HeadRank: Decoding-Free Passage Reranking via Preference-Aligned Attention Heads

Anonymous ACL submission

Abstract

Decoding-free reranking methods that read relevance signals directly from LLM attention weights offer significant latency advantages over autoregressive approaches, yet suffer from attention score homogenization: middle-context documents receive near-identical scores, destroying the fine-grained distinctions required for ranking. We propose HeadRank, a framework that lifts preference optimization from discrete token space into the continuous attention domain through entropy-regularized head selection, hard adjacent-level preference pairs, and a distribution regularizer that jointly sharpen discriminability in the homogenized middle zone. Depth truncation at the deepest selected layer further reduces inference to $\mathcal{O}(1)$ forward passes. Across 14 benchmarks on three Qwen3 scales (0.6B–4B) using only 211 training queries, HeadRank achieves the highest average NDCG@10 at every scale, outperforming both generative and decoding-free baselines on the majority of benchmarks with 100% formatting success. At 4B, 57.4% of relevant middle-zone documents reach the top quartile versus 14.2% for irrelevant ones—a 43-percentage-point selectivity gap that demonstrates the effectiveness of attention-space preference alignment for listwise reranking.

1 Introduction

Reranking with Large Language Models (LLMs) has become a dominant paradigm in information retrieval (Zhu et al., 2023). Generative approaches—listwise (Sun et al., 2023), pointwise (Nogueira et al., 2020), pairwise (Qin et al., 2024), and setwise (Zhuang et al., 2024)—require at least one decoding step per candidate, incurring latency that scales with list length and risking malformed outputs that silently corrupt rankings.

A recent line of work sidesteps generation entirely by reading relevance signals from attention weights during the prefill pass (Chen et al., 2025; Tran et al., 2025; Na et al., 2025)—no decoding, no formatting risk. However, we find that these methods hit a wall as the candidate list grows due to what we call *attention score homogenization*: attention weights for middle-context documents become nearly uniform, destroying the fine-grained distinctions that ranking requires. Current head-selection heuristics (Tran et al., 2025; Zhang et al., 2025b) exacerbate this by latching onto noisy, high-entropy heads never trained for ranking.

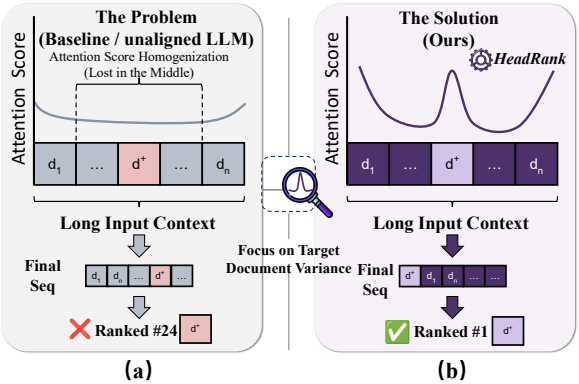


Figure 1: **Comparison of reranking paradigms.**

We resolve this with **HeadRank**, which combines three complementary ideas: (i) entropy-regularized head selection that isolates stable, low-entropy retrieval heads; (ii) Adjacent-Level Preference Sampling (ALPS) that constructs hard contrastive pairs from adjacent relevance grades; and (iii) an attention-space DPO objective with a distribution regularizer that explicitly fights score flatlining. At inference time, the forward pass is truncated at the deepest selected layer $l_{\max}$, yielding $\mathcal{O}(1)$-pass complexity. Figure 1 contrasts the failure mode with our solution: baseline methods produce flat attention curves that bury relevant documents in the middle zone, whereas HeadRank restores discriminative peaks.

Our main contributions are twofold: **(1)** We identify *attention score homogenization* as the mechanistic bottleneck of decoding-free reranking and

propose HeadRank, which lifts DPO into continuous attention space with entropy-regularized head discovery and Adjacent-Level Preference Sampling (ALPS), enabling preference alignment without autoregressive decoding. A depth-truncation strategy further reduces inference cost to $l_{\max}/L$ of a full forward pass. **(2)** Experiments on 14 benchmarks show that HeadRank trained on only 211 queries achieves the best average performance across both generative and decoding-free baselines at all three model scales (0.6B–4B), with 100% formatting success rate and substantial latency savings.

2 Related Work

LLM-Based Re-ranking. LLM reranking methods are broadly categorized into generative and attention-based paradigms. Generative approaches prompt LLMs to produce ranked outputs through autoregressive decoding: RankGPT (Sun et al., 2023) outputs listwise permutations, while pointwise (Nogueira et al., 2020), pairwise (Qin et al., 2024), and setwise (Zhuang et al., 2024) variants reduce output length but still require at least one decoding step per candidate. Recent efforts apply reinforcement learning to generative rerankers (Zhuang et al., 2025; Zhang et al., 2025a), yet these methods remain in token space and inherit decoding overhead and formatting risks. In contrast, attention-based reranking sidesteps generation entirely. ICR (Chen et al., 2025) first demonstrated that aggregated attention scores from a single prefill pass can serve as relevance signals, eliminating decoding latency. Subsequent work identifies specialized "retrieval heads" responsible for information recall (Wu et al., 2025; Yin and Steinhardt, 2025): CoRe (Tran et al., 2025) discovers such heads via a contrastive metric, QR-R (Zhang et al., 2025b) uses causal interventions, and NIAH (Wu et al., 2025) leverages needle-in-a-haystack probing. However, all existing attention-based methods select heads through post-hoc heuristics without explicit preference optimization, leaving them vulnerable to score homogenization when the candidate list grows.

Attention Patterns and Positional Bias in LLMs. The premise that attention weights encode relevance is empirically supported (Serrano and Smith, 2019; Wiegreffe and Pinter, 2019), and mechanistic interpretability has revealed functional specialization among attention heads—with only a sparse subset being critical for specific tasks (Wu et al., 2025; Yin and Steinhardt, 2025). A closely related challenge is the "Lost in the Middle" phenomenon (Liu et al., 2024; Peysakhovich and Lerer, 2023), where LLMs systematically under-attend to middle-context information. This behavior is linked to "attention sinks" that divert probability mass to structurally prominent tokens (Xiao et al., 2023). Permutation ensembling (Tang et al., 2023) mitigates positional bias but multiplies latency proportionally. Our work provides a mechanistic explanation for this phenomenon at the level of individual attention heads—what we term *attention score homogenization*—and directly resolves it through preference alignment rather than inference-time ensembling.

Preference Optimization for Ranking. Direct Preference Optimization (DPO) (Rafailov et al., 2023) aligns LLMs via token-level generation probabilities without explicit reward models, and has been adapted for generative reranking. However, applying DPO directly to reranking requires autoregressive decoding, defeating the latency benefits of attention-based approaches. HeadRank projects preference alignment from discrete token space into continuous attention space, enabling DPO-style optimization over attention head activations without any generation step.

3 Methodology

3.1 Problem Formulation and Preliminaries

Document reranking aims to compute a relevance score $s(q, d_i)$ for each document in a candidate set $\mathcal{D} = \{d_1, d_2, \ldots, d_N\}$, which is typically retrieved by an initial ranker such as BM25, to produce a sorted permutation. Modern LLMs approach this task through stacked Transformer layers, whose internal attention distributions provide a natural relevance signal.

Generative reranking methods inherently rely on computationally expensive autoregressive decoding. In contrast, attention-based approaches utilize internal activations during the prefill phase. Following the paradigm of In-Context Re-ranking (ICR), the input sequence is formatted by placing documents before the query, denoted as $X = [\text{Instruction} \oplus d_1 \oplus \cdots \oplus d_N \oplus q]$, to satisfy the causal attention mask requirements of decoder-only LLMs. This arrangement ensures that query tokens can attend to preceding document tokens.

For a specific attention head h in layer l, let $A_{i,j}^{(l,h)}$ denote the attention weight from the i-th

Table 1: NDCG@10 across in-domain and out-of-domain benchmarks with BM25 top-40 retrieval across Qwen3 model sizes. **SR (%)** denotes the formatting Success Rate of the generated outputs.

		In-Domain		Out-of-Domain										
Model	**Method**	**DL19**	**DL20**	**COVID**	**ArguAna**	**News**	**FiQA**	**SciDocs**	**NFC**	**NQ**	**DBPedia**	**FEVER**	**Avg.**	**SR (%)**
Baseline	BM25 (top40)	50.58	47.96	59.47	29.99	39.52	23.61	14.90	33.75	30.55	31.80	65.13	38.84	100.0
Qwen3-0.6B	RankGPT	49.17	48.54	60.28	26.44	37.81	23.22	14.60	33.48	30.89	32.18	60.85	37.95	76.6
	ICR	50.09	41.75	63.28	32.43	41.13	22.70	13.75	34.06	33.62	32.57	65.59	39.18	100.0
	CoRe	50.46	44.24	65.53	33.61	42.55	25.66	15.48	33.22	36.61	34.20	71.00	41.14	100.0
	QR-R	51.96	43.32	67.84	34.39	40.04	24.42	13.93	34.28	35.45	33.22	66.18	40.46	100.0
	NIAH	46.55	38.87	59.76	33.87	39.34	23.53	14.58	32.40	33.83	31.82	71.65	38.75	100.0
	HeadRank	**59.93**	**51.71**	**68.71**	**34.62**	**44.02**	**25.99**	**16.36**	**35.09**	**43.25**	**37.02**	**73.47**	**44.56**	**100.0**
Qwen3-1.7B	RankGPT	56.32	52.77	68.15	27.47	46.30	25.95	16.46	28.82	39.50	36.44	54.69	41.17	97.1
	ICR	52.11	46.14	65.42	30.63	43.47	27.17	15.59	34.65	37.09	33.24	69.39	41.35	100.0
	CoRe	51.26	45.81	68.67	34.59	42.71	**28.49**	16.57	34.91	40.00	34.39	71.02	42.58	100.0
	QR-R	52.36	46.56	67.77	33.52	43.89	27.61	16.00	34.70	40.00	33.36	65.45	41.93	100.0
	NIAH	48.98	43.25	63.39	33.00	41.22	26.29	15.44	33.42	35.98	32.15	**72.59**	40.52	100.0
	HeadRank	**61.32**	**57.85**	**70.81**	**36.59**	**46.40**	28.11	**16.96**	**36.23**	**43.32**	**37.06**	71.86	**46.05**	**100.0**
Qwen3-4B	RankGPT	63.58	59.52	71.12	27.72	**49.32**	27.23	16.68	30.24	44.86	**41.34**	62.18	44.89	95.8
	ICR	54.38	48.17	66.63	23.68	42.69	24.53	14.47	34.61	37.67	32.81	61.06	40.06	100.0
	CoRe	56.67	50.89	70.08	32.93	43.32	**29.88**	16.52	35.13	42.80	35.84	**73.39**	44.31	100.0
	QR-R	57.45	53.84	73.04	32.35	44.46	28.18	15.98	35.68	44.30	36.27	69.82	44.67	100.0
	NIAH	48.13	40.94	63.64	24.91	38.30	25.00	14.69	32.26	34.40	30.82	70.83	38.54	100.0
	HeadRank	**64.33**	**59.89**	**73.94**	**35.30**	46.27	28.73	**16.89**	**36.44**	**45.36**	37.24	69.69	**46.73**	**100.0**

token to the j-th token, quantifying the pairwise information flow. Rather than relying on standard decoding, we directly extract these internal distributions. The relevance score $\alpha_d^{(l,h)}$ for a specific head (l, h) is defined as the aggregated attention mass from query tokens to document tokens:

$$\alpha_d^{(l,h)} = \frac{1}{|\mathcal{I}_q|} \sum_{i \in \mathcal{I}_q} \sum_{j \in \mathcal{I}_d} A_{i,j}^{(l,h)} \tag{1}$$

where $\mathcal{I}_q$ and $\mathcal{I}_d$ denote the token index sets for the query and the document, respectively.

