising relevance.

To summarize, our contributions are as follows:

- **First authority-aware GenIR**: To our knowledge, this is the first work to systematically incorporate authority into GenIR, addressing the need for trustworthiness.
- **Comprehensive training framework**: We introduce a three-stage training strategy that progressively instills the concept of authority using multimodal scores and preference optimization.
- **Real-world impact**: AuthGR is thoroughly validated across three complementary evaluations: it matches a 14B baseline that is 4.7× larger in offline settings; it boosts user engagement in large-scale online A/B tests; and it achieves superior quality ratings in human evaluations.

2 Related Work

2.1 Generative Information Retrieval

Generative Information Retrieval (GenIR) directly generates document identifiers (DocIDs) (Metzler et al., 2021; Li et al., 2024a). Early research explored diverse DocIDs, such as numeric IDs (Tay et al., 2022), URLs (Ren et al., 2023), or semantic keywords (Lee et al., 2023). Subsequent work shifted towards optimizing ranking quality via ranking-aware loss (Li et al., 2024b; Mekonnen et al., 2025) or reinforcement learning (Zhou et al., 2023; Wen et al., 2025). Despite advancements, GenIR predominantly focuses on *semantic relevance*, overlooking *document authority*. This is critical as GenIR expands into industrial domains, *e.g.*, finance (Shen et al., 2025; Chen et al., 2024) or e-commerce (Pang et al., 2025), where prioritizing relevance over authority exposes users to potentially inaccurate content. See Appendix C for broader discussion.

2.2 Trustworthiness in Information Retrieval

Traditional information retrieval has long emphasized *trustworthiness* by quantifying *authority* via link structures such as PageRank (Brin and Page, 1998) and HITS (Kleinberg, 1999), as well as content signals like TrustRank (Gyöngyi et al., 2004; Dong et al., 2015). Recent initiatives such as the TREC Health Misinformation Track (Clarke et al., 2020, 2021) further underscore its necessity in high-stakes domains, requiring diverse credibility indicators including source expertise (Zhang et al., 2022). Despite this rich history, GenIR systems have yet to adapt these sophisticated assessment methods. Our work bridges this gap by systematically incorporating authority signals into generative retrieval, ensuring reliability for real-world deployment.

3 Proposed Method

We propose an ***Auth**ority-aware **G**enerative **R**etriever (**AuthGR**)*, the first framework to systematically integrate authority into GenIR. As depicted in Figure 2, AuthGR comprises three key components: (i) **Multimodal Authority Scoring** to quantify trustworthiness via textual and visual cues; (ii) a **Three-Stage Training Pipeline** to progressively instill authority awareness; and (iii) a **Hybrid Ensemble Pipeline** for robust deployment.

3.1 Multimodal Authority Scoring

To quantify document authority at scale, we propose Multimodal Authority Scoring that automates human-like assessment. Traditional systems rely on fragmentary signals like link structures (Brin and Page, 1998), which often fail to capture trustworthiness holistically. In contrast, human judgment

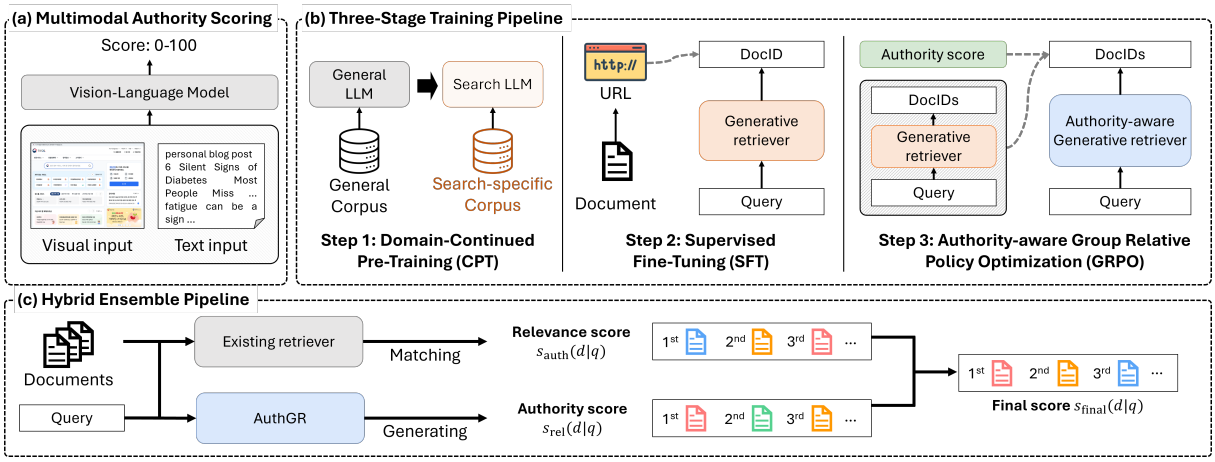

Figure 2: Overall architecture of AuthGR. (a) *Multimodal Authority Scoring* quantifies document trustworthiness based on textual and visual signals. (b) *Three-stage training pipeline* progressively instills authority awareness into the retriever. (c) The *Hybrid Ensemble Pipeline* integrates AuthGR with existing rankers for deployment.

integrates textual content, visual design, and advertisement patterns. To replicate this intuition, we employ a Vision-Language Model (VLM) as a scalable proxy for human evaluators.

Concretely, the VLM jointly processes: (i) *textual signals* including document title, body text, and URL metadata; and (ii) *visual signals* from a page-level screenshot. This integration is vital since promotional content often mimics authoritative language, making text alone deceptive. Visual cues, *e.g.*, ad intrusiveness and layout quality, are essential to distinguish genuine credibility from sophisticated mimicry, as illustrated in Appendix A.5. We prompt the VLM using a comprehensive rubric that evaluates the three core dimensions of Expertise, Officialness, and Public Interest, further supplemented by checks for commercial intent and harmfulness (*e.g.*, spam, illegal content)[1]. Formally, the authority score for a document d is defined as:

$$\texttt{Authority}(d) = f_{\text{VLM}}(T(d), V(d)) \in [0, 100], \quad (1)$$

where f_{VLM} is the scoring function from the VLM, and $T(d)$ and $V(d)$ denote textual and visual features. The model yields both a score and a concise natural-language rationale. The score subsequently serves as rewards during the group relative policy optimization stage (Section 3.2.3) to explicitly prioritize authoritative documents. Our validation also confirms that authority scores align strongly with human judgment (see Appendix A.2 for details).

[1]The full prompt is in Appendix A.1.

3.2 Three-Stage Training Pipeline

We systematically embed authority into retrievers with three distinct stages. For the generation target, we utilize host-level URLs as document identifiers[2]. The host-level granularity minimizes noise and exposes source identity, providing a stable foundation for authority modeling.

3.2.1 Domain-Continued Pre-Training (CPT)

The first stage adapts a general-purpose LLM to the search domain through domain-continued pre-training. This stage bridges the gap between broad linguistic knowledge and structured query–document correlations, as pointed out in the prior work (Ye et al., 2025)[3]. Specifically, we leverage large-scale search logs formatted as `[Query + URL + Title + Body]`. This structure enables the model to internalize associations between content and source identity. For instance, it learns that domains like "`.gov`" correlate with official institutions. Formally, the model parameters θ is updated via the standard language modeling objective:

$$\mathcal{L}_{\text{CPT}} = -\sum_{t} \log p_\theta(x_t \mid x_{<t}), \quad (2)$$

where x_t denotes the t-th token in the concatenated sequence. By treating URLs as meaningful semantic units rather than random strings, CPT establishes a robust prior for the subsequent supervised mapping and authority alignment stages.

[2]For instance, we use "`plus.gov.kr`" instead of "`https://plus.gov.kr/portal/ntcmttr`".

[3]See Appendix G for further discussion.

