ReAttn: Improving Attention-based Re-ranking via Attention Re-weighting

Yuxing Tian[1], **Fengran Mo**[1], **Weixu Zhang**[2], **Yiyan Qi**[3], **Jian-Yun Nie**[1]
[1]University of Montreal; [2]McGill University & MILA;
[3]International Digital Economy Academy

Abstract

The strong capabilities of recent Large Language Models (LLMs) have made them highly effective for zero-shot re-ranking task. Attention-based re-ranking methods, which derive relevance scores directly from attention weights, offer an efficient and interpretable alternative to generation-based re-ranking methods. However, they still face two major limitations. First, attention signals are highly concentrated a small subset of tokens within a few documents, making others indistinguishable. Second, attention often overemphasizes phrases lexically similar to the query, yielding biased rankings that irrelevant documents with mere lexical resemblance are regarded as relevant. In this paper, we propose **ReAttn**, a post-hoc re-weighting strategy for attention-based re-ranking methods. It first compute the cross-document IDF weighting to down-weight attention on query-overlapping tokens that frequently appear across the candidate documents, reducing lexical bias and emphasizing distinctive terms. It then employs entropy-based regularization to mitigate over-concentrated attention, encouraging a more balanced distribution across informative tokens. Both adjustments operate directly on existing attention weights without additional training or supervision. Extensive experiments demonstrate the effectiveness of our method.

1 Introduction

Information retrieval (IR) serves as a cornerstone of modern intelligent systems, powering applications in search, recommendation, and retrieval-augmented generation (Lewis et al., 2020). A typical IR pipeline adopts a two-stage paradigm in which an initial retriever gathers a broad set of candidate documents using sparse (Robertson and Zaragoza, 2009) or dense retriever (Karpukhin et al., 2020), and a subsequent re-ranker refines these candidate documents to produce the final ranked list (Nogueira et al., 2019, 2020). The re-ranking stage is particularly critical, as it governs retrieval precision and shapes the contextual information accessible to downstream reasoning or generation models (Gao et al., 2023).

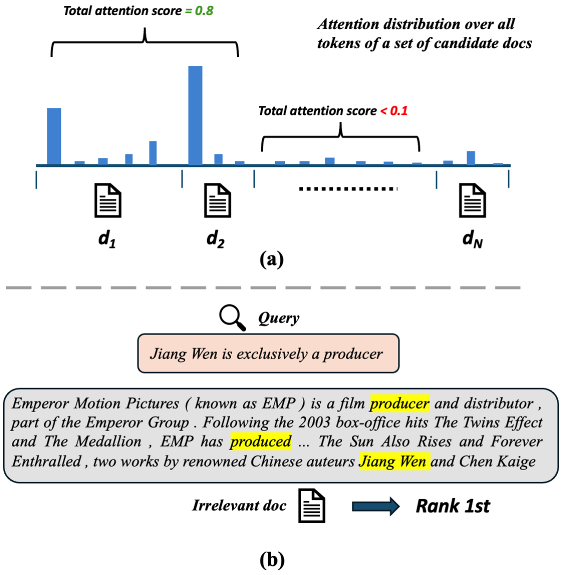
Total attention score = 0.8
Attention distribution over all tokens of a set of candidate docs
Total attention score < 0.1
d_1
d_2
d_N
(a)
Query
Jiang Wen is exclusively a producer
Emperor Motion Pictures (known as EMP) is a film producer and distributor , part of the Emperor Group . Following the 2003 box-office hits The Twins Effect and The Medallion , EMP has produced ... The Sun Also Rises and Forever Enthralled , two works by renowned Chinese auteurs Jiang Wen and Chen Kaige
Irrelevant doc
Rank 1st
(b)

Figure 1: Illustration of two issues in attention-based re-ranking. (a) Signal concentration: the total attention mass is heavily concentrated on a few tokens of a few documents, leaving most candidates with negligible attention scores. (b) Lexical bias: attention disproportionately highlights tokens that are lexically similar to the query (e.g., producer, Jiang Wen), causing irrelevant documents with lexical overlap to the query to receive inflated relevance scores and misleading the ranking.