3.2 Entropy-Regularized Core Head Selection

Identifying the right attention heads is critical: contrastive metrics alone can select heads that are nominally discriminative but exhibit chaotic, high-entropy distributions, introducing noise during alignment. Concurrent work on retrieval-head contrasting for hallucination mitigation underscores the importance of principled head identification. We therefore propose an entropy-regularized discovery mechanism.

Given a relevant document d^+ and a set of irrelevant documents D^-, we follow CoRe (Tran et al., 2025) and first compute a base discriminative score $S_{\text{disc}}^{(l,h)}$ for head h in layer l. This score employs a temperature-scaled softmax function to measure the effectiveness of the head in distinguishing the positive document from the negative documents:

$$S_{\text{disc}}^{(l,h)} = \frac{e^{\alpha_{d^+}^{(l,h)}/\tau}}{\sum_{d \in \{d^+\} \cup D^-} e^{\alpha_d^{(l,h)}/\tau}} \tag{2}$$

To penalize heads with dispersed attention, we define an entropy gating factor. Let $A_t^{(l,h)} = \frac{1}{|\mathcal{I}_q|} \sum_{i \in \mathcal{I}_q} A_{i,t}^{(l,h)}$ denote the query-averaged attention weight to the t-th token, obtained by marginalizing $A_{i,j}^{(l,h)}$ over query positions. The Shannon entropy of this distribution over the sequence length L_{seq} is denoted as $\mathcal{H}^{(l,h)} = -\sum_{t=1}^{L_{\text{seq}}} A_t^{(l,h)} \log A_t^{(l,h)}$. This value is normalized by the maximum possible entropy $\log L_{\text{seq}}$ to define the gating factor $G_{\text{ent}}^{(l,h)}$:

$$G_{\text{ent}}^{(l,h)} = 1 - \lambda \frac{\mathcal{H}^{(l,h)}}{\log L_{\text{seq}}} \tag{3}$$

where λ represents a hyperparameter that scales the penalty.

The final selection criterion $\Phi(l, h)$ combines both components to favor heads with a sharp focus while downplaying those with high-entropy distributions:

$$\Phi(l, h) = S_{\text{disc}}^{(l,h)} \cdot G_{\text{ent}}^{(l,h)} \tag{4}$$

The top K heads are selected based on $\Phi(l, h)$, and the maximum layer index among these heads determines the depth for inference termination, denoted as $l_{\max}$.

Table 2: Recall@2 and Recall@5 on knowledge-intensive multi-hop reasoning benchmarks (HotpotQA, 2WikiMultihopQA, MuSiQue) with ColBERTv2 top-20 retrieval. **SR** denotes the formatting Success Rate of generative methods. Reference rows (*Mistral 7B*, *Llama-3.1 8B*, *GPT-3.5 Turbo*) run RankGPT on their respective models.

		HotpotQA		2Wiki		MuSiQue		Avg.		
Model	**Method**	**R@2**	**R@5**	**R@2**	**R@5**	**R@2**	**R@5**	**R@2**	**R@5**	**SR**
	ColBERTv2	64.65	79.30	59.23	68.20	37.94	49.52	53.94	65.67	—
Qwen3-0.6B	RankGPT	52.85	62.90	39.50	46.70	30.42	37.49	40.92	49.03	80.1%
	ICR	57.50	75.35	51.28	62.95	31.23	43.69	46.67	60.66	100%
	HeadRank	**67.50**	**85.25**	**60.98**	**72.12**	36.72	**51.02**	**55.07**	**69.46**	100%
Qwen3-1.7B	RankGPT	57.25	73.20	46.45	59.15	35.02	47.83	46.24	60.06	93.7%
	ICR	66.95	82.10	59.65	70.60	**39.61**	51.07	55.40	67.92	100%
	HeadRank	**69.35**	**86.20**	**62.65**	**72.88**	39.05	**54.47**	**57.02**	**71.18**	100%
Qwen3-4B	RankGPT	71.60	84.25	**65.18**	73.28	41.25	52.99	59.34	70.17	98.9%
	ICR	57.45	74.20	53.93	66.00	34.23	46.52	48.54	62.24	100%
	HeadRank	**72.65**	**86.45**	63.60	**73.75**	**42.67**	**55.10**	**59.64**	**71.77**	100%
Mistral 7B	RankGPT	64.80	79.80	56.30	67.10	36.20	48.70	52.40	65.20	20.5%
Llama-3.1 8B	RankGPT	69.90	84.80	60.80	72.00	39.30	51.40	56.60	69.40	98.2%
GPT-3.5 Turbo	RankGPT	71.80	84.90	61.60	72.80	38.70	54.10	57.40	70.60	98.6%

3.3 Attention-based Direct Preference Optimization (HeadRank)

Standard DPO (Rafailov et al., 2023) aligns token-level generation probabilities, so applying it directly to reranking requires autoregressive decoding and its attendant latency. We therefore project the alignment paradigm from discrete token space into continuous attention space.

Let $s_\theta^{(d)}$ denote the relevance score computed exclusively by aggregating the attention mass from the selected core heads. For a preference pair, let s_θ^+ and s_θ^- be the scores for the chosen and rejected documents, respectively. To prevent the objective from degenerating into disjoint heuristics, we unify the margin constraints and anti-homogenization penalties into a single loss. Let $\Delta s_\theta = s_\theta^+ - s_\theta^-$ denote the predicted score margin, and $\mathbf{\Delta}_{\text{ref}} = [s_\theta^+ - s_{\text{ref}}^+, s_\theta^- - s_{\text{ref}}^-]^\top$ denote the reference divergence vector. The unified HeadRank objective is formulated as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{align}}(\Delta s_\theta) + \mathcal{L}_{\text{prox}}(\mathbf{\Delta}_{\text{ref}}) + \Omega(\boldsymbol{s}_\theta) \quad (5)$$

where $\mathcal{L}_{\text{prox}}(\mathbf{\Delta}_{\text{ref}}) = \frac{\beta}{2}\|\mathbf{\Delta}_{\text{ref}}\|_2^2$ is the proximal penalty, and the pairwise alignment objective $\mathcal{L}_{\text{align}}$ and the distribution regularizer $\Omega(\boldsymbol{s}_\theta)$ are formally defined as:

$$\mathcal{L}_{\text{align}}(\Delta s_\theta) = -\log\sigma(\Delta s_\theta) + [m - \Delta s_\theta]_+ - \alpha\Delta s_\theta \quad (6)$$

$$\Omega(\boldsymbol{s}_\theta) = \gamma\mathcal{H}(\boldsymbol{p}) - \eta\,\mathrm{Var}(\boldsymbol{s}_{\text{mid}}) \quad (7)$$

Here, $[x]_+ = \max(0, x)$ denotes the standard hinge operator with margin constraint m, preventing overconfidence. The linear term $-\alpha\Delta s_\theta$ complements the sigmoid: as Δs_θ grows, $-\log\sigma(\cdot)$ saturates and its gradient vanishes, whereas the linear term maintains a steady gradient that continues to encourage margin expansion; the proximal penalty $\mathcal{L}_{\text{prox}}$ bounds this push by anchoring scores near the reference model. To maintain the pre-trained retrieval capabilities and prevent representation collapse, we introduce a **Proximal Policy Retention** ($\mathcal{L}_{\text{prox}}$) term. Rather than the KL divergence used in standard DPO, we adopt an L_2 proximal penalty $\|\mathbf{\Delta}_{\text{ref}}\|_2^2$, which prevents the aligned policy from deviating too far from the frozen reference in the continuous attention domain.

$\Omega(\boldsymbol{s}_\theta)$ is a distribution regularizer that directly targets "Lost in the Middle" score homogenization. $\mathcal{H}(\boldsymbol{p})$ is the Shannon entropy of the document score distribution $\boldsymbol{p}$; minimizing it sharpens the overall score spread. Simultaneously, maximizing $\mathrm{Var}(\boldsymbol{s}_{\text{mid}})$—the variance of intermediate document scores[1]—breaks the score-flattening effect among middle-ranked candidates, restoring fine-grained discriminability across the full context window.

3.4 Early-Exit Inference via Depth Truncation

HeadRank is entirely *decoding-free*: relevance scores are extracted from attention weights during the prefill phase, bypassing autoregressive generation entirely. Since all K core heads reside at or below layer $l_{\max}$, the forward pass safely exits at that depth, consuming only $l_{\max}/L$ of the full-model FLOPs. Together, the decoding-free ar-

[1]The "middle zone" comprises documents at BM25 ranks in the 25th–75th percentile (positions 11–30 for top-40, positions 6–15 for top-20).

chitecture and early termination yield $\mathcal{O}(1)$-pass inference with zero formatting failures.

3.5 Data Construction and Iterative Alignment

To bypass the heavy data reliance of generative models and synergize with the distribution regularizer $\Omega(\boldsymbol{s}_\theta)$, an efficient data construction pipeline is established using 211 queries from MS MARCO V2. The top-100 documents are retrieved via BM25 to form the context. To ensure pair quality, a margin-aware Adjacent-Level Preference Sampling (ALPS) strategy is applied. Motivated by the finding that fine-grained relevance labels improve zero-shot LLM rankers (Zhuang et al., 2023a), preference pairs exhibiting a relevance gap larger than one level (e.g., 3 versus 0 or 3 versus 1) are strictly discarded, restricting the construction exclusively to adjacent-level pairs (e.g., 3 versus 2). This formulation forces the attention heads to learn fine-grained, discriminative ranking features, directly complementing the intermediate score variance objective.

These same pairs are used for both core head selection and optimization. Because the attention topology shifts during training, an iterative refinement step is employed. After an initial training round, we re-run head selection on the updated weights to recalibrate the core set, keeping it aligned with the post-training attention topology. The complete training and inference procedure is summarized in Algorithm 1 (Appendix H).