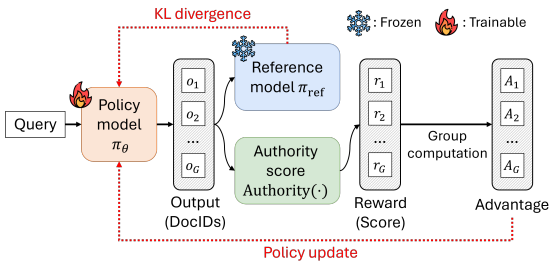

Figure 3: Illustration of the GRPO stage.

3.2.2 Supervised Fine-Tuning (SFT)

The supervised fine-tuning stage optimizes the model for generate relevant DocIDs given a query. This step transforms the latent query-URL associations from CPT into a robust ranking capability. Formally, for a query-document pair (q, d) from the dataset $\mathcal{D}$, we minimize the negative log-likelihood of the ground truth DocID sequence:

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{(q,d)\sim\mathcal{D}}\big[\log P_\theta(d \mid q)\big], \qquad (3)$$

where $P_\theta(d \mid q) = \prod_{t=1}^{L} p_\theta(y_t \mid q, y_{<t})$ denotes the probability of generating the DocID sequence $y = (y_1, \ldots, y_L)$[4].

To construct scalable training data, we utilize real-world search click logs from a major commercial search engine. These logs provide continuously updated signals of user intent and emerging trends, yet inherently suffer from noise due to position bias and accidental clicks. To mitigate this, we employ a hybrid filtering strategy: (i) frequency-based pruning to remove unstable long-tail queries and (ii) relevance verification using an auxiliary ranker to discard semantically mismatched pairs. This approach effectively balances data scale with quality, establishing a reliable foundation for the subsequent authority alignment stage.

3.2.3 Authority-Aware Ranking with Group Relative Policy Optimization (GRPO)

The final stage employs preference optimization to align the model with authority signals. Since SFT treats all valid documents equally, it fails to capture the inherent relativity of trustworthiness. While simple alternatives like Weighted Cross-Entropy exist, they remain pointwise and lack the exploration mechanisms necessary for effective ranking[5]. To overcome these limitations, we adopt Group Relative Policy Optimization (GRPO) (Shao et al., 2024), enabling the model to explicitly prioritize authoritative sources within a group.

[4]Refer Appendix D.3 for the input prompt.

[5]We further discuss the limitations in Appendix D.6.

As illustrated in Figure 3, we sample a group of G candidate DocIDs $O = \{o_1, \ldots, o_G\}$ for a given query q using the current policy $\pi_{\theta_{\text{old}}}$. Each output o_i receives a scalar reward $r_i = \texttt{Authority}(d_i)$. We then compute the advantage $A_i = \frac{r_i - \text{mean}(\mathbf{r})}{\text{std}(\mathbf{r})}$ by normalizing rewards within the group. The policy is optimized via the following objective:

$$\begin{aligned}\mathcal{L}_{\text{GRPO}} = \mathbb{E}\Big[q \sim P(Q),\ \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(O|q)\Big] \\ \frac{1}{G}\sum_{i=1}^{G}\Big[\min(\rho_i A_i,\ \text{clip}(\rho_i, 1-\epsilon, 1+\epsilon)A_i) \\ - \beta\, D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}})\Big],\end{aligned} \qquad (4)$$

where $\rho_i = \frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}$ is the likelihood ratio. π_θ and π_{ref} are the policy and reference model initialized with the SFT model. The hyperparameters β and ϵ regulate the KL penalty and the clipping range. This stage enables the model to distinguish authority levels beyond simple relevance by effectively balancing exploration and exploitation. For training stability, we utilize a strictly filtered dataset of high-frequency queries using pre-computed rewards.

3.3 Hybrid Ensemble Pipeline for Deployment

For production, we employ a Hybrid Ensemble Pipeline that synergizes existing retrievers with our model. This balances high recall and high precision in real-world scenarios. While existing retrievers (Kwon et al., 2025) provides relevance scores $S_{\text{rel}}(d \mid q)$ for broad coverage, AuthGR generates a concise set of DocIDs $\mathcal{D}_{\text{auth}}(q)$ aligned with authority. For $d \in \mathcal{D}_{\text{auth}}(q)$, we derive a normalized score via linear decay:

$$S_{\text{auth}}(d \mid q) = \frac{N - \texttt{rank}(d) + 1}{N}, \qquad (5)$$

where N is the number of generated DocIDs. This formulation assigns higher weights to top-ranked candidates, effectively boosting authoritative documents in the final results.

The final score is computed as follows:

$$S_{\text{final}}(d \mid q) = S_{\text{rel}}(d \mid q) + \lambda \cdot S_{\text{auth}}(d \mid q) \cdot \mathbb{I}\left[d \in \mathcal{D}_{\text{auth}}(q)\right] \qquad (6)$$

where $\mathbb{I}[\cdot]$ is the indicator function and λ controls the strength of authority signal. This formulation preserves the recall of existing rankers while injecting the knowledge of the AuthGR, yielding results that are both semantically relevant and trustworthy.

Model	Size	P@3	R@5	R@10
In-context Learning (ICL)				
Gemma 3	27B	0.1255	0.1732	0.2285
EXAONE 3.5	32B	0.0825	0.1253	0.1612
K-EXAONE	236B	0.1366	0.1918	0.2656
Qwen3	32B	0.0821	0.1176	0.1570
LLaMA 3.1	405B	0.1413	0.1974	0.2590
LLaMA 4 Scout	109B	0.1066	0.1555	0.2128
LLaMA 4 Maverick	400B	0.1274	0.1841	0.2483
DeepSeek-R1	671B	0.0891	0.1328	0.1718
DeepSeek-V3	671B	0.1359	0.1932	0.2626
DeepSeek-V3.2	685B	0.1398	0.2027	0.2729
GPT-4o	-	0.1700	0.2348	0.3170
Supervised Fine-tuning (SFT)				
HyperCLOVAX	0.5B	0.3470	0.4933	0.6634
HyperCLOVAX	1.5B	0.3573	0.5058	0.6708
LLaMA 3.2	1B	0.3479	0.4948	0.6679
LLaMA 3.2	3B	0.3602	0.5108	0.6892
T5Gemma 2	0.5B	0.3280	0.4646	0.6151
T5Gemma 2	2B	0.3399	0.4805	0.6384
Qwen3	1.7B	0.3433	0.4853	0.6582
Qwen3	4B	0.3581	0.5053	0.6843
HyperCLOVAX	14B	0.3854	0.5508	0.7289
Ours				
AuthGR (SFT)	3B	0.3555	0.5058	0.6899
AuthGR (CPT+SFT)	3B	0.3725	0.5293	0.7031
AuthGR (Full)	3B	**0.3856**	**0.5464**	**0.7175**

Table 1: Offline evaluation results. Bold denotes the best performance of our method. 'AuthGR (Full)' denotes the final model incorporating CPT, SFT, and GRPO.

	Production	Hybrid Ensemble
Label score	3.06	**3.41**

Table 2: Human evaluation results in a blind side-by-side test, comparing (i) the production system and (ii) its integration with AuthGR via Hybrid Ensemble.

Metrics	Control	Treatment
Pages with clicks	+0.08%	+21.36%
Total document clicks	+0.22%	+22.07%
Top 1 document CTR	+0.87%	+22.83%
Top 3 document CTR	+0.81%	+22.68%
Top 5 document CTR	+0.81%	+22.76%

Table 3: Online A/B test results showing relative gain in user engagement. 'Control' group uses the production system. 'Treatment' group receives results enhanced by AuthGR through the Hybrid Ensemble.