Traditional re-ranking methods have predominantly relied on supervised learning with cross-encoder architectures, which require substantial labeled data and task-specific fine-tuning. The advent of large language models (LLMs) has profoundly transformed this landscape. Owing to their strong zero-shot reasoning capabilities and ability to comprehend long contexts, LLMs can effectively perform zero-shot re-ranking task (Sachan

Findings of the Association for Computational Linguistics: EACL 2026, pages 1282–1295
March 24-29, 2026 ©2026 Association for Computational Linguistics

et al., 2022; Sun et al., 2023a; Qin et al., 2024). Early attempts to apply LLMs to re-ranking can be viewed as generation-based approaches (Sun et al., 2023b), as they directly prompt the model to generate an ordered list of candidate documents identifiers, effectively framing re-ranking as a text generation task. Although these approaches exploit the generative capabilities of LLMs to reason about document relevance, they also introduce several practical challenges: high computational overhead from autoregressive decoding; strong sensitivity to prompt design; and low instruction-following reliability, which frequently results in inconsistencies or omissions in the generated outputs. Recent research has increasingly focused on non-generative re-ranking approaches that derive relevance score of documents directly from internal signals within LLMs. Chen et al. (2025) has shown that the attention signal in LLMs implicitly reflect relevance between queries and documents: query tokens tend to allocate greater attention weights to tokens in relevant documents. Specifically, they first aggregate attention weights that tokens in each document receive from all query tokens across all layers and heads of the LLM to obtain the token-level relevance scores. These scores are then summed within each document to produce a document-level relevance score used for ranking. Zhang et al. (2025c) further identify specialized query-focused retrieval heads that attend selectively to information most relevant to the query. Compared with Chen et al. (2025), which aggregates attention weights across all heads, using only these specialized heads obtains more reliable relevance score and leads to better re-ranking performance.

Despite the promising performance of attention-based re-ranking methods, two inherent issues constrain its effectiveness. The first is signal concentration (Figure 1(a)): a small number of tokens within a few documents absorb most of the total attention mass, while most documents receive minimal attention contributions. This imbalance limits the re-ranker's ability to distinguish and rank the less salient documents. The second is lexical bias (Figure 1(b)), where attention tends to overemphasize tokens that are lexically similar to the query. As a consequence, documents containing such overlapping expressions, even when semantically irrelevant, can receive high attention scores. While attention mechanisms are capable of modeling semantic relevance, this lexical preference can partially constrain their effectiveness, leading to suboptimal ranking of truly relevant documents.

To this end, we propose **ReAttn**, a post-hoc adjustment strategy for attention-based re-ranking that re-weighting token- and document-level attention score without additional training. We first introduce a cross-document inverse document frequency (IDF) weighting to mitigate lexical bias. Intuitively, if a query token appears across most documents, it provides little evidence for distinguishing truly relevant documents and dilutes the discriminative strength of the attention signal. Thus, we measure the frequency of each query token appear across the candidate set and compute corresponding cross-document IDF weights. These weights are then used to rescale token-level attention scores, reducing the influence of ubiquitous tokens and emphasizing distinctive ones that better indicate relevance. Next, we introduce an entropy-based regularization. For each document, we normalize its token-level relevance scores into a probability distribution and compute its entropy as a measure of attention dispersion. Documents with higher entropy reflect broader coverage across informative tokens, whereas lower entropy signals an overly narrow focus. Then we apply a document-level weighting based on this measure, decreasing the relevance scores of low entropy documents and increasing those of high entropy documents. This yields a more balanced allocation of attention across documents and improves discrimination among moderately relevant documents.