4 Experiments

4.1 Experimental Setup

Datasets and Metrics. We evaluate reranking effectiveness, out-of-domain generalization, and long-context reasoning across 14 benchmarks, organized into three groups. The first group comprises standard passage retrieval benchmarks, specifically TREC-DL19 and TREC-DL20, to assess in-domain alignment. The second group evaluates cross-domain transferability using seven diverse datasets from the BEIR benchmark: TREC-COVID (Voorhees et al., 2021), ArguAna, TREC-News, FiQA, SciDocs, NFCorpus (NFC), and DBPedia. The third group evaluates knowledge-intensive and multi-hop reasoning capabilities over long contexts using Natural Questions (NQ), FEVER, HotpotQA, 2WikiMultihopQA (2Wiki), and MuSiQue. For the passage retrieval and out-of-domain benchmarks, as well as NQ and FEVER, the top-40 candidate documents are initially retrieved using BM25 (Robertson et al., 2009). For multi-hop reasoning datasets, we follow the methodology of Gutiérrez et al. (Gutiérrez et al., 2024) and employ ColBERTv2 (Santhanam et al., 2022) to retrieve the top-20 candidate passages. For evaluation metrics, Normalized Discounted Cumulative Gain at rank 10 (NDCG@10) serves as the primary metric for the BM25-retrieved datasets, while Recall@2 and Recall@5 are measured for the ColBERTv2-retrieved multi-hop datasets. We also report Formatting Success Rate (SR) for generative methods—the fraction of queries yielding a correctly formatted ranking—to quantify output reliability.

Baselines. The proposed HeadRank framework is compared against five state-of-the-art reranking methods, encompassing both generative and decoding-free attention-based paradigms. For generative reranking, **RankGPT** (Sun et al., 2023) is adopted as a representative baseline, outputting sorted permutations via autoregressive decoding (prompt template in Appendix F). Supervised fine-tuned rerankers such as RankLLaMA (Ma et al., 2024) and query-likelihood approaches (Zhuang et al., 2023b) are orthogonal to our decoding-free paradigm and therefore excluded from the main comparison. For decoding-free attention-based methods, we compare against four closely related approaches. **ICR** (Chen et al., 2025) serves as the foundational baseline by uniformly aggregating global attention scores. **NIAH** (Wu et al., 2025) evaluates reranking using specific heads identified via the Needle-In-A-Haystack capability. **QR-R** (Zhang et al., 2025b) leverages causal interventions to isolate query-dependent attention pathways. Finally, **CoRe** (Tran et al., 2025) represents the current state-of-the-art, discovering retrieval heads based on a discrete contrastive metric without continuous alignment optimization.

Implementation Details. We build HeadRank on three Qwen3 checkpoints (Yang et al., 2025) (0.6B, 1.7B, and 4B parameters). Training data consists of 211 queries from the MS MARCO V2 passage corpus; for each query, ALPS constructs adjacent-level preference pairs from BM25 top-100 candidates, yielding $\sim$91K training pairs for the 0.6B/1.7B models and $\sim$200K for the 4B model (the larger model's richer attention topology produces more head-level contrasts that pass the entropy gate, expanding the effective pair set). We

Table 3: Ablation study of the HeadRank framework (Qwen3-0.6B). **Ours** utilizes Top-8 core heads, ALPS data construction, and the full joint DPO objective. NQ and FiQA are evaluated with NDCG@10; 2Wiki and Hotpot with Recall@5.

Method	NQ	FiQA	2Wiki	Hotpot
Ours (Full Framework)	**43.25**	**25.99**	**72.12**	**85.25**
w/ SFT (RankNet)	37.39	22.92	69.65	79.60
w/ all heads	20.52	8.97	67.05	54.80
w/o Entropy Gating G_{ent}	39.50	24.12	69.80	82.15
w/o ALPS strategy	40.85	23.50	70.45	83.20
w/o regularizer $\Omega(\boldsymbol{s}_\theta)$	38.60	23.95	68.20	80.50
w/o proximal penalty $\mathcal{L}_{\text{prox}}$	36.45	21.30	68.85	80.10

select $K{=}8$ core retrieval heads via entropy gating; head distributions before and after alignment are reported in Appendix E. All models are trained for a single epoch. Because several attention-based baselines (e.g., ICR) rely on calibration (Zhao et al., 2021) to remove positional bias, we apply the same calibration protocol uniformly across all methods—including HeadRank—at both training and inference time, ensuring a fair comparison. At inference, all attention-based methods process the full BM25 top-40 list in a single forward pass with batch size 16, avoiding the latency and contextual fragmentation of iterative sliding-window decoding used by RankGPT (window size 20, stride 10). Full hyperparameters (β, α, λ, τ, learning rates, hardware) are in Appendix D.

4.2 In-Domain and Out-of-Domain Retrieval.

Table 1 presents NDCG@10 performance across two in-domain TREC-DL benchmarks and nine out-of-domain datasets (including BEIR subsets, NQ, and FEVER). HeadRank achieves the highest average NDCG@10 across all Qwen3 model scales (0.6B, 1.7B, and 4B), outperforming generative and decoding-free baselines as well as BM25.

For in-domain alignment on TREC-DL19 and TREC-DL20, HeadRank at 4B matches or exceeds every baseline, including the generative upper bound RankGPT. This demonstrates that projecting preference constraints into continuous attention space sharpens discriminability in long contexts.

HeadRank generalizes robustly out-of-domain despite training on only 211 MS MARCO queries. Combining ALPS and $\Omega(\boldsymbol{s}_\theta)$ yields consistent gains across medicine (TREC-COVID), encyclopedias (DBPedia), and fact-checking (FEVER). Under Qwen3-1.7B, HeadRank reaches 43.32 on NQ, substantially outperforming zero-shot attention baselines like ICR (37.09) and NIAH (35.98) that lack explicit preference alignment. This suggests the model learns generalizable semantic relevance rather than overfitting.

On a few benchmarks, generative methods retain an edge: RankGPT (4B) outperforms HeadRank on News (+3.1) and DBPedia (+4.1), likely because these knowledge-grounded datasets benefit from full autoregressive reasoning. Similarly, CoRe marginally leads on FiQA at 1.7B (+0.38). However, these per-dataset advantages do not transfer to the aggregate: HeadRank leads in overall average NDCG@10 at every scale (44.56 vs. 41.14 at 0.6B; 46.05 vs. 42.58 at 1.7B; 46.73 vs. 44.89 at 4B), confirming that attention-space alignment provides more *consistent* cross-domain gains.

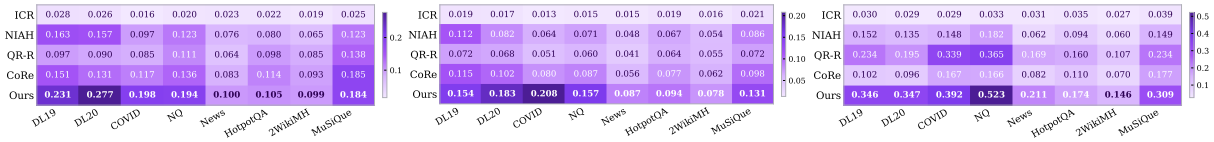
Moreover, all attention-based methods, including HeadRank, achieve a 100% formatting success rate by construction, whereas RankGPT drops to 76.6% SR on Qwen3-0.6B due to malformed autoregressive outputs.

4.3 Multihop Reasoning Capabilities.

Table 2 presents the results for knowledge-intensive multihop reasoning tasks (HotpotQA, 2Wiki, and MuSiQue). Unlike single-hop retrieval, these tasks require the reranker to identify multiple complementary evidence documents whose relevance emerges only through cross-document inference chains—a setting where attention-based scoring can capture inter-document bridging entities that generative methods must reason over token-by-token. HeadRank consistently improves retrieval across all Qwen3 model scales, achieving the highest average recall. Specifically, utilizing the Qwen3-4B model, our method obtains an average Recall@2 of 59.64 and Recall@5 of 71.77. Notably, this 4B model outperforms larger generative baselines—Llama-3.1 8B (56.60 R@2) and GPT-3.5 Turbo (57.40 R@2) running RankGPT—despite being roughly half their size. Furthermore, complex multihop contexts severely disrupt the formatting stability of generative methods, as evidenced by Mistral 7B dropping to a 20.5% Success Rate (SR). In contrast, HeadRank guarantees a 100% SR, confirming the absolute robustness of the decoding-free alignment paradigm.

4.4 Ablation Studies

To dissect HeadRank's components, we conduct ablations on the Qwen3-0.6B architecture (Table



(a) Qwen3-0.6B (b) Qwen3-1.7B (c) Qwen3-4B

Figure 2: **Middle-zone normalized attention-score standard deviation** (↑ better) across five methods, eight datasets, and three model scales. Lighter cells indicate more severe attention homogenization.

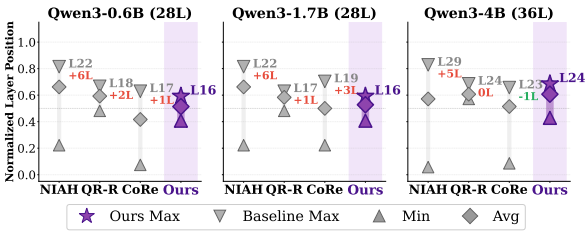


Figure 3: **Depth distribution of selected core heads** (Qwen3-0.6B). Heads above the dashed line at $l_{\max}$ are pruned for early-exit inference.

3).

Core Heads & Entropy Gating. Removing core head isolation (*w/ all heads*) causes catastrophic degradation (e.g., FiQA drops from 25.99 to 8.97), proving global attention is heavily contaminated by noise. Omitting entropy gating (*w/o* G_{ent}) consistently decreases performance, confirming that purely contrastive selection isolates unstable, high-entropy heads.

ALPS & Distribution Regularizer. Replacing ALPS with random pairs (*w/o ALPS*) degrades accuracy, validating that adjacent-level contrasting is essential for fine-grained discrimination. Removing the distribution regularizer (*w/o* $\Omega(\boldsymbol{s}_\theta)$) causes a clear drop on long-context multihop tasks, confirming its role in preventing score homogenization among middle-ranked candidates.

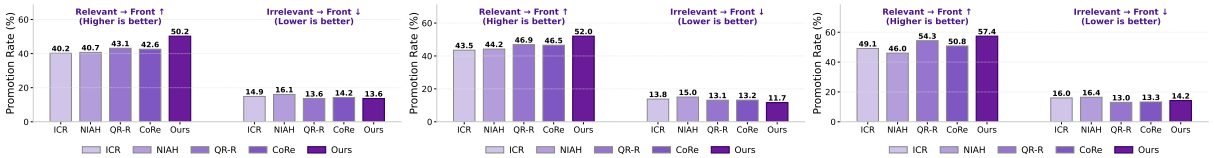
Proximal Penalty. Finally, discarding the proximal constraint (*w/o* $\mathcal{L}_{\text{prox}}$) causes representation collapse and catastrophic forgetting, severely impairing zero-shot out-of-domain generalization on NQ and FiQA.