4 Experimental Setup

Datasets. We constructed three industrial-scale datasets from proprietary logs of a large-scale commercial web search engine in Korea. For CPT, we utilized 9.85M query–document pairs formatted as `[Query; URL; Title; Body]` from the web crawl of the search engine. For SFT, we curated 3.95M pairs from high-stakes domains such as health or finance with weekly query counts (QC) > 50, filtering 63% of noisy interactions. For GRPO, we selected 13.81K queries (QC > 200) and utilized pre-computed authority scores across 3.75M host URLs as rewards. Please refer to Appendix D.1.

Baselines. We adopt two categories of baselines considering Korean as the target language. (i) In-context learning: Large-scale foundation models including **Gemma 3** (Kamath et al., 2025), **EXAONE 3.5** (LG AI Research, 2024), **K-EXAONE** (LG AI Research, 2025), **Qwen3** (Yang et al., 2025), **LLaMA 3.1** (Dubey et al., 2024), **LLaMA 4** (Meta AI, 2025), **DeepSeek-R1/V3** (Guo et al., 2025; Liu et al., 2024), and **GPT-4o** (Hurst et al., 2024). (ii) Supervised fine-tuning: Compact models including **HyperCLOVAX** (Yoo et al., 2024; NAVER Cloud, 2025), **T5Gemma 2** (Zhang et al., 2025a), **LLaMA 3.2** (Dubey et al., 2024), and **Qwen3** (Yang et al., 2025). See Appendix D.2 for details.

Implementation Details. We initialized a 3B decoder-only transformer from the language component of HyperCLOVAX-SEED-Vision-Instruct (Yoo et al., 2024). For GRPO, a rollout group size was $G = 256$ and KL coefficient was $\beta = 0.2$. During inference, we employed beam search with a size of 10, and the coefficient of the ensemble was set to $\lambda = 0.6$ based on validation tuning. Further details are in Appendix D.4.

Evaluation Protocols. We validate AuthGR via three evaluation protocols. (i) **Offline Evaluation**: We utilize 3,000 expert queries with human-labeled ground truth to measure Precision@3 and Recall@{5,10}. (ii) **Human Evaluation**: A blind side-by-side comparison on 500 queries assesses the quality of AuthGR-enhanced pipeline over production system. (iii) **Online A/B Test**: We analyze user engagement metrics across millions of interactions on the large-scale web search platform. We compare the production system against AuthGR-enhanced pipeline, monitoring key metrics including click-through rate and total document clicks in mid-2025. See Appendix D.5 for details.

5 Experimental Results

5.1 Main Results

Offline Evaluation. Table 1 demonstrates the offline performance, where AuthGR achieves supe-

Model	SFT	CPT	GRPO	P@3	R@5	R@10
AuthGR (0.5B)	✓			0.3470	0.4933	0.6634
	✓		✓	**0.3515**	**0.4989**	**0.6651**
AuthGR (3B)	✓			0.3555	0.5058	0.6899
	✓		✓	0.3660	0.5216	0.6986
	✓	✓		0.3725	0.5293	0.7031
	✓	✓	✓	**0.3856**	**0.5464**	**0.7175**

Table 4: Ablation study of the training stages.

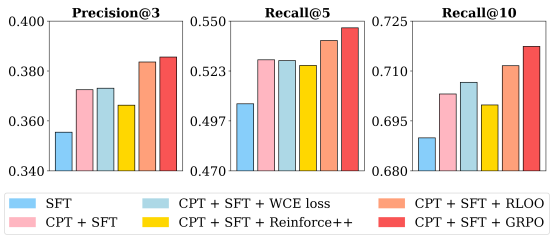

Figure 4: Performance of ranking optimization methods. 'CPT+SFT+GRPO' represents our AuthGR.

rior accuracy with the highest P@3 among all baselines. Remarkably, our 3B model performs on par with the 14B model despite being 4.7× smaller, highlighting the parameter efficiency of our framework. Moreover, the effectiveness of three-stage pipeline is validated, where CPT ensures domain adaptation and GRPO explicitly optimizes authority. Furthermore, the limited performance of ICL baselines underscores the necessity of task-specific fine-tuning for effective generative retrieval.

Human Evaluation. Table 2 presents the results of a blind side-by-side comparison between the commercial search engine's production system and AuthGR-enhanced ensemble pipeline. Each result was evaluated on a 1–5 point scale across relevance and authority. Remarkably, AuthGR achieves 11.4% gain in average score, confirming that explicitly optimizing for authority yields results that users perceive as more relevant and trustworthy.

Online A/B Test. As reported in Table 3, AuthGR delivers substantial improvements in an online A/B test. Compared to the production baseline, user engagement metrics surged, where 'Pages with clicks' increased by 21.36% and 'Total document clicks' by 22.07%. Notably, the 'Top-1 document CTR' is enhanced by 22.83%, indicating that authority-aware ranking significantly enhances the quality of the top-ranked results.

5.2 In-depth Analysis

Impact of Training Stages. Table 4 illustrates an ablation study for the training stages of AuthGR,

Training stage	Authority scores		# docs by authority	
	Mean(↑)	Median(↑)	Low(↓)	High(↑)
CPT+SFT	87.2	90.0	118	2,754
+ GRPO (Binary)	88.0	90.0	115	2,804
+ GRPO (Linear)	**90.4**	**95.0**	**106**	**2,877**

Table 5: Authority score distribution in generated documents. Mean and Median are computed on a 0-100 scale. 'Low' and 'High' indicate the number of documents scoring 0-60 and 90-100, respectively.

highlighting four key insights. (i) The full pipeline achieves the best performance, yielding a total gain of 8.5% in P@3 over the SFT baseline for the 3B model. (ii) The CPT stage delivers a dominant improvement of 4.8%, creating a strong domain-specific foundation for the subsequent GRPO stage, which contributes an additional 3.0%. (iii) The benefit of GRPO scales with model capacity. The 3B model exhibits a larger gain of 3.1% compared to 1.3% for the 0.5B model, suggesting that larger models better internalize the concept of authority. (iv) A clear synergy exists between stages, GRPO achieves a higher gain of 3.5% when applied after CPT, compared to 3.0% for the SFT-only baseline.

Comparison of Ranking Optimization. Figure 4 compares the effectiveness of GRPO against representative pointwise methods such as Weighted Cross-Entropy (WCE) and Reinforce++ (Hu et al., 2025), as well as groupwise alternatives like RLOO (Ahmadian et al., 2024). The key findings are as follows: (i) GRPO emerges as the most effective approach, yielding the highest gain of 3.5% in P@3 over the CPT + SFT baseline. (ii) Groupwise methods significantly outperform pointwise methods. GRPO and RLOO deliver clear gains of 3.5% and 3.0% in P@3, respectively. In contrast, WCE provides only a marginal 0.2% gain, and Reinforce++ degrades performance by 1.7%. This confirms that modeling the relative order of candidates is better suited for the ranking tasks.

Impact on Authority Scores. Table 5 shows that GRPO successfully reshapes the authority distribution of generated documents. The mean score increases from 87.2 to 90.4, with the median reaching 95.0. This distributional shift is driven by a 10.2% reduction in low-authority generations (0–60) and a 4.5% increase in high-authority generations (90–100). This confirms that the model systematically learns to prioritize trustworthy sources.

6 Conclusion

We propose AuthGR, a generative retrieval framework that explicitly integrates document authority via multimodal scoring and progressive training. By employing a hybrid ensemble strategy, we successfully deployed this 3B model to a commercial search engine, where it matched the performance of 14B baselines and significantly improved real-world user engagement and satisfaction. Our work underscores the critical role of authority in generative retrieval, paving the way for the development of trustworthy real-world search applications.

7 Limitations

While AuthGR demonstrates significant improvements in authority-aware generative retrieval, we acknowledge several limitations that provide directions for future work. (i) We primarily use multimodal authority scores as reward signals for policy optimization. Although these scores prove the effectiveness in capturing site trustworthiness, incorporating more diverse and granular reward signals could further refine the model's understanding. For instance, integrating implicit user feedback, such as dwell time or scroll depth, could further enhance user satisfaction. (ii) We observed that authority reasoning capabilities scale with model size, as the 3B model outperformed the 0.5B model. However, we did not scale beyond 3B due to the computational overhead. By leveraging advanced engineering techniques for inference efficiency, such as model quantization or speculative decoding, future research could explore the scalability of AuthGR with much larger foundation models to further unlock complex authority reasoning potential.