Our contributions are summarized as follows:

- We identify two core issues that limit the effectiveness of attention-based re-ranking methods: signal concentration and lexical bias.
- We propose ReAttn, which introduces a cross-document IDF weighting to reduce lexical bias and an entropy-based regularization to alleviate score concentration, jointly improving attention reliability and ranking stability.
- Extensive experiments on multiple datasets of re-ranking and long context reasoning task demonstrate that ReAttn consistently improves ranking performance over attention-based re-ranking baselines with minimal computational overhead.

2 Related Work

Zero-Shot Re-Ranking with LLMs. Zero-shot re-ranking with LLMs broadly categorized into three primary paradigms: point-wise, pair-wise, and list-wise approaches. In point-wise re-ranking, each candidate document is evaluated independently, typically using generation probabilities or scalar relevance scores derived from the LLM's logits (Sachan et al., 2022; Liang et al., 2022). Although computationally efficient, these methods often fall short in performance because they fail to capture interactions or relative preferences among documents. Pair-wise methods (Qin et al., 2024) address this limitation by comparing document pairs to infer preference relations, which are then aggregated to produce the final ranking. While this pairwise comparison generally enhances discriminative accuracy, it also introduces a quadratic computational cost with respect to the number of candidate documents. In contrast, list-wise approaches (Ma et al., 2023; Sun et al., 2023a; Chen et al., 2025) process the entire candidate set simultaneously, enabling the model to capture holistic dependencies and optimize ranking quality at the list level. The main challenge of list-wise modeling lies in its requirement for extended context handling. However, recent advances in long-context LLM architectures (Chen et al., 2023; Jin et al., 2024a; Fu et al., 2024; Mo et al., 2025) have made such modeling increasingly practical. As a result, list-wise re-ranking has emerged as both a scalable and highly effective strategy. Building on this trend, our work focuses on attention-based list-wise re-ranking (Chen et al., 2025), which combines the efficiency of non-generative architectures with the stability of attention-driven scoring, outperforming generation-based re-ranking methods in both computational and empirical robustness (Ma et al., 2023; Sun et al., 2023a; Mo et al., 2023).

Attention Mechanisms. Attention mechanisms form the foundation of transformer-based language models, facilitating token-to-token information exchange and offering a potential lens for interpreting model behavior. Since the seminal work of Bahdanau et al. (2015), attention weights have often been employed as proxies for token-level importance, providing insights into how models distribute focus during inference. Despite ongoing debate regarding their faithfulness as explanations of internal reasoning (Serrano and Smith, 2019; Wiegreffe and Pinter, 2019; Zhang et al., 2024, 2025a,b; Mo et al., 2026), attention patterns have demonstrated considerable utility in practice. For example, Izacard and Grave (2021) exploit cross-attention between queries and passages to enhance retriever training, while Peysakhovich and Lerer (2023) leverage attention-based document sorting to improve text generation quality. Building on this intuition, in-context re-ranking (ICR) (Chen et al., 2025) interprets attention as an implicit signal of document relevance, enabling LLMs to reorder retrieved passages directly within their context windows, without the need for additional supervision or parameter updates.

Concurrently, research in mechanistic interpretability has examined the functional specialization and redundancy of individual attention heads in transformers. Early investigations demonstrated that many attention heads contribute little to overall performance, with only a small subset being functionally critical (Michel et al., 2019; Voita et al., 2019). Subsequent analyses have identified distinct behavioral roles, such as induction heads that capture recurring patterns across sequences (Olsson et al., 2022; Yin and Steinhardt, 2025; Ren et al., 2024), and others that help manage knowledge conflicts (Shi et al., 2024; Jin et al., 2024b) or mitigate contextual distractions (Zhu et al., 2025). More recently, attention-based retrieval behavior has emerged as a focal point of study: Wu et al. (2025b) identify retrieval heads that copy relevant answer tokens from long contexts, while Zhang et al. (2025c) introduce query-focused retrieval heads (QR heads) that model fine-grained interactions between queries and supporting evidence.