Attention-based DPO vs. Standard SFT. Replacing DPO with pairwise SFT (*w/ SFT*, RankNet-style $\sigma(s_c - s_r)$) consistently degrades all four evaluation sets. Without the frozen reference penalty $\mathcal{L}_{\text{prox}}$, unconstrained SFT aggressively fits hard targets, distorting the pre-trained attention topology and triggering representation collapse; the proximal anchor preserves linguistic priors while steering toward relevance preferences.

4.5 Deep Analysis

Attention Homogenization Across Methods and Scales. To determine the pervasiveness of score flatlining in the middle context zone and assess whether any existing method effectively mitigates it, Figure 2 diagnoses this phenomenon across five methods, eight datasets, and three model scales using middle-zone normalized standard deviation. The results reveal a systemic vulnerability in baseline approaches: ICR rows remain near-white throughout, indicating a complete flatlining of its scoring function. Similarly, NIAH and CoRe recover only marginal variance, while QR-R exhibits moderate but inconsistent improvements. In stark contrast, **HeadRank** consistently occupies the darkest row across all 24 experimental conditions, with its advantage widening substantially at the 4B scale. This robust discriminability is directly attributable to the $\mathrm{Var}(\boldsymbol{s}_{\text{mid}})$ regularization term within $\Omega(\boldsymbol{s}_\theta)$, which serves as a structural wedge to forcibly break attention homogenization. Furthermore, our entropy gating mechanism, G_{ent}, ensures that this discriminative pressure is concentrated exclusively on stable, low-entropy heads, preventing noisy heads from dominating the variance and stabilizing the optimization process. Cross-dataset universality and scaling behavior are further validated in Appendix C, with per-dataset radar profiles in Appendix B.

From Variance to Ranking Precision. Does this structural variance actually translate into correct positional assignments? We quantify *middle-to-front promotion rates*—the fraction of relevant vs. irrelevant documents originally in the middle zone (BM25 ranks in the 25th–75th percentile) that are reranked into the top quartile, averaged across eight datasets (Figure 4). HeadRank's selectivity gap (relevant − irrelevant promotion) grows monotonically with scale: 36.6 pp at 0.6B (+7.1 over QR-R), 40.4 pp at 1.7B (achieving the lowest irrelevant promotion rate among all methods), and 43.1 pp at 4B. This monotonic scaling mirrors the variance trends observed earlier, confirming that the score



(a) Qwen3-0.6B (b) Qwen3-1.7B (c) Qwen3-4B

Figure 4: **Middle-to-front promotion rates** averaged across eight datasets. Documents in the middle zone (25th–75th percentile of BM25 ranks) are checked for promotion to the top quartile after reranking. Left bars: relevant documents promoted (↑ better); right bars: irrelevant documents promoted (↓ better).

Table 4: Token-level attention visualization for FEVER. Claim: "Jeb Bush's mother is Laura Bush." Deeper purple indicates higher attention weight. BM25/ICR/RankGPT columns show the rank assigned by each method.

Rank	Passage (token-level attention coloring)	BM25	ICR	RankGPT	Rel.
#1	**Jeb Bush** Jeb Bush John Ellis " " Bush Sr . (born February ,) is an American businessman and politician who served as the 43rd Governor of Florida from to . Bush , who grew up in Houston , is the second son of former President George H . W . . .	13	11	22	✓
#2	**Barbara Bush** Barbara Bush nee Pierce ; born June ,) is the wife of George H . W . Bush , the 41st President of the United States , and served as First Lady of the United States from to . She is the mother of George W . Bush , . . .	1	16	3	–
#3	**Political positions of Jeb Bush** Political positions of Jeb Bush is a Republican politician in the United States . Bush was governor of Florida from to . He was a candidate for the Republican nomination for president of the United States in the election .	2	4	12	–
#4	**Miss Beazley (dog)** Miss Beazley (dog) October – May ,) was a Scottish Terrier which belonged to former U.S . President George W . Bush and former U.S . First Lady Laura Bush . Miss Beazley 's father , a Scottish terrier named Clinton , was born on November , ...	9	1	18	–
#5	**Governor Bush** Governor Bush may refer to : George W . Bush (born) , 46th Governor of Texas (–) and 43rd President of the United States . Jeb Bush (born) , 43rd Governor of Florida (–) Category : and name disambiguation pages	12	37	21	–

headroom generated by $\Omega(s_\theta)$ translates directly into ranking precision. ALPS hard-sample pairs reinforce these dynamics by supplying contrastive signals between adjacent-grade documents within the middle zone, effectively transforming structural score spread into downstream ranking gains.

Dynamic Layer Pruning and Early-Exit Efficiency. HeadRank achieves these discriminability gains while reducing computational overhead. Figure 3 shows that preference alignment compresses core heads into shallower layers compared to baselines (e.g., NIAH and QR-R reach L_{22} in Qwen3-0.6B). By optimizing designated heads to capture relevance signals directly, HeadRank localizes discriminability to earlier stages, enabling safe truncation at $l_{\max} = 16$ (28L models) or $l_{\max} = 24$ (36L model). This early-exit inference translates directly into wall-clock speedups, placing HeadRank on the Pareto frontier of latency versus NDCG@10 (Figure 5, Appendix A).

Case study. Table 4 visualizes token-level attention for the FEVER claim *"Jeb Bush's mother is Laura Bush"* (false—his mother is Barbara Bush). BM25 ranks *Barbara Bush* at #1 via a term-frequency trap ("Bush" + "mother"), burying the gold *Jeb Bush* article at #13. ICR ranks *Miss Beazley* (a presidential pet dog) at #1 by surface-matching "Laura Bush," compounded by lost-in-the-middle positional bias. **HeadRank** instead concentrates attention on the relational chain "second (74) → son (82) → of (46)," promoting the gold document to #1 (NDCG@10: 1.0 vs. BM25 0.08, ICR 0.63). Crucially, the selected heads track *dependency structure* rather than surface tokens: "second son of" encodes a familial relation that unambiguously identifies parentage—something neither keyword matching nor unaligned attention can resolve. Additional case studies covering seven benchmarks are in Appendix G, with multi-hop examples in Appendix G.2.

5 Conclusion

We presented **HeadRank**, which lifts preference optimization into continuous attention space to overcome score homogenization in decoding-free rerankers. Trained on only 211 queries, HeadRank achieves the best average ranking quality across 14 benchmarks at three Qwen3 scales (0.6B–4B) with $\mathcal{O}(1)$-pass inference, outperforming both generative and attention-based baselines—including models twice its size on multi-hop reasoning tasks.

Limitations

Training-Time Efficiency. Although HeadRank achieves significant *inference* speedup through depth truncation, its *training* cost remains comparable to full-model DPO because gradients still propagate through all layers up to $l_{\max}$. Since all selected core heads reside within layers 11–16 (out of 28 for Qwen3-0.6B), a natural extension is to freeze layers beyond $l_{\max}$ during training, potentially reducing optimizer states and backward-pass FLOPs by over 40%. We leave this head-guided layer freezing strategy to future work.

Static Head Assignment. The core head set is fixed after iterative selection and shared across all queries at inference time. A query-adaptive head selection mechanism could yield further gains for heterogeneous query types but would introduce additional overhead.

References

Shijie Chen, Bernal Jimenez Gutierrez, and Yu Su. 2025. Attention in large language models yields efficient zero-shot re-rankers. In *The Thirteenth International Conference on Learning Representations*.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. Hipporag: Neurobiologically inspired long-term memory for large language models. In *Advances in Neural Information Processing Systems*, volume 37.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Xueguang Ma, Liang Liang, Nan Wang, Hao Yang, and Jimmy Lin. 2024. Fine-tuning LLaMA for multi-stage text retrieval. In *Proceedings of the 47th International ACM SIGIR Conference*.

Jeongwoo Na, Jun Kwon, Eunseong Choi, and Jongwuk Lee. 2025. Multi-view-guided passage reranking with large language models. *arXiv preprint arXiv:2509.07485*.

Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. 2020. Document ranking with a pretrained sequence-to-sequence model. *arXiv preprint arXiv:2003.06713*.

Alexander Peysakhovich and Adam Lerer. 2023. Attention sorting combats recency bias in long context language models. *arXiv preprint arXiv:2310.01427*.

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.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36.

Stephen Robertson, Hugo Zaragoza, and 1 others. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3715–3734. Association for Computational Linguistics.

Sofia Serrano and Noah A Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951.

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. Association for Computational Linguistics.

Raphael Tang, Xinyu Zhang, Xueguang Ma, Jimmy Lin, and Ferhan Ture. 2023. Found in the middle: Permutation self-consistency improves listwise ranking in large language models. *arXiv preprint arXiv:2310.07712*.

Linh Tran, Yulong Li, Radu Florian, and Wei Sun. 2025. Contrastive retrieval heads improve attention-based re-ranking. *arXiv preprint arXiv:2510.02219*.

Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. 2021. TREC-COVID: Constructing a pandemic information retrieval test collection. *ACM SIGIR Forum*.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on EMNLP-IJCNLP*, pages 11–20.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. 2025. Retrieval head mechanistically explains long-context factuality. In *The Thirteenth International Conference on Learning Representations*.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2023. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Wang, Bowen Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Kayo Yin and Jacob Steinhardt. 2025. Which attention heads matter for in-context learning? In *Forty-Second International Conference on Machine Learning*.

Le Zhang, Bo Wang, Xipeng Qiu, Siva Reddy, and Aishwarya Agrawal. 2025a. REARANK: Reasoning re-ranking agent via reinforcement learning. *arXiv preprint arXiv:2505.20046*.

Wuwei Zhang, Fangcong Yin, Howard Yen, Danqi Chen, and Xi Ye. 2025b. Query-focused retrieval heads improve long-context reasoning and re-ranking. *arXiv preprint arXiv:2506.09944*.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pages 12697–12706. PMLR.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Zhicheng Dou, and Ji-Rong Wen. 2023. Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107*.

Honglei Zhuang, Zhen Qin, Kai Hui, Junru Wu, Le Yan, Xuanhui Wang, and Michael Bendersky. 2023a. Beyond yes and no: Improving zero-shot LLM rankers via scoring fine-grained relevance labels. *arXiv preprint arXiv:2310.14122*.

Shengyao Zhuang, Xueguang Ma, Bevan Koopman, Jimmy Lin, and Guido Zuccon. 2025. Rank-R1: Enhancing reasoning in LLM-based document rerankers via reinforcement learning. *arXiv preprint arXiv:2503.06034*.