Ethics Statement

This work fully complies with the ACL Ethics Policy. We declare that there are no ethical issues in this paper. The scientific artifacts we have utilized are publicly available for research under permissive licenses, and the utilization of these tools is consistent with their intended applications.

Acknowledgments

This work was supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. IITP-RS-2022-II220680, NRF-RS-2025-00564083, IITP-RS-2019-II190421, IITP-2026-RS-2024-00437633, each contributing 25% to this research).

References

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. 2024. Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms. *CoRR*, abs/2402.14740.

Anirudhan Badrinath, Prabhat Agarwal, Laksh Bhasin, Jaewon Yang, Jiajing Xu, and Charles Rosenberg. 2025. Pinrec: Outcome-conditioned, multi-token generative retrieval for industry-scale recommendation systems. *CoRR*, abs/2504.10507.

Michele Bevilacqua, Giuseppe Ottaviano, Patrick S. H. Lewis, Scott Yih, Sebastian Riedel, and Fabio Petroni. 2022. Autoregressive search engines: Generating substrings as document identifiers. In *NeurIPS*.

Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Comput. Networks*, 30(1-7):107–117.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive entity retrieval. In *ICLR*.

Jiangui Chen, Ruqing Zhang, Jiafeng Guo, Yiqun Liu, Yixing Fan, and Xueqi Cheng. 2022. Corpusbrain: Pre-train a generative retrieval model for knowledge-intensive language tasks. In *CIKM*, pages 191–200.

Wei Chen, Yixin Ji, Zeyuan Chen, Jia Xu, and Zhongyi Liu. 2024. LLMGR: large language model-based generative retrieval in alipay search. In *SIGIR*, pages 2847–2851.

Charles L. A. Clarke, Maria Maistro, and Mark D. Smucker. 2021. Overview of the TREC 2021 health misinformation track. In *TREC*, volume 500-335 of *NIST Special Publication*. National Institute of Standards and Technology (NIST).

Charles L. A. Clarke, Saira Rizvi, Mark D. Smucker, Maria Maistro, and Guido Zuccon. 2020. Overview of the TREC 2020 health misinformation track. In *TREC*, volume 1266 of *NIST Special Publication*. National Institute of Standards and Technology (NIST).

Jiaxin Deng, Shiyao Wang, Kuo Cai, Lejian Ren, Qigen Hu, Weifeng Ding, Qiang Luo, and Guorui Zhou. 2025. Onerec: Unifying retrieve and rank with generative recommender and iterative preference alignment. *CoRR*, abs/2502.18965.

Xin Luna Dong, Evgeniy Gabrilovich, Kevin Murphy, Van Dang, Wilko Horn, Camillo Lugaresi, Shaohua Sun, and Wei Zhang. 2015. Knowledge-based trust:

Estimating the trustworthiness of web sources. *Proc. VLDB Endow.*, 8(9):938–949.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2022. Recommendation as language processing (RLP): A unified pretrain, personalized prompt & predict paradigm (P5). In *RecSys*, pages 299–315.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.

Zoltán Gyöngyi, Hector Garcia-Molina, and Jan O. Pedersen. 2004. Combating web spam with trustrank. In *VLDB*, pages 576–587.

Ruining He, Lukasz Heldt, Lichan Hong, Raghunandan H. Keshavan, Shifan Mao, Nikhil Mehta, Zhengyang Su, Alicia Tsai, Yueqi Wang, Shao-Chuan Wang, Xinyang Yi, Lexi Baugher, Baykal Cakici, Ed H. Chi, Cristos Goodrow, Ningren Han, He Ma, Rómer Rosales, Abby Van Soest, and 4 others. 2025. PLUM: adapting pre-trained language models for industrial-scale generative recommendations. *CoRR*, abs/2510.07784.

Jian Hu, Jason Klein Liu, and Haotian Xu Wei Shen. 2025. Reinforce++: A simple and efficient approach for aligning large language models. *CoRR*, abs/2501.03262.

Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *CoRR*, abs/2410.21276.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, and 1 others. 2025. Gemma 3 technical report. *CoRR*, abs/2503.19786.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *EMNLP*, pages 6769–6781.

Jon M. Kleinberg. 1999. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632.

Ohjoon Kwon, Changsu Lee, Jihye Back, Lim Sun Suk, Inho Kang, and Donghyeon Jeon. 2025. QUPID: Quantified understanding for enhanced performance, insights, and decisions in Korean search engines. In *ACL (Industry Track)*, pages 541–552.

Geon Lee, Bhuvesh Kumar, Mingxuan Ju, Tong Zhao, Kijung Shin, Neil Shah, and Liam Collins. 2026. Sequential data augmentation for generative recommendation. In *WSDM*, pages 303–312. ACM.

Sunkyung Lee, Minjin Choi, Eunseong Choi, Hyeyoung Kim, and Jongwuk Lee. 2025a. GRAM: generative recommendation via semantic-aware multi-granular late fusion. In *ACL*, pages 33294–33312.

Sunkyung Lee, Minjin Choi, and Jongwuk Lee. 2023. GLEN: generative retrieval via lexical index learning. In *EMNLP*, pages 7693–7704.

Sunkyung Lee, Seongmin Park, Jonghyo Kim, Mincheol Yoon, and Jongwuk Lee. 2025b. Enhancing time awareness in generative recommendation. In *EMNLP (Findings)*, pages 23917–23933.

LG AI Research. 2024. Exaone 3.5: Series of large language models for real-world use cases. *CoRR*, abs/2412.04862.

LG AI Research. 2025. K-exaone technical report. *CoRR*, abs/2601.01739.

Yongqi Li, Xinyu Lin, Wenjie Wang, Fuli Feng, Liang Pang, Wenjie Li, Liqiang Nie, Xiangnan He, and Tat-Seng Chua. 2024a. A survey of generative search and recommendation in the era of large language models. *CoRR*.

Yongqi Li, Nan Yang, Liang Wang, Furu Wei, and Wenjie Li. 2024b. Learning to rank in generative retrieval. In *AAAI*, pages 8716–8723.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv: 2412.19437*.

Kidist Amde Mekonnen, Yubao Tang, and Maarten de Rijke. 2025. Lightweight and direct document relevance optimization for generative information retrieval. In *SIGIR*, pages 1327–1338.

Meta AI. 2025. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation. https://ai.ai.meta.com/blog/llama-4-multimodal-intelligence/.

Donald Metzler, Yi Tay, Dara Bahri, and Marc Najork. 2021. Rethinking search: making domain experts out of dilettantes. *SIGIR Forum*, 55(1):13:1–13:27.

NAVER Cloud. 2025. Hyperclova x think technical report. *arXiv preprint arXiv: 2506.22403*.

Ming Pang, Chunyuan Yuan, Xiaoyu He, Zheng Fang, Donghao Xie, Fanyi Qu, Xue Jiang, Changping Peng, Zhangang Lin, Zheng Luo, and Jingping Shao. 2025. Generative retrieval and alignment model: A new paradigm for e-commerce retrieval. In *WWW Companion*, pages 413–421.

Gustavo Penha, Ali Vardasbi, Enrico Palumbo, Marco De Nadai, and Hugues Bouchard. 2024. Bridging search and recommendation in generative retrieval: Does one task help the other? In *RecSys*, pages 340–349.

Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Hulikal Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Q. Tran, Jonah Samost, and 1 others. 2023. Recommender systems with generative retrieval. In *NeurIPS*, pages 10299–10315.