Shengyao Zhuang, Honglei Zhuang, Bevan Kober, and Guido Zuccon. 2023b. Open-source large language models are strong zero-shot query likelihood models for document ranking. *arXiv preprint arXiv:2310.13243*.

Shengyao Zhuang, Honglei Zhuang, Bevan Kober, 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*.

Appendix

Table of Contents

A Inference Efficiency **11**

B Per-Dataset Radar Profiles **11**

C Cross-Dataset Universality and Scaling Behavior **11**

D Hyperparameter Details **11**

E Core Head Selection Details **12**

F RankGPT Prompt Template **12**

G Token-Level Attention Visualization **13**

G.1 Additional Token-Level Attention Visualizations 13

G.2 Multi-Hop Reasoning Attention Visualizations 15

H Algorithm Pseudocode **16**

A Inference Efficiency

Figure 5 plots inference latency against NDCG@10 across all methods and model scales. **HeadRank** occupies the Pareto frontier at every latency tier: it delivers the highest NDCG@10 among all compared methods while avoiding the autoregressive decoding overhead of RankGPT. Depth truncation at layer $l_{\max}$ (Figure 3) enables early exit after processing only $l_{\max}/L$ of the model's layers, yielding substantial wall-clock inference speedups that scale with model size.

B Per-Dataset Radar Profiles

Figure 6 presents per-dataset radar charts of normalised middle-zone standard deviation at 0.6B, 1.7B, and 4B scales. HeadRank traces the outermost polygon across all eight axes at both scales, confirming that the variance gains observed in the heatmaps (Section 4.5) are consistent across ad-hoc retrieval (DL19, DL20, COVID, NQ, News) and multi-hop reasoning benchmarks (HotpotQA, 2WikiMultiHop, MuSiQue).

C Cross-Dataset Universality and Scaling Behavior

Having established in Section 4.5 that HeadRank resolves middle-zone homogenization, a natural

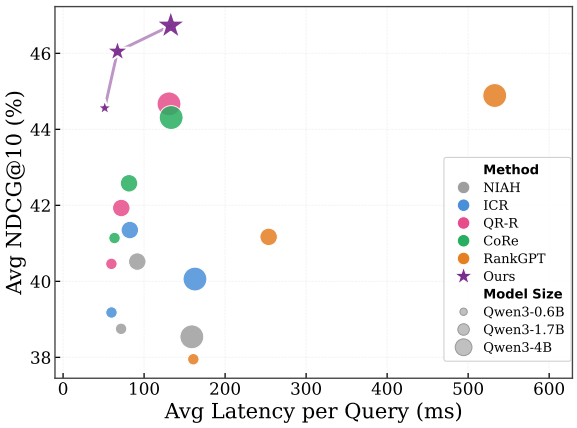
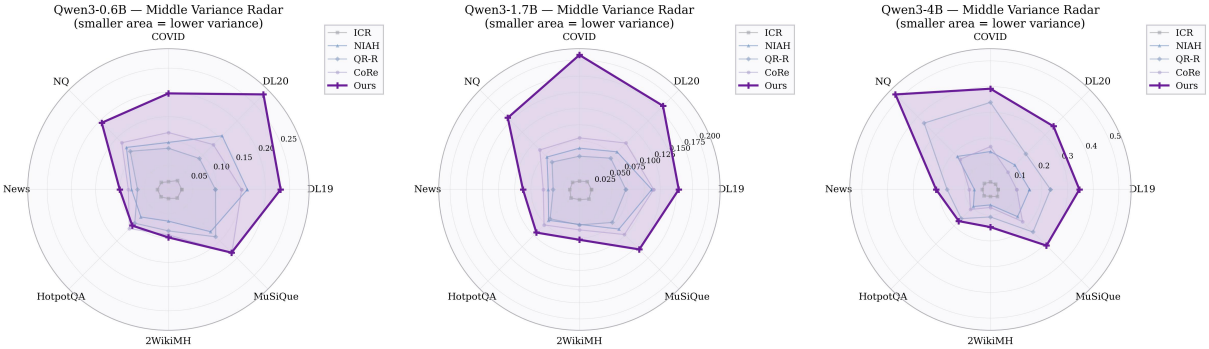


Figure 5: **Inference latency vs. NDCG@10** across all methods and model scales. **HeadRank** occupies the Pareto frontier, delivering the highest NDCG@10 at each latency tier without generative decoding overhead.

follow-up is whether this advantage generalizes beyond any single benchmark and how it interacts with model capacity. Per-dataset radar profiles (Figure 6) confirm that discriminability gains are not dataset-specific artifacts: at both 0.6B and 4B, HeadRank traces the outermost polygon across all eight radar axes—spanning both ad-hoc retrieval and multi-hop reasoning—while ICR collapses to the innermost ring and no baseline achieves comparable cross-dataset coverage. Figure 7 further reveals that HeadRank's middle-zone variance increases monotonically with model size on six of eight datasets, whereas baselines plateau or regress on multi-hop collections. This scaling trajectory suggests that larger checkpoints surface richer retrieval-head signal that G_{ent} can isolate with increasing precision, enabling the $\mathrm{Var}(\boldsymbol{s}_{\text{mid}})$ penalty to amplify progressively finer score differences as model capacity grows.

D Hyperparameter Details

All experiments are conducted on $8 \times$ NVIDIA H20 GPUs (97,871 MiB VRAM each) using the Megatron-LM distributed training framework. Tensor parallelism is set to TP=4 for the 0.6B and 1.7B scales and TP=8 for the 4B scale, with a micro-batch size of 1 per GPU and a global batch size of 32. Scale-specific learning rates are 1×10^{-5} (0.6B), 5×10^{-6} (1.7B), and 2×10^{-6} (4B). The proximal regularisation coefficient is β=0.05, the contrastive loss weight is α=0.05, and attention scores are aggregated over K=8 core heads via summation. Gradient norms are clipped at 5.0, and all models are trained for a single epoch. Conver-



(a) Qwen3-0.6B (b) Qwen3-1.7B (c) Qwen3-4B

Figure 6: Per-dataset radar profiles of normalised middle-zone standard deviation. HeadRank consistently occupies the outermost polygon across all three model scales.

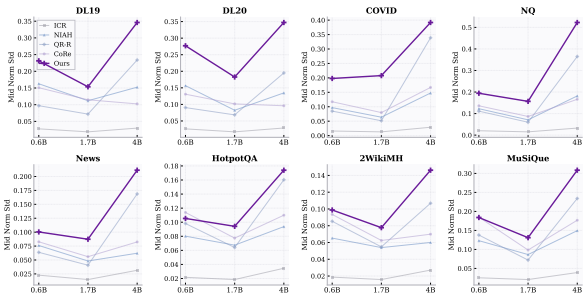


Figure 7: **Scaling behavior of middle-zone normalized std** (↑ better) as model capacity increases from 0.6B → 1.7B → 4B, reported per dataset (2×4 grid). **HeadRank** exhibits monotonic improvement on six of eight datasets and dominates all baselines at every scale.

gence is reached at approximately 1,800 steps for the 0.6B model, 2,200 steps for the 1.7B model, and 2,800 steps for the 4B model.

E Core Head Selection Details

Core retrieval heads are selected by the entropy-regularized scoring function (Eq. (4)) with λ=0.1 and τ=0.001. We apply the procedure *twice*: once on the pretrained base model (Before DPO) and once on the final aligned checkpoint (After DPO), selecting the top-K=8 heads each time. Table 5 reports all selected heads as (Layer, Head) pairs ranked by descending score; heads that appear in both sets are marked with ✓ in the Shared column.

Several patterns emerge from Table 5. **(i) DPO consistently shifts heads toward shallower layers.** Before alignment, selected heads span a wide range at every scale: layers 6–21 for 0.6B, 11–19 for 1.7B, and 20–24 for 4B. After DPO, each model *compresses* its head distribution into a tighter, shallower band—0.6B narrows from 6–21 to 11–16, 1.7B from 11–19 to 11–16 (dropping the L19 outlier and concentrating four of eight heads in layer 16), and 4B introduces L15-H11 while retaining five core heads in layers 21–24, effectively extending its depth floor downward by five layers. This uniform shallowing suggests that HeadRank discovers that lower-layer heads—closer to surface lexical features and with lower entropy—carry cleaner ranking signals once preference-aligned. **(ii) Scale-dependent depth bands persist after alignment.** Despite the consistent shallowing trend, each scale retains a characteristic operating depth: 0.6B settles in layers 11–16, 1.7B similarly in 11–16, and 4B in 15–24. The depth band shifts upward with model capacity, indicating that larger models encode retrieval signal in progressively deeper representations. **(iii) Stability increases with scale.** The shared-head ratio rises from 2/8 (0.6B) to 4/8 (1.7B) and 5/8 (4B), consistent with the hypothesis that larger models develop well-differentiated retrieval heads during pretraining, which DPO further sharpens without displacing.

F RankGPT Prompt Template

We present the zero-shot RankGPT prompt (Sun et al., 2023) used as the generative baseline throughout our evaluation. The prompt follows a multi-turn dialogue structure consisting of a system prompt, iterative per-passage user turns with assistant acknowledgements, and a closing post-prompt that requests the final ranked output.

System Prompt

You are RankGPT, an intelligent assistant that can rank passages based on their relevancy to the query. I will provide you with `[NUM]` pas-

Table 5: Top-8 core retrieval heads selected before and after HeadRank alignment for each Qwen3 scale, using entropy gating (λ=0.1, τ=0.001). Entries are sorted by post-DPO score. ✓ marks heads appearing in both sets (highlighted in purple).

Rank	Qwen3-0.6B Before	Qwen3-0.6B After	Qwen3-0.6B Shared	Qwen3-1.7B Before	Qwen3-1.7B After	Qwen3-1.7B Shared	Qwen3-4B Before	Qwen3-4B After	Qwen3-4B Shared
1	L11-H2	L14-H14		L14-H14	L11-H2	✓	L21-H8	L24-H29	✓
2	L11-H3	L11-H3	✓	L11-H2	L14-H6	✓	L21-H10	L21-H8	✓
3	L16-H7	L16-H13		L11-H3	L16-H8		L22-H4	L21-H10	✓
4	L21-H9	L11-H2	✓	L13-H14	L11-H3	✓	L23-H10	L23-H10	✓
5	L21-H8	L14-H6		L14-H6	L16-H7		L24-H29	L22-H4	✓
6	L6-H7	L13-H6		L16-H14	L16-H15		L21-H11	L23-H30	
7	L17-H12	L16-H8		L16-H13	L14-H15		L20-H16	L21-H18	
8	L13-H14	L16-H14		L19-H5	L16-H14	✓	L20-H15	L15-H11	
Shared	2 / 8			4 / 8			5 / 8		

sages, each indicated by number identifier []. Rank the passages based on their relevance to query: [QUERY].

Zero-shot User Instruction

Iterative User Message (per passage): [RANK] [Passage Content (truncated)]

Iterative Assistant Message (per passage): Received passage [RANK].

Post Prompt: Search Query: [QUERY]. Rank the [NUM] passages above based on their relevance to the search query. The passages should be listed in descending order using identifiers. The most relevant passages should be listed first. The output format should be [] > [], e.g., [1] > [2]. Only provide the ranking results, do not say any word or explain.

Sliding-window deployment. The prompt is deployed via a *sliding window* (size 20, stride 10) over the BM25 top-40 candidate list, sweeping from the bottom of the list upward so that each document is ranked at least twice. Per-window responses are parsed to extract integer identifiers; a window is counted as a formatting success if $\geq 50\%$ of the expected identifiers are recovered. The same prompt is reused without modification for the multihop (HotpotQA, 2Wiki, MuSiQue) and MLDR variants, with only the passage content and query language changing.

G Token-Level Attention Visualization

Tables 6–10 visualize the token-level attention distribution of **HeadRank** across seven diverse benchmarks. Each token is highlighted with a purple background whose intensity reflects its attention weight—deeper purple indicates higher attention. The BM25, ICR, and RankGPT columns show the rank each method assigns to the same passage for comparison.

Compositional entity retrieval (Table 6). The query *"famous computer scientists disappeared at sea"* requires matching both a profession and a fate. **HeadRank** places an extremely concentrated attention spike on the token *"sea"* in the Jim Gray article (the sole gold document), promoting it from BM25 rank #4 to #1 (NDCG@10 = 1.0). In contrast, RankGPT preserves the BM25 top-3 (generic list pages) unchanged (NDCG@10 = 0.43), while ICR ranks an irrelevant "Disappearance of Genette Tate" (a missing child case) at #1 (NDCG@10 = 0.63)—matching "disappeared" but ignoring "computer scientists".

Financial comparison reasoning (Table 7). For *"What are the tax benefits of dividends vs selling stock,"* attention concentrates on comparison-relevant tokens ("selling", "stock", "dividends"), enabling **HeadRank** to achieve perfect NDCG@10 = 1.0, versus ICR (0.31) and RankGPT (0.26).

G.1 Additional Token-Level Attention Visualizations

We present additional token-level attention visualizations across five diverse benchmarks, each illustrating a distinct failure mode that DPO-aligned attention overcomes.

Precise counting and entity disambiguation (Table 11). For the query *"how many sons robert kraft has,"* the gold passage states "They had **four**

Table 6: Token-level attention visualization for DBPedia. Query: “famous computer scientists disappeared at sea” Deeper purple indicates higher attention weight. BM25/ICR/RankGPT columns show the rank assigned by each method.

Rank	Passage (token-level attention coloring)	BM25	ICR	RankGPT	Rel.
#1	**Jim Gray (computer scientist)** Jim Gray (computer scientist) James Nicholas "Jim " (born January , ; lost at sea January , ; declared deceased May ,) was an American computer scientist who received the Turing Award in "for seminal contributions to database and transaction processing res	4	2	4	✓
#2	**Computer scientist** Computer scientist A computer is a scientist who has acquired knowledge of computer science , the study of the theoretical foundations of information and computation and their application.Computer scientists typically work on the theoretical side of computer systems , as oppo	5	5	5	–
#3	**Albert Lindsey Zobrist** Albert Lindsey Zobrist (born February ,) is an American computer scientist , games researcher , and inventor of the famous Zobrist Hashing , which was published in . He is further author of the first Go program in as part of his PhD Thesis on patte	10	19	10	–
#4	**Disappearance of Genette Tate** Disappearance of Genette Tate Louise (5 May – disappeared August) is an English girl whose disappearance became a famous missing person case when she went missing at age while delivering newspapers in Aylesbeare , Devon , England .	27	1	26	–
#5	**Douglas T. Ross** Douglas T . Ross Taylor "Doug " (21 December – January) was an American computer scientist pioneer , and Chairman of SofTech , Inc . He is most famous for originating the term CAD for computer-aided design , and is considered to be the father of Automatically Prog	11	13	11	–

Table 7: Token-level attention visualization for FiQA. Query: “What are the tax benefits of dividends vs selling stock” Deeper purple indicates higher attention weight. BM25/ICR/RankGPT columns show the rank assigned by each method.

Rank	Passage (token-level attention coloring)	BM25	ICR	RankGPT	Rel.
#1	"In the US , dividends are presently taxed at the same rates as capital gains , however selling stock could lead to less tax owed for the same amount of cash raised , because you are getting a return of basis or can elect to engage in a ""loss harvesting"" …	4	3	18	✓
#2	The benefit is not in taxes . When you sell a portion of your stock , you no longer have a portion of your stock . When you get a dividend , you still have a portion of your stock . Dividends are distributed from the net profits of a …	6	12	4	✓
#3	Not sure how this has got this far with no obvious discussion about the huge tax advantages of share buy backs vs dividend paying . Companies face a very simple choice with excess capital - pay to shareholders in the form of a taxable dividend , invest in future growth …	36	1	23	–
#4	Assuming a USA taxable account : Withdrawing funds from a brokerage account has nothing to do with taxes . Taxes are owed on the profit when you sell a stock , no matter what you do with the funds . Taxes are owed on any dividends the stock produces , …	18	36	14	–
#5	"They are similar in the sense that they are transferring money from the company to shareholders , but that’s about it . There is different tax treatment , yes , but that’s because they are fundamentally different . Dividends transfer money equally to all shareholders , but that also reduces …	5	4	3	–

sons:” followed by their names. Our attention peaks sharply at `four` (20)→`sons` (100)→`:` (52), precisely locating the numeric answer, then traces each son’s name with moderate weights (Jonathan, Daniel, Joshua Kraft). A confusing passage about a *different* Kraft family member (“two sons and a daughter”) also activates `sons` (51), but the model distinguishes it via the differing count. BM25 buries the gold passage at rank #24; our method promotes it to #1 (NDCG@10: Ours = 1.0 vs. ICR = 0.41).

Near-synonym disambiguation: *filmed* vs. *set* (Table 8). The query *“where is the show shameless filmed”* asks for the *production location* (Los Angeles), not the fictional setting (Chicago). In the gold passage, attention concentrates on the filming clause: `filmed` (67)→`in` (100)→`Los` (22)→`Angeles` (33), while the setting clause “set in Chicago” receives minimal weight (`set` = 24, `Chicago` = 10). Conversely, in the UK-version distractor (#4), attention peaks on `is` (100)→`set` (96)—the story setting rather than the filming location—correctly down-ranking it. ICR fails catastrophically, ranking gold passages at positions 39 and 40 (NDCG@10: Ours = 0.73 vs. ICR = 0.15).

Medical evidence assessment (Table 9). For the biomedical query about dexamethasone as a COVID-19 treatment, attention in the top-ranked passage concentrates on the conclusion:`dexamethasone` (45)→`reduce` (46)→`mortality` (51)→`COVID-19` (93)→`patients` (100)→`only` (51). The notable weight on the qualifier “`only`” indicates the model captures the *scope limitation* of the evidence (severe patients only), going beyond simple keyword matching (NDCG@10: Ours = 0.89 vs. ICR = 0.52).

Verb semantics in entity retrieval (Table 12). For *"Who created Wikipedia?,"* the #2 passage title contains "How A Bunch of Nobodies **Created** The World's Greatest Encyclopedia," where `Created` receives full attention weight (100), directly aligning with the query verb. In the #1 Wikipedia main page, attention shifts to the organizational creator: `Wikimedia`(39)→`Foundation`(96). BM25 ranks these gold documents at 33 and 14, overwhelmed by numerous language-edition pages that mention "Wikipedia" frequently (NDCG@10: Ours = 1.0 vs. RankGPT = 0.28).

Definition coverage (Table 10). For *"define visceral?,"* all top-5 passages are relevant, showcasing the model's ability to retrieve multiple valid definitions. Attention in #1 mirrors the query exactly: `Define`(100)→`visceral`(82)→`:`(79), while #3 activates the medical sense `viscera)`(100), demonstrating polysemy-aware coverage (NDCG@10: Ours = 1.0 vs. ICR = 0.82).

G.2 Multi-Hop Reasoning Attention Visualizations

Multi-hop questions require chaining evidence across multiple documents through bridge entities that may not appear in the query itself. We visualize six representative cases from HotpotQA, 2WikiMultiHopQA, and MuSiQue, where **HeadRank** achieves perfect or near-perfect Recall@2 while baseline methods largely fail. These examples reveal how the aligned attention distribution learns to *hop* between documents by shifting its focus from entity identification in one document to relationship tracing in another.

Shared-attribute discovery (Table 13). The query asks what occupation David Yates and Pietro Germi share. **HeadRank** (R@2 = 1.0) places both biographical documents in the top 2, with attention peaking on identity tokens *"Yates"*(100) and the sentence-final period in the Germi article (100), confirming each person's role after reading the full occupational description (*"filmmaker"* and *"director"*). Both ICR and RankGPT score R@2 = 0.0, instead surfacing *works about* Germi (e.g., *The Testimony*) rather than Germi's own biography—a failure to distinguish "about a person" from "mentioning a person's name."

Adjective attribution tracing (Table 14). To answer which author "Ortonesque" refers to, the model must trace the adjective back to playwright Joe Orton. **HeadRank** (R@2 = 1.0) ranks both Orton's and Henry James's biographies at the top, with attention concentrating on *"Orton"*(100) and *"playwright"*(43) in the Orton article. ICR (R@2 = 0.0) ranks Orton's biography at position 14, instead promoting his plays (*The Ruffian on the Stair*, rank 1) and a biographical film (*Prick Up Your Ears*, rank 2)—documents that mention "Joe Orton" frequently but describe works, not the person.

Creator–work chain resolution (Table 15). This two-hop question requires identifying the composer of the 1926 film *Camille* (William Axt) and then locating where Axt died. Attention in the film document peaks on *"film"*(100), while in the composer's biography it shifts to *"scores"*(100) and *"Born"*(49)—demonstrating a clear focal-point switch between hops. **HeadRank** achieves R@2 = 1.0, whereas both ICR and RankGPT manage only 0.5.

Genealogical reasoning (Table 16). To find Marianus V of Arborea's mother, the model must first locate his father (Brancaleone Doria) and then identify the father's wife (Eleanor of Arborea). In the Brancaleone Doria article, *"husband"*(100) receives the highest attention—directly encoding the family-relationship chain "X's husband's wife = X's mother." ICR ranks this critical bridging document at position 15, failing to capture the genealogical inference.