Ruiyang Ren, Wayne Xin Zhao, Jing Liu, Hua Wu, Ji-Rong Wen, and Haifeng Wang. 2023. TOME: A two-stage approach for model-based retrieval. In *ACL*, pages 6102–6114.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300.

Yedan Shen, Kaixin Wu, Yuechen Ding, Jingyuan Wen, Hong Liu, Mingjie Zhong, Zhouhan Lin, Jia Xu, and Linjian Mo. 2025. Alleviating llm-based generative retrieval hallucination in alipay search. In *SIGIR*, pages 4294–4298.

EuiYul Song, Sangryul Kim, Haeju Lee, Joonkee Kim, and James Thorne. 2024. Re3val: Reinforced and reranked generative retrieval. In *EACL (Findings)*, pages 393–409.

Weiwei Sun, Lingyong Yan, Zheng Chen, Shuaiqiang Wang, Haichao Zhu, Pengjie Ren, Zhumin Chen, Dawei Yin, Maarten de Rijke, and Zhaochun Ren. 2023. Learning to tokenize for generative retrieval. In *NeurIPS*.

Yi Tay, Vinh Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Prakash Gupta, Tal Schuster, William W. Cohen, and Donald Metzler. 2022. Transformer memory as a differentiable search index. In *NeurIPS*, pages 21831–21843.

Yujing Wang, Yingyan Hou, Haonan Wang, Ziming Miao, Shibin Wu, Qi Chen, Yuqing Xia, Chengmin Chi, Guoshuai Zhao, Zheng Liu, and 1 others. 2022. A neural corpus indexer for document retrieval. In *NeurIPS*.

Haoyang Wen, Jiang Guo, Yi Zhang, Jiarong Jiang, and Zhiguo Wang. 2025. On synthetic data strategies for domain-specific generative retrieval. In *ACL*, pages 7961–7976.

Yanjing Wu, Yinfu Feng, Jian Wang, Wenji Zhou, Yunan Ye, and Rong Xiao. 2024. Hi-gen: Generative retrieval for large-scale personalized e-commerce search. In *ICDM*, pages 893–898.

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. *CoRR*, abs/2505.09388.

Dezhi Ye, Junwei Hu, Jiabin Fan, Bowen Tian, Jie Liu, Haijin Liang, and Jin Ma. 2025. Best practices for distilling large language models into BERT for web search ranking. In *COLING (Industry Track)*, pages 128–135.

Kang Min Yoo, Jaegeun Han, Sookyo In, Heewon Jeon, Jisu Jeong, Jaewook Kang, Hyunwook Kim, Kyung-Min Kim, Munhyong Kim, Sungju Kim, and 1 others. 2024. Hyperclova X technical report. *CoRR*, abs/2404.01954.

Hansi Zeng, Chen Luo, Bowen Jin, Sheikh Muhammad Sarwar, Tianxin Wei, and Hamed Zamani. 2024a. Scalable and effective generative information retrieval. In *WWW*, pages 1441–1452.

Hansi Zeng, Chen Luo, and Hamed Zamani. 2024b. Planning ahead in generative retrieval: Guiding autoregressive generation through simultaneous decoding. In *SIGIR*, pages 469–480.

Biao Zhang, Paul Suganthan, Gaël Liu, Ilya Philippov, Sahil Dua, Ben Hora, Kat Black, Gus Martins, Omar Sanseviero, Shreya Pathak, Cassidy Hardin, Francesco Visin, Jiageng Zhang, Kathleen Kenealy, Qin Yin, Olivier Lacombe, Armand Joulin, Tris Warkentin, and Adam Roberts. 2025a. T5gemma 2: Seeing, reading, and understanding longer.

Dake Zhang, Amir Vakili Tahami, Mustafa Abualsaud, and Mark D. Smucker. 2022. Learning trustworthy web sources to derive correct answers and reduce health misinformation in search. In *SIGIR*, pages 2099–2104. ACM.

Fuwei Zhang, Xiaoyu Liu, Xinyu Jia, Yingfei Zhang, Zenghua Xia, Fei Jiang, Fuzhen Zhuang, Wei Lin, and Zhao Zhang. 2025b. HierGR: Hierarchical semantic representation enhancement for generative retrieval in food delivery search. In *ACL (Industry Track)*, pages 444–455.

Hailin Zhang, Yujing Wang, Qi Chen, Ruiheng Chang, Ting Zhang, Ziming Miao, Yingyan Hou, Yang Ding, Xupeng Miao, Haonan Wang, and 1 others. 2023. Model-enhanced vector index. In *NeurIPS*.

Bowen Zheng, Yupeng Hou, Hongyu Lu, Yu Chen, Wayne Xin Zhao, Ming Chen, and Ji-Rong Wen. 2024. Adapting large language models by integrating collaborative semantics for recommendation. In *ICDE*, pages 1435–1448.

Yujia Zhou, Zhicheng Dou, and Ji-Rong Wen. 2023. Enhancing generative retrieval with reinforcement learning from relevance feedback. In *EMNLP*, pages 12481–12490.

Shengyao Zhuang, Houxing Ren, Linjun Shou, Jian Pei, Ming Gong, Guido Zuccon, and Daxin Jiang. 2023. Bridging the gap between indexing and retrieval for differentiable search index with query generation. In *Gen-IR@SIGIR*.

Metric	Value
Point-biserial correlation (r)	0.495 ($p < 0.001$)
ROC-AUC	0.915
Matthews Correlation (MCC)	0.526

Table 6: Quantitative alignment between VLM-generated authority scores and human expertise labels.

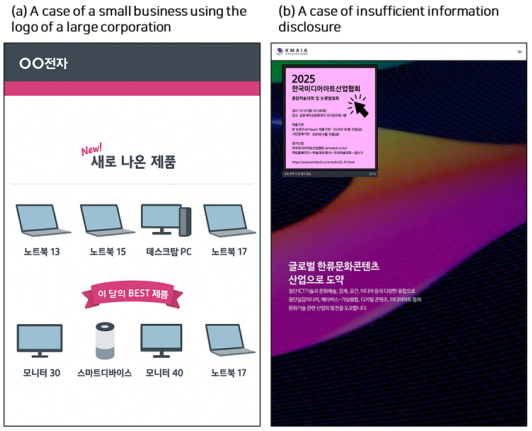

Figure 5: Examples where the Text+Image input yields more appropriate results than the image-only input. (a) A case of a small business using the logo of a large corporation. The Korean text reveals a generic vendor name, correcting visual bias. (b) A case of insufficient information disclosure. The sparse visual layout is clarified by text confirming its official status.

A Additional Details on Multimodal Authority Scoring

A.1 VLM Prompt and Full Output Schema

As shown in Figure 6, the VLM prompt is designed with a systematic structure to ensure reliability and consistency. It comprises six key components: *Context* provides the background information for explaining the situation; *Role* defines the expert persona; *Task* specifies the concrete action the model is expected to perform; *Output format* prescribes the structure of the response; *Instruction* offers detailed guidance; and *Constraints* outline the strict conditions. This structured design enables consistent authority evaluation across diverse websites. The VLM processes multimodal inputs including textual data and screenshots of websites. Multiple sites are grouped and processed through a batch API request. The results are returned as a strictly formatted JSON array, and **this predefined structure eliminates complex parsing overhead**, thereby significantly enhancing data processing efficiency. Note that unevaluable scores -1 are mapped to 0.

A.2 Reliability of VLM Scores

To validate the VLM scores as a reliable proxy for human judgment, we analyzed their correlation with binary human expertise labels (true/false). As summarized in Table 6, **all metrics consistently demonstrate strong and statistically significant correlation between VLM scores and human assessments**. The Point-biserial correlation and Matthews Correlation Coefficient confirm robust positive alignment between the VLM scores and the expert-labeled expertise. Additionally, the ROC-AUC demonstrates the discriminative power of the model in distinguishing authoritative sites. These results verify that the VLM effectively functions as a high-quality automated assessor that closely reflects human evaluators' intuition.