Hidden bridge-entity discovery (Table 17). The query "When was Lady Godiva's birthplace abolished?" is complicated by the strong BM25 distractor *Lady Godiva Rides Again* (a 1951 film, BM25 rank 1). **HeadRank**'s attention in the Spalding Priory document peaks sharply on *"Godiva"*(94) and *"Countess"*(100)—the historical figure's formal title that bridges to Mercia—while the film document receives high attention only on the surface match *"Godiva"*(100) with near-zero scores elsewhere, correctly deprioritizing it.

Cross-lingual entity relay (Table 18). Finding the birth date of the performer of the Swedish song *Till dom ensamma* (Mauro Scocco) is a challenging case: Scocco's biography article does not contain the song title and is ranked last by both BM25 (rank 20) and ICR (rank 20). **HeadRank** promotes it to rank 2 (R@2 = 1.0), achieving an 18-position leap. Attention in the song article concentrates on the Swedish title tokens *"Till"*(65),

Table 8: Token-level attention visualization for DL20. Query: “where is the show shameless filmed” Deeper purple indicates higher attention weight. BM25/ICR/RankGPT columns show the rank assigned by each method.

Rank	Passage (token-level attention coloring)	BM25	ICR	RankGPT	Rel.
#1	Shameless (U.S . TV series) Shameless is an American comedy-drama television series which airs on Showtime . This remake of the British series is set in Chicago , although filmed in Los Angeles , with the exterior scenes shot in Chicago.[1	18	39	4	✓
#2	(See Shameless UK TV Series). The show first aired on Showtime on January , , and is currently renewed for season which is set to premiere in January , . Although filmed in Los Angeles , Shameless is set in ChicagoÂĢÂĻs Canaryville neighborhood on the South Side .	7	11	3	✓
#3	Shameless is an American television comedy-drama which airs on Showtime . This remake of the award-winning British series is set in Chicago ’s Canaryville neighborhood on the South Side , although filmed in Los Angeles , with the exterior scenes filmed in Chicago .	40	40	31	✓
#4	Shameless goes Stateside : The new U.S . version of the UK favourite show premieres tonight on the Showtime cable channel . The original : The UK version of Shameless is set in the gritty Chatsworth estate . The role was played by Anne-Marie Duff for two series in the …	31	12	6	–
#5	ÂĢÂĺShamelessÂĢÂĻ filming in Chicago next week , Extras needed . According to a new casting call , Shameless will return to Chicago next week for a five day shoot . The show is primarily filmed in Los Angeles but they spend a week or two each season filming .. According …	4	4	14	✓

Table 9: Token-level attention visualization for COVID. Query: “what evidence is there for dexamethasone as a treatment for COVID-19?” Deeper purple indicates higher attention weight. BM25/ICR/RankGPT columns show the rank assigned by each method.

Rank	Passage (token-level attention coloring)	BM25	ICR	RankGPT	Rel.
#1	**Dexamethasone for COVID-19? Not so fast** Dexamethasone for COVID-19? Not so fast Recent announcements indicated without sharing any distinct published set of results , that the corticosteroid dexamethasone may reduce mortality of severe COVID-19 patients only . The recent Coronavirus [severe acute respiratory syndrome (SARS)	16	1	7	✓
#2	**Effect of Dexamethasone in Hospitalized Patients with COVID-19: Preliminary Report** Effect of Dexamethasone in Hospitalized Patients with COVID-19 : Preliminary Report Background : Coronavirus disease (COVID-19 is associated diffuse lung damage . Corticosteroids may modulate immune-mediated lung injury and reducing progression to respiratory failure and deat	5	5	3	✓
#3	**Dexamethasone for COVID-19? Not so fast.** Dexamethasone for COVID-19? Not so fast . Recent announcements indicated without sharing any distinct published set of results , that the corticosteroid dexamethasone may reduce mortality of severe COVID-19 patients only . The recent Coronavirus [severe acute respiratory syndrome (SARS	17	8	8	✓
#4	**Short-Term Corticosteroids in SARS-CoV2 Patients: Hospitalists’ Perspective** Short-Term Corticosteroids in SARS-CoV2 Patients : Hospitalists ’ Perspective Background : Dexamethasone a synthetic glucocorticoid has anti-inflammatory and immunosuppressive properties . There is a hyperinflammatory response involved in the clinical course of patients with pneumonia	12	29	18	–
#5	**Short-Term Dexamethasone in Sars-CoV-2 Patients.** Short-Term Dexamethasone in Sars-CoV-2 Patients . BACKGROUND a synthetic glucocorticoid has anti-inflammatory and immunosuppressive properties . There is a hyperinflammatory response involved in the clinical course of patients with pneumonia due to SARS-CoV-2 . To date ,	7	25	4	–

“ensamma”(52) for matching, and on *“by”*(32) for linking to the performer. In the biography, attention peaks on the sentence-final period (100) after scanning the full identity description, with secondary activation on *“Swedish”*(53) and *“described”*(62).

H Algorithm Pseudocode

Table 10: Token-level attention visualization for DL19. Query: “define visceral?” Deeper purple indicates higher attention weight. BM25/ICR/RankGPT columns show the rank assigned by each method.

Rank	Passage (token-level attention coloring)	BM25	ICR	RankGPT	Rel.
#1	Define visceral : felt in or as if in the internal organs of the body : deep ; not intellectual : instinctive , unreasoning ÂĢÂĶ visceral in a sentence	2	5	2	✓
#2	Definition of visceral „ : felt in or as if in the internal organs of the body : deep a visceral conviction . : not intellectual : instinctive , unreasoning visceral drives . : dealing with crude or elemental emotions : earthy a visceral novel . : of , relating . . .	22	19	8	✓
#3	The reason is that there are two different kinds of fat in your belly and visceral fat is only one of them . In a health setting , the word visceral means in or near your vital organs (your viscera). These are the organs deep in your gut , like . . .	21	9	7	✓
#4	Definition of 'visceral'. visceral (vÃīÂªsÃīÂĻrÃīÂĻl) Visceral feelings are feelings that you feel very deeply and find it difficult to control or ignore , and that are not the result of thought . I never overcame a visceral antipathy for the monarchy . ...the sheer visceral joy of being alive . . .	18	2	5	✓
#5	Definition of visceral „ : felt in or as if in the internal organs of the body : deep a visceral conviction . : not intellectual : instinctive , unreasoning visceral drives . : dealing with crude or elemental emotions : earthy a visceral novel . : of , relating . . .	23	20	29	✓

Table 11: Token-level attention visualization for DL20. Query: “how many sons robert kraft has” Deeper purple indicates higher attention weight. BM25/ICR/RankGPT columns show the rank assigned by each method.

Rank	Passage (token-level attention coloring)	BM25	ICR	RankGPT	Rel.
#1	They had four sons : Jonathan A . Kraft , born March , , is president of The Kraft Group and the New England Patriots „ Daniel A . Kraft is president of International Forest Products founded in by his father „ Joshua Kraft is president and CEO of the . . .	24	4	18	✓
#2	Kraft also served as the owner/investor of the San Jose Earthquakes from 1999ÂĢÂ¡2000 , the two years which the Kraft Group owned the team . Personal life . He has children , two sons and a daughter „ In , Kraft married Patricia Lipoma in a Jewish ceremony at the . . .	19	8	17	–
#3	They had four sons : Jonathan A . Kraft , born March , , is president of The Kraft Group and the New England Patriots „ Daniel A . Kraft is president of International Forest Products founded in by his father „ Joshua Kraft is president and CEO of the . . .	26	19	28	✓
#4	No Robert Kraft the Owner of the Patriots is not related in any way to Kraft foods . Patriots owner made his money from manufacturing , he started International Forest Products : ÂĢÂ¦ owns mills , manufactures and distributes paper and packaging products in countries „ Kraft Foods was formed . . .	36	13	7	–
#5	As the New England Patriots head to Super Bowl LI , owner Robert Kraft has a net worth of $5 .2 billion . He says the key is ambition As the New England Patriots head to Super Bowl LI , owner Robert Kraft has a net worth of $5 .2 . . .	4	2	11	–

Table 12: Token-level attention visualization for DBPedia. Query: “Who created Wikipedia?” Deeper purple indicates higher attention weight. BM25/ICR/RankGPT columns show the rank assigned by each method.

Rank	Passage (token-level attention coloring)	BM25	ICR	RankGPT	Rel.
#1	**Wikipedia** Wikipedia is a free-access , free-content Internet encyclopedia , supported and hosted by the non-profit Wikimedia Foundation „ Those who can access the site can edit most of its articles . Wikipedia is ranked among the ten mos	33	14	23	✓
#2	**The Wikipedia Revolution** The Wikipedia Revolution : How A Bunch of Nobodies Created The World's Greatest Encyclopedia is a popular history book by new media researcher and writer Andrew Lih.At the time of its publication it was "the only narrative account " of the online encyclope	14	1	14	✓
#3	**Swedish Wikipedia** Swedish Wikipedia The (Swedish : svenska Wikipedia , also svenskspr̊kiga Wikipedia) is the Swedish language edition of Wikipedia . Started in May , it is the fourth-oldest edition of Wikipedia , after the English Wikipedia „ German Wikipedia , and Catalan Wikipedia .	35	32	24	–
#4	**Vietnamese Wikipedia** Vietnamese Wikipedia The (Vietnamese : Wikipedia) is the Vietnamese-language edition of Wikipedia , a free , publicly editable , online encyclopedia supported by the Wikimedia Foundation „ As with other language editions of Wikipedia , the project's content	5	8	5	–
#5	**Wikipedia Review** Wikipedia Review is an Internet forum and blog for the discussion of Wikimedia projects , in particular the content and conflicts of Wikipedia . An InformationWeek Grok on Google blog described Wikipedia Review as "one of a number of Wikipedia watchdog " websites , "dedi	40	40	29	–

Table 13: Token-level attention visualization for HotpotQA. Query: “What occupation was shared by David Yates and Pietro Germi?” Deeper purple indicates higher attention weight. BM25/ICR/RankGPT columns show the rank assigned by each method.

Rank	Passage (token-level attention coloring)	BM25	ICR	RankGPT	Rel.
#1	**David Yates** David Yates (born (1963–)08) is an English filmmaker who has directed feature films , short films , and television productions .	2	4	7	✓
#2	**Pietro Germi** Pietro Germi ; September – December) was an Italian actor , screenwriter , and director . Germi was born in Genoa , Liguria , to a lower-middle-class family . He was a messenger and briefly attended nautical school before deciding on a career in acting .	1	7	6	✓
#3	**Ottavio Alessi** Ottavio Alessi Born in Cammarata Province of Agrigento , Alessi entered the film industry in as an assistant director . In he started an intense career as a screenwriter , alternating between genre films and art films and collaborating with Pietro Germi , Franco Rossi , Folco Q	3	1	5	–
#4	**The Testimony (1946 film)** The Testimony (1946 film) (Italian :Il testimone is Italian crime film directed by Pietro Germi and starring Roldano Lupi , Marina Berti and Ernesto Almirante . The film was made at the Cines Studios in Rome . It is one of several films regarded as an antecedent of th	5	2	1	–
#5	**Guido Coen** Guido Coen (1915–2010 was an Italian-born British film producer and film subtitler . He and his family were interned in Douglas on the Isle of Man during the Second World War . He began his career working for Filippo Del Giudice and Two Cities Films . When Two Cities was …	6	3	17	–

Table 14: Token-level attention visualization for HotpotQA. Query: “Between Joe Orton and Henry James, which author is the adjective "Ortonesque" used to refer to his work?” Deeper purple indicates higher attention weight. BM25/ICR/RankGPT columns show the rank assigned by each method.