A.3 Impact of Modalities on Scoring

To assess the contribution of different modalities, we evaluated the VLM on 40,000 websites using three settings: text-only (URL, title, body, Wikipedia), image-only (screenshots), and text+image. Each website was independently annotated by human annotators, who labeled whether the site exhibited expertise (true or false). We then considered a site to be authoritative if the VLM produced a high authority score, and measured agreement with the human judgments. The VLM achieved approximately 81% accuracy with text-only inputs and 92% with image-only inputs. Notably, **the multimodal setting reached 97% accuracy, demonstrating that multimodal inputs substantially improve authority detection**.

A.4 Efficiency and Scalability of Scoring

Cost Efficiency. We adopt three optimization strategies. First, we employ **batch requests**, where multiple queries are grouped and sent to the API simultaneously. This reduces the number of calls, minimizes network overhead, and benefits from lower per-request costs compared to real-time queries. Second, **multi-sample requests** are utilized by consolidating multiple sites into a single prompt, thereby reducing prompt token usage and improving overall response efficiency. Finally, we enforce **structured JSON output formats** instead of free-form natural language. This eliminates unnecessary descriptive text, lowers post-processing overhead, and enables seamless integration into automated pipelines. Together, these strategies significantly enhance the scalability and economic viability of VLM-based authority evaluation.

Scalability and Maintenance. For production-scale deployment, we conduct a full database re-

VLM Prompt

Context:
The task is to determine which web documents are valuable enough to be included in search engine results. The goal is to provide users with high-quality web documents that demonstrate reliability and expertise. Such documents typically originate from authoritative institutions or organizations, or consist of content created with significant effort and originality.

Role:
You are a site authority evaluation expert.

Task:
1. **Site type:** Classify the website by its structure/function.
2. **Topic:** Categorize the main subject or information the site covers.
3. **Score:** Assign an integer score between -1 and 100.
4. **Evidence:** Provide 1–2 sentences in explaining the reason for the score.

Output Format:
All output must be a JSON Array as shown below:

```
[{"url": "", "site_type": "", "topic": "",
"score": 0, "evidence": ""}, ...]
```

- Do not include any text outside of the JSON Array.
- The number of output items must match the number of input sites.

Site Type Instruction:

- **Commerce**: A site where users can directly purchase products, with visible product listings showing prices.
- **SNS**: An official website where social interaction and content sharing are the main features.
- **Community**: An official site for discussions and information exchange with forums, comments, or membership.
- **News/Media**: An official news or broadcasting site run by governments, newspapers, or broadcasters.
- **Blog**: Blog platforms.
- **General**: Any site not fitting the above categories, or where the type is ambiguous.

Topic Instruction:
- **Health**: Medical info, diseases, healthcare, medicine, devices, nutrition, rehab, diet.
- **Education**: Schools, learning, careers, exams, certificates, training, textbooks, courses.
- **IT**: Tech, telecom, electronics, digital devices, industry news, trends.
- ...

Score Instruction:
- 80–100: High-quality sites
- 50–79: Mid-quality sites
- 10–49: Low-quality sites
- 0: Spam, gambling, illegal, adult content, `warning.or.kr`, or policy-violating content
- -1: No information, inaccessible site, or unable to evaluate

Constraints:
- Strictly adhere to the JSON output format.
- The `evidence` field must always be written in Korean.
- Prioritize the screenshot image over other signals when evaluating.
- Do not classify `site_type` solely by URL.

Figure 6: Prompts used for Multimodal Authority Scoring.

Model	Size	Latency (ms)	Throughput
HyperCLOVAX (SFT)	14B	2,881	1.30
AuthGR (Ours)	**3B**	**1,225**	**3.26**

Table 7: Comparison of inference efficiency. Latency denotes the average response time, and throughput indicates the number of requests processed per second.

scoring quarterly, requiring approximately 30M tokens per 100K sites. While we used a high-capacity VLM internally, we confirmed that open-source VLMs effectively reproduce our results. This validates generalizability of the rubric beyond proprietary models. To ensure freshness, we monitor content changes during crawling and perform incremental weekly updates. We identify target sites using a Gradient Boosted Regression Tree model trained on VLM labels; sites whose predicted scores deviate significantly from current values are flagged for re-scoring. Combined with manageable retraining cost, this two-tier update strategy ensures sustainable real-world deployment.

A.5 Qualitative Examples of Scoring

Relying solely on visual cues can lead to misjudgments when screenshots are misleading or insufficient. Figure 5(a) illustrates a small vendor website selling products from a large corporation may display the corporation's logo on its site. The image-only model may mistakenly identify it as the corporation's official website. The VLM produced the incorrect evidence, "*the official OO Electronics website — a corporate, authoritative site*," and consequently judged the site to exhibit high expertise. In contrast, in the text+image condition, the VLM correctly recognizes it as "*the homepage of an individual or small-scale seller of OO Electronics products.*", assigning appropriate low expertise.

Similarly, Figure 5(b) shows a site with sparse visual content. The image-only model may dismiss it as assessed low authority with the reasoning "*the information site of the Art and Science Convergence Project Group, but the content is limited to a brief notice page.*" In contrast, by leveraging both texts and images, the model correctly identifies it as "*official site of the Korea Media Art Industry Association, providing industry information and networking opportunities*," yielding a high score.

B Efficiency Analysis

B.1 Training Efficiency

The three-stage pipeline is designed for periodic retraining at a manageable cost. On 8×A100 GPUs, the full process completes in approximately 83 hours: (i) CPT: 57h (9.85M samples), (ii) SFT: 20h (3.95M samples), and (iii) GRPO: 6h (13.8K samples). The minimal overhead of the GRPO stage is particularly advantageous for frequent authority-alignment updates.

B.2 Inference Efficiency

Table 7 compares the inference latency of AuthGR (3B) against the HyperCLOVAX-14B (SFT) baseline. On an NVIDIA A100 environment, AuthGR 3B achieves a 2.35× reduction in latency and 2.51× higher throughput while maintaining comparable ranking performance. This demonstrates that AuthGR delivers high-quality retrieval at a fraction of the computational cost of larger models.

C Extended Related Work on GenIR

DocID Design and Variations. Beyond basic numeric IDs (Wang et al., 2022; Zhuang et al., 2023) and URLs (Chen et al., 2022), various DocIDs have been explored to bridge the gap between identifiers and document semantics. These include N-grams for substring-based retrieval (Cao et al., 2021; Bevilacqua et al., 2022), or hierarchical codebooks (Zhang et al., 2023; Zeng et al., 2024a,b).

Optimization and Learning Paradigms. The evolution of GenIR has shifted from simple tasks to complex ranking optimization. Apart from the ranking losses (Sun et al., 2023), distillation have been employed to transfer ranking capabilities from large-scale teacher models to generative retrievers (Zeng et al., 2024a). Furthermore, the generative paradigm has been successfully adapted to recommendation systems, where models generate item IDs for personalized discovery (Rajput et al., 2023; Geng et al., 2022; Zheng et al., 2024; Lee et al., 2025a,b; Deng et al., 2025; Lee et al., 2026).

Applications and Feedback Signals. Recent research incorporates diverse feedback signals, such as user preference rankings, to fine-tune generative models through reinforcement learning (Song et al., 2024). While these methods have been rapidly expanded into real-world industrial applications (Penha et al., 2024; He et al., 2025), such as financial services (Chen et al., 2024), or visual

discovery (Badrinath et al., 2025), they remain limited to relevance-centric optimization. Our work extends this trajectory by incorporating authority signals into the GenIR pipeline.