Rank	Passage (token-level attention coloring)	BM25	ICR	RankGPT	Rel.
#1	**Henry James** Henry James OM ((1843–)15 – (1916–)28) was an American author regarded as a key transitional figure between literary realism and literary modernism , and is considered by many to be among the greatest novelists in the English language . He was the son of Henry	2	3	7	✓
#2	**Joe Orton** Joe Orton John Kingsley "Joe " Orton (1 January – August) was an English playwright and author . His public career was short but prolific , lasting from until his death three years later . During this brief period he shocked , outraged , and amused audiences with his …	1	14	6	✓
#3	**What the Butler Saw (play)** What the Butler Saw (play) is a farce written by the English playwright Joe Orton . It was premiered at the Queen’s Theatre in London on March . It was Orton’s final play and the second to be performed after his death , following "Funeral Games " in .	4	5	2	–
#4	**The Ruffian on the Stair** The Ruffian on the Stair On is a play by British playwright Joe Orton which was first broadcast on BBC Radio in August . It is an unsympathetic yet comedic one-act portrayal of working class England , as played out by a couple and a mysterious young man who t	3	1	1	–
#5	**Prick Up Your Ears** Prick Up Your Ears is a British film , directed by Stephen Frears , about the playwright Joe Orton and his lover Kenneth Halliwell . The screenplay was written by Alan Bennett , based on the biography by John Lahr . The film stars Gary Oldman as Orton , Alfred …	6	2	3	–

Table 15: Token-level attention visualization for 2WikiMultiHopQA. Query: “Where did the composer of film Camille (1926 Feature Film) die?” Deeper purple indicates higher attention weight. BM25/ICR/RankGPT columns show the rank assigned by each method.

Rank	Passage (token-level attention coloring)	BM25	ICR	RankGPT	Rel.
#1	**Camille (1926 feature film)** Camille (1926 feature film) is a American silent film based on the play adaptation of "La Dame aux CameliasThe Lady of the Camellias") by Alexandre Dumas , "fils", first published in French as a novel in and as a play in . Adapted by Fred de Gresac , George …	1	1	11	✓
#2	**William Axt** William Axt (April – February ,) was an American composer of nearly two hundred film scores . Born in New York City , Axt graduated from DeWitt Clinton High School in The Bronx and studied at the National Conservatory of Music of America . He served as an ass	2	5	2	✓
#3	**Marcel Varnel** Marcel Varnel (16 October – July) was a film director . He was born Marcel Hyacinthe le Bozec in Paris , France . Varnel started his working life on the Paris stage , soon becoming a director of musical comedies . In he moved to New York City working …	8	11	10	–
#4	**Philippe Arthuys** Philippe Arthuys November – January) was a French composer and film director . He worked on over films between and . His film " The Glass Cage " was entered into the 4th Moscow International Film Festival .	5	3	16	–
#5	**Fred Raymond** Fred Raymond aka Raimund Friedrich Vesely (20 April – January) was an Austrian composer . Raymond , born in Vienna , was the third child (after two daughters) of Vinzenz Vesely , an employee of the Austrian state railway system , and his wife Henriette , nee Dluho	17	15	5	–

Table 16: Token-level attention visualization for 2WikiMultiHopQA. Query: “Who is Marianus V Of Arborea’s mother?” Deeper purple indicates higher attention weight. BM25/ICR/RankGPT columns show the rank assigned by each method.

Rank	Passage (token-level attention coloring)	BM25	ICR	RankGPT	Rel.
#1	**Marianus V of Arborea** Marianus V of Arborea (1378 or – was the Judge of Arborea from until his death . His surname was Doria , but since he belonged to the ruling house of Arborea he is often dynastically called Bas-Serra , or Doria-Bas . Younger brother and successor of Frederick ,	1	1	3	✓
#2	**Brancaleone Doria** Brancaleone Doria was the husband of Eleanor of Arborea . He was a scion of an influential family (the Doria) of the Republic of Genoa , the son of the elder Brancaleone and a woman named Giacomina . On March , he became a vassal of Peter IV of …	7	15	2	✓
#3	**William II of Narbonne** William II of Narbonne was Viscount (1397-) and the nominal Judge of Arborea (1407- 1420). He was the grandson of Beatrice , youngest daughter of Marianus IV of Arborea and Timbra de RocabertÃŃ , and Aimery VI of Narbonne (married 1363). When Marianus V , the	2	2	1	–
#4	**Marianus I of Arborea** Marianus I of Arborea known as Mariano de Zori , was an early Judge of Arborea . The exact date of his reign is unknown . Francisco de Vico , followed by , placed it in – without any documentary evidence . Giovanni Francesco Fara , after analysing the documents …	3	3	20	–
#5	**Comita II of Arborea** Comita II of Arborea or III)(died) was the " giudice " of the Giudicato of Arborea , from until his death . He was the son of Constantine I of Arborea , first ruler of Arborea of the Lacon dynasty . Married Elena de Orrubu , mother of …	5	5	6	–

Table 17: Token-level attention visualization for MuSiQue. Query: “When was Lady Godiva’s birthplace abolished?” Deeper purple indicates higher attention weight. BM25/ICR/RankGPT columns show the rank assigned by each method.

Rank	Passage (token-level attention coloring)	BM25	ICR	RankGPT	Rel.
#1	**Spalding Priory** Spalding Priory It was founded as a cell of Croyland Abbey , in , by Leofric , Earl of Mercia and his wife , Godiva , Countess of Leicester . It was supported by Leofric’s eldest son , ÃĨlfgĂgr , Earl of Mercia and the monks were confirmed in their …	2	1	1	✓
#2	**Mercia** Mercia When ÃĨthelfld died in , ÃĨlfwynn , her daughter by ÃĨthelred , succeeded as ’Second Lady of the Mercians’, but within six months Edward had deprived her of all authority in Mercia and taken her into Wessex .	14	5	6	✓
#3	**Lady Godiva Rides Again** Lady Godiva Rides Again The film is most notable for the presence of actresses who were later to become famous . Diana Dors , who appears as a beauty queen , was later marketed as the film’s star . It also features Joan Collins in her film debut as an …	1	2	7	–
#4	**Kavangoland** Kavangoland like other homelands in South West Africa , was abolished in May at the start of the transition to independence .	4	4	2	–
#5	**Ruth Williams Khama** Ruth Williams Khama Lady Khama (9 December – May) was the wife of Botswana’s first president Sir Seretse Khama , the Paramount Chief of its Bamangwato tribe . She served as the inaugural First Lady of Botswana from to .	16	17	14	–

Table 18: Token-level attention visualization for MuSiQue. Query: “What is the Till dom ensamma performer’s birth date?” Deeper purple indicates higher attention weight. BM25/ICR/RankGPT columns show the rank assigned by each method.

Rank	Passage (token-level attention coloring)	BM25	ICR	RankGPT	Rel.
#1	**Till dom ensamma** Till dom ensamma is a song written by Mauro Scocco , and recorded by himself on the album Dr . Space dagbok , and released as a single the same year .	1	1	2	✓
#2	**Mauro Scocco** Mauro Scocco (born September) is a Swedish pop artist of Italian descent . He has been described as "one of the sharpest songwriters in Sweden". Scocco was the singer for the pop group Ratata (1980–83) transformed into a duo with Johan Ekelund (1983–89). After Rat	20	20	3	✓
#3	**Elias Tillandz** Elias Tillandz (1640–1693), born "Tillander", was a Swedish born doctor and botanist in Finland . He was the professor of medicine at the Academy of Turku . He wrote the country’s first botanical work , the "Catalogus Plantarum", which was first published in . As a do	19	18	1	–
#4	**Ecce Ancilla Domini** Ecce Ancilla Domini (Latin : "Behold the handmaiden of the Lord"), or The Annunciation , is an oil painting by the English artist Dante Gabriel Rossetti , first painted in and now in Tate Britain in London . The Latin title is a quotation from the Vulgate text of	2	3	20	–
#5	**Giovanni Cifolelli** Giovanni Cifolelli was an Italian mandolin virtuoso and dramatic composer whose date and place of birth are unknown . In he made his appearance in Paris as a mandolin virtuoso and was highly esteemed , both as a performer and teacher . He published his "Method fo	14	7	13	–

Algorithm 1 HeadRank Training and Inference

Require: Query set $\mathcal{Q}$; BM25 top-N candidates per query; pre-trained LLM θ_0
Ensure: Ranking permutation π for each query

Phase 1: ALPS Data Construction
for each query $q \in \mathcal{Q}$ **do**
  Retrieve top-100 documents via BM25 ▷ Training: top-100
  Construct preference pairs (d_w, d_l) using only *adjacent* relevance levels
end for

Phase 2: Initial Core Head Selection
for each head (l, h) in the Transformer **do**
  Compute discriminability $S_{\text{disc}}^{(l,h)}$ and entropy gate $G_{\text{ent}}^{(l,h)}$ using the constructed data
  Compute selection score $\Phi(l, h) = S_{\text{disc}}^{(l,h)} \cdot G_{\text{ent}}^{(l,h)}$ ▷ Eq. (4)
end for
Select top-K heads by Φ; set $l_{\max} \leftarrow \max\{l : (l, h) \in \text{top-}K\}$

Phase 3: HeadRank Training
Freeze reference model $\theta_{\text{ref}} \leftarrow \theta_0$
for each mini-batch of preference pairs **do**
  Extract attention scores $s_\theta(d)$ from core heads for d_w, d_l
  Compute $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{align}}(\Delta s_\theta) + \mathcal{L}_{\text{prox}}(\boldsymbol{\Delta}_{\text{ref}}) + \Omega(\boldsymbol{s}_\theta)$
  Update θ via gradient descent on $\mathcal{L}_{\text{total}}$
end for

Phase 4: Iterative Head Recalibration
Re-run **Phase 2** on updated θ to recalibrate the core head set and $l_{\max}$

Phase 5: Inference (Early-Exit)
for each test query q **do**
  Run forward pass truncated at layer $l_{\max}$ ▷ Inference: top-40
  Extract attention scores from core heads; rank by $s_\theta(d)$
end for
return ranking permutation π