D Additional Experimental Setup

D.1 Datasets

- **CPT Phase**: To adapt the model for generative retrieval, we designed diverse sequence formats. In addition to standard search log sequences such as `[Query; Title]` and `[Query; Snippet]`, we added `[Query; URL; Title; Body]`. This enables the model to learn the structural relationship between URLs and content reliability while modeling query-document relevance.
- **SFT Phase**: We applied a two-stage refinement: (i) rule-based filtering to exclude low-authority platforms (*e.g.*, personal blogs, inactive wikis) and (ii) model-based cleaning using a retriever (Kwon et al., 2025). Although this removed 63% of raw click logs, our preliminary study showed that training on the curated subset improved P@3 by +16.62% and accelerated convergence by 40% compared to using the full, noisy dataset.
- **GRPO Phase**: We focused on high-frequency queries across eight high-stakes domains: *health, education, information technology, finance, parenting, society, animals,* and *recruitment*. From weekly logs, we sampled queries with QC > 50 and further filtered for unambiguous user intents using an LLM (`gemma-3-27b-it`), resulting in 13,814 high-quality samples to ensure clear reward signals during policy optimization.

D.2 Baselines

All baselines were sourced from the Hugging Face Hub. The specific versions are listed below.

(i) In-context Learning:

- **Gemma 3**: `google/gemma-3-27b-it`
- **EXAONE 3.5**: `LGAI-EXAONE/EXAONE-3.5-32B-Instruct`
- **K-EXAONE**: `LGAI-EXAONE/K-EXAONE-236B-A23B`
- **Qwen3**: `Qwen/Qwen3-32B`
- **LLaMA 3.1**: `meta-llama/Llama-3.1-405B`
- **LLaMA 4**: `meta-llama/Llama-4-Scout-17B-16E-Instruct`, `meta-llama/Llama-4-Maverick-17B-128E-Instruct`
- **DeepSeek-R1**: `deepseek-ai/DeepSeek-R1`
- **DeepSeek-V3**: `deepseek-ai/DeepSeek-V3`
- **DeepSeek-V3.2**: `deepseek-ai/DeepSeek-V3.2`

(ii) Supervised Fine-tuning:

- **HyperCLOVAX**: `naver-hyperclovax/HyperCLOVAX-SEED-Text-Instruct-0.5B`, `naver-hyperclovax/HyperCLOVAX-SEED-Text-Instruct-1.5B`, `naver-hyperclovax/HyperCLOVAX-SEED-Think-14B`
- **LLaMA 3.2**: `meta-llama/Llama-3.2-1B-Instruct`, `meta-llama/Llama-3.2-3B-Instruct`
- **T5Gemma 2**: `google/t5gemma-2-270m-270m`, `google/t5gemma-2-1b-1b`
- **Qwen3**: `Qwen/Qwen3-1.7B`, `Qwen/Qwen3-4B`

SFT Instruction Prompt

Instruction:
You are a model that generates authoritative websites. Based on the user query, generate the most relevant and authoritative website URL.

Conditions:
1. The website must be a trustworthy domain (e.g., .gov, .edu, official institutions, etc.).
2. Generate a website that contains appropriate information to fulfill the user's request.
3. The output should only include the website.
4. Even if the input is incomplete, analyze the user's intent to generate the most suitable website.
5. If no suitable website is found, return an empty output.

Input:
- Query: `{query}`

Output:
- Site: `[Website URL ID]`

Figure 7: Prompts used for the SFT stage. The `{query}` field is a placeholder for the user query.

D.3 Prompts for SFT Stage

Figure 7 shows the instruction prompt used for the SFT stage. For fair comparison, we utilize the identical prompt for all SFT baselines in Table 1.

D.4 Implementation Details

We implemented the model in PyTorch and trained it on 8 NVIDIA A100 GPUs using DeepSpeed ZeRO-3. We utilized AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.95$, weight decay $= 0.01$, a cosine scheduler and a warmup ratio of 0.03. The maximum input length was 2,048 tokens. For the CPT stage, we used a global batch size of 256 and peak learning rate 5.0×10^{-6}. In the SFT stage, the

model was trained for 3 epochs with a global batch size of 512 and a learning rate of 1×10^{-6}. For the GRPO stage, the model was trained for one epoch. During the rollout, we sampled candidate DocIDs using a temperature of 1.5, a top-p of 0.8, and a top-k of 50. For the WCE baseline (Equation 7), we set $\alpha = 4.0$. During inference, we used beam search following existing works (Tay et al., 2022; Wang et al., 2022) with a beam size of 10. The maximum generation length per DocID was 50 tokens, sufficient to cover host-level URLs.

D.5 Evaluation Protocol

Offline Quantitative Evaluation. For the 3,000 expert queries, we retrieved 30 candidate documents per query and instructed human annotators to identify the top-3 relevant ones. To ensure fair comparison, both ground truth labels and model predictions were evaluated at the host-URL level.

Human Evaluation. In a blind side-by-side comparison, annotators evaluated the top-1 results from the production baseline and our hybrid system. We used a 5-point scale that simultaneously weights semantic relevance (providing sufficient information) and authority (originating from credible sources). For each query, annotators made a direct preference judgment ("*Which result is better overall?*") to compute the win ratio. Detailed annotation guidelines are in Appendix F.

Online A/B Testing. We conducted an A/B test over several days in mid-2025. The control group received results from the existing production system, while the treatment group received results from the same system augmented with AuthGR through the hybrid ensemble pipeline. We monitored three engagement metrics: (i) *pages with clicks* (the number of pages receiving at least one click), (ii) *total document clicks* (the total number of document clicks), and (iii) *top@k document click-through rate (CTR)* for $k \in \{1, 3, 5\}$ to specifically quantify user engagement with high-ranked authoritative documents.

D.6 Details of the Weighted Cross-Entropy Baseline

To incorporate authority signals, we examine a pointwise Weighted Cross-Entropy (WCE), as shown in Figure 4. The core idea is to scale the standard cross-entropy loss for each DocID by its authority score, thereby encouraging the model to assign higher probabilities to more authoritative sources. Formally, given a document d, its authority score is transformed into a weight:

$$w_d = 1 + \alpha \cdot \frac{\texttt{Authority}(d)}{100} \tag{7}$$

where α is a hyperparameter that controls the weight of authority. The objective minimizes the weighted negative log-likelihood over the training data $\mathcal{D}$:

$$\mathcal{L}_{\text{WCE}} = -\mathbb{E}_{(q,d)\sim\mathcal{D}}\big[w_d \log P_\theta(d \mid q)\big] \tag{8}$$

However, WCE has critical limitations: (i) pointwise loss misaligns with the ranking tasks, and (ii) it lacks exploration capability. These limitations motivated our adoption of the group-wise GRPO.

Metric	Gain	p-value	95% CI
P@3	2.31%	0.0277	[0.0012, 0.0158]
R@5	1.65%	0.0483	[-0.0002, 0.0170]

Table 8: Statistical significance between AuthGR and HyperCLOVAX 14B. Gain represents the relative difference in metrics between AuthGR and the 14B model.

Training stages	Gain	p-value	95% CI
SFT → CPT+SFT	+4.38%	<0.0001	[0.0082, 0.0221]
CPT+SFT → Full	+3.00%	<0.0001	[0.0062, 0.0155]
SFT → Full	+7.51%	<0.0001	[0.0189, 0.0334]

Table 9: Statistical significance of the three-stage training. Gain denotes the relative improvement in metric values achieved by each incremental stage on P@3.

E Additional Experimental Results

E.1 Statistical Significance Analysis

To ensure that performance gains are robust, we conducted statistical validation on 3,041 queries using paired t-tests with bootstrap resampling (5,000 iterations). We also report the 95% Confidence Interval (CI) to estimate the range of the true mean performance difference.

Comparison with 14B Baseline. Table 8 shows that AuthGR 3B significantly outperforms the 14B baseline in top-rank metrics including P@3 and R@5 with $p < 0.05$. Despite its smaller size, our model achieves statistically superior precision. This accuracy, combined with lower operational costs, underscores the practical suitability of AuthGR for production-scale deployment.

Effectiveness of Training Pipeline. Table 9 demonstrates that every stage yields statistically significant gains. Notably, the full pipeline achieves a +7.51% relative improvement in P@3 over the

Scaling strategies	P@3	R@5	R@10
Binary	0.3820	0.5375	0.7133
Binning	0.3841	0.5420	0.7151
Sigmoid	0.3833	0.5377	0.7147
Linear (AuthGR)	**0.3856**	**0.5464**	**0.7175**

Table 10: Performance comparison over various scaling strategies for authority scores.

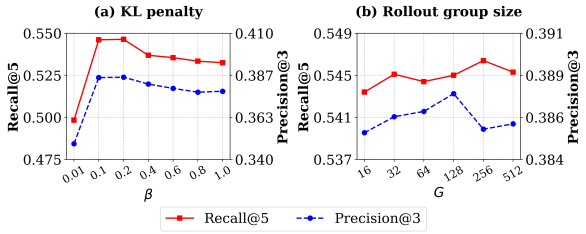

Figure 8: Performance of AuthGR under different hyperparameters: (a) varying coefficient of the KL penalty β and (b) varying size of rollout group G.

SFT baseline, validating that our progressive training approach effectively enhances authority-aware reasoning beyond standard supervised fine-tuning.

E.2 Impact of Reward Scaling Strategies

To effectively leverage our fine-grained signals, we evaluate four reward scaling strategies in GRPO: (i) *Binary* (Shao et al., 2024), mapping scores above a fixed threshold (e.g., 80) to 1 and others to 0; (ii) *Binning*, discretizing scores into five ordinal classes; (iii) *Sigmoid*, nonlinear transformation to emphasize higher scores; and (iv) *Linear*, which directly uses the raw scores. Table 10 demonstrates that the Linear strategy consistently outperforms all alternatives across all metrics. This suggests that preserving fine-grained authority differences is more effective than coarse approximations.

E.3 Hyperparameter Study

Figure 8 shows the performance of AuthGR when varying the coefficient of the KL penalty β and varying the size of the rollout group n. (i) The optimal β lies in the range of 0.1-0.2. When setting β as 0.2 from 0.01, the performance improves by 10.6% and 9.7% in Precision@3 and Recall@5, respectively. It highlights that model performance is highly sensitive to β. Moreover, explicitly regularizing the KL divergence between the trained and reference policies is crucial for preserving the relevance-based ID generation ability acquired during the SFT stage. (ii) Regarding rollout size, the best performance is observed when n ranges from 128 to 512. In contrast, too few rollouts provide insufficient information to compute meaningful relative reward advantages within each group.

Training stage	Recall	Avg. rank (↓)	Avg. confidence (↑)
CPT+SFT (Baseline)	7,298	3.9246	0.2263
+ GRPO (Binary)	7,438 (+1.9%)	**3.9091**	0.2279
+ GRPO (Linear)	**7,817 (+7.1%)**	3.9478	**0.2288**

Table 11: Impact of GRPO on accuracy and confidence. Recall counts queries with ground-truth DocID in the outputs. Avg. rank and confidence measure the position of ground-truth DocIDs and generation probability.

Data filtering	P@3	R@5	R@10
w/o filtering	0.3194	0.4642	0.6299
w/ filtering	0.3725	0.5293	0.7031
Gain (%)	16.62	14.02	11.62

Table 12: Performance comparison with and without data filtering in the SFT stage. "Gain" represents the relative percentage improvement achieved after filtering.

E.4 Impact on Retrieval Accuracy and Confidence

To analyze the impact of GRPO on authority and confidence, Table 11 compares the CPT+SFT baseline against GRPO-Binary (mapping scores > 80 to 1) and GRPO-Linear (using raw scores). The Linear formulation proves most effective, achieving 7.1% gain in Recall over the baseline and increased average confidence. This confirms that fine-grained linear rewards provide superior calibration compared to coarse binary signals, enabling the model to retrieve ground-truth documents with greater confidence.

E.5 Impact of Data Filtering in SFT

Since raw click logs are inherently noisy, we investigated the impact of data quality on the SFT stage as shown in Table 12. Empirically, the best performance was obtained at a threshold of 2.1. At this point, approximately 63% of the original data was filtered out, reducing the number of training samples from nearly 11 million to just over 4 million. Despite this substantial reduction, model performance improved markedly across all metrics with Precision@3 rising by 16.6% and Recall@5 by 14.0%, and Recall@10 by 11.6%. These results provide strong empirical evidence that the *quality* of supervision data is more critical than its *quantity* in generative retrieval. By removing noise, the model can learn robust mappings between queries and documents.

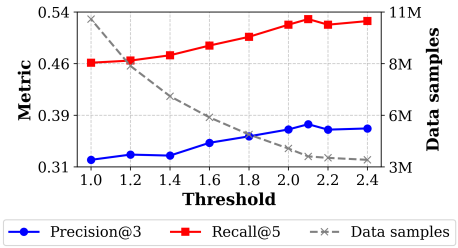

Figure 9: Performance over varying filtering thresholds. The numbers of data samples are shown in millions.

E.6 Performance over Data Filtering Thresholds

Figure 9 illustrates the impact of relevance-based data filtering on model performance. We filter low-quality query-document pairs from raw click logs, varying the threshold from 1.0 to 2.2. As the threshold increases, the training dataset size decreases from 10.99M to 3.95M samples. Performance peaks at 2.1, which we adopt as the optimal setting. Beyond this point, aggressive filtering degrades results due to insufficient training signals. It confirms that data quality is critical for learning robust query-to-DocID mappings.

F Annotation Guidelines for Human Evaluation

Annotators performed a blind side-by-side comparison to assess the overall quality of top-ranked search result. We used a 5-point Likert scale designed to holistically capture both information relevance and source authority. The detailed criteria are as follows:

- **5 (Excellent)**: The document fully satisfies the user's information needs with comprehensive and accurate content. The source is highly authoritative (e.g., official government website, public institution, or major corporation). This is the ideal result for the query.
- **4 (Good)**: The document provides high-quality and relevant information that effectively addresses the need. The source is trustworthy but lacks top-tier official status (e.g., well-maintained expert blog, reputable news organization, or major community platform). The content is trustworthy, but the source itself lacks official status.
- **3 (Fair)**: The document partially satisfies the needs, but coverage is limited. The source has low or indeterminate authority (e.g., personal blog, forum post, or small commercial site).
- **2 (Poor)**: The document is marginally relevant with very limited information, requiring additional searches to satisfy the need. The source quality is typically low.
- **1 (Bad)**: The document is irrelevant and very low quality (e.g., spam, heavy ads), or inaccessible (e.g., broken links).

In addition to the 1-5 rating, annotators provided a direct preference judgment ("Which result is better overall?") for each query to calculate the win ratio used in our main analysis.

G Role of Continued Pre-Training in Search Domain

Continued pre-training (CPT) on domain-specific corpora is essential for adapting LLMs to search applications, bridging the gap between generic linguistic knowledge and relevance-oriented retrieval tasks. Without CPT, LLMs often remain biased toward generic next-token distributions, failing to capture implicit user relevance signals that imply retrieval and ranking. Ye et al. (2025) demonstrate that applying CPT on a click-stream corpus, structured as *(query, clicked title, summary)* triples, effectively aligns models with user interaction patterns and domain-specific semantics before fine-tuning. Following CPT, their framework fine-tunes the LLM with a pairwise ranking loss and subsequently distills knowledge into a BERT-based encoder via a hybrid loss combining absolute score regression and relative ranking margins. Their experiments confirm that CPT yields statistically significant gains, boosting NDCG@5 from 0.8709 to 0.8793. Importantly, this CPT-based pipeline was also deployed in a commercial search product, highlighting its practical necessity beyond controlled benchmarks. These findings establish CPT not as an optional enhancement, but as an indispensable phase for ensuring domain alignment and performance in real-world search systems.