3 Preliminary: Attention-Based Re-ranking with LLMs

Formally, let a query be denoted as Q and its corresponding set of retrieved candidate documents as $D = \{d_1, d_2, \ldots, d_N\}$. The objective of re-ranking is to assign each document $d_i \in D$ a relevance score with respect to the query Q and a final ranking sorted in descending order of relevance score. In attention-based re-ranking, we first construct the sequence $X = [d_1, d_2, ..., d_N, Q]$ by concatenating all candidate documents followed by the query, and provide it as input to a LLM with L transformer layers and H attention heads. For each document d_i, the token-level relevance score of its j-th token, denoted as $s_{d_{i,j},Q}$, is obtained by aggregating the attention weights that this token receives from all query tokens across all layers and attention heads:

$$s_{d_{i,j},Q} = \frac{1}{|\mathcal{I}_Q|} \sum_{l=1}^{L} \sum_{h=1}^{H} \sum_{k \in \mathcal{I}_Q} a_{j,k}^{l,h} \tag{1}$$

where $\mathcal{I}_Q$ denotes the set of query tokens, and $a_{j,k}^{l,h}$ represents the attention weight from the k-th token (in the query) to the j-th token (in document d_i) by the h-th attention head at l-th layer. Summing token-level scores within a document yields the relevance score for document d_i.

However, raw attention do not always provide a faithful measure of relevance. Prior work has shown that LLMs exhibit intrinsic biases in attention allocation (Gallegos et al., 2024),including disproportionate weighting toward longer documents and meaningless tokens (e.g., punctuation). These intrinsic biases can undermine the reliability of attention-based relevance computation, especially when attention is directly used as a ranking signal. Following Zhao et al. (2021), ICR (Chen et al., 2025) use *"N/A"* as the calibration query Q_c and calculate calibration scores $\{s_{d_{i,j},Q_{cal}}\}$ via attention aggregation for each document. The calibration scores capture strong attention weights biased towards meaningless tokens in the documents, such as punctuation, which should not affect the relevance of the documents. Then we can subtract it from ranking scores from the actual query $\{s_{d_{i,j},Q}\}$ to obtain the calibrated ranking score $\{s_{d_{i,j}}\}$:

$$s_{d_{i,j}} = s_{d_{i,j},Q} - s_{d_{i,j},Q_c}. \tag{2}$$

After that, we filter out such tokens with abnormally negative calibrated scores. Finally, we sum the calibrated scores for all tokens in each document to obtain the final ranking score s_{d_i}, which measures the change of attention weights that each document receives when the query changes from the content-free calibration query to the actual query:

$$\begin{aligned} \mathcal{S}_{d_i} &= \{d_{i,j}\} \\ \mathcal{S}_{d_i}^* &= \{d_{i,j} | d_{i,j} > \bar{\mathcal{S}}_{d_i} - 2\sigma_{\mathcal{S}_{d_i}}\} \\ s_{d_i} &= \sum_{s \in \mathcal{S}_{d_i}^*} s \end{aligned} \tag{3}$$

where σ denotes standard deviation.

4 ReAttn

We propose **ReAttn**, a post-hoc re-weighting strategy for attention-based re-ranking that re-weights both token-level and document-level attention scores without any additional training. ReAttn aims to address two key issues: (*i*) *lexical bias*, where tokens overlapping with the query receive disproportionate attention, and (*ii*) *signal concentration*, where only top-ranked documents receive strong signals, leaving lower-ranked documents poorly differentiated. The method consists of two sequential steps: cross-document IDF reweighting at the token level and entropy-based regularization at the document level. We detail each component below.

4.1 Cross-Document IDF Weighting

The first step mitigates lexical bias by downweighting contributions from query tokens that frequently appear across the candidate set. Intuitively, if a query token occurs frequently across candidate documents, it provides limited discriminative information and can weaken the attention-based relevance estimation.

Formally, for each token $t_{i,j}$ in document d_i that also appears in the query Q (i.e., $t_{i,j} \in \mathcal{I}_Q$), we compute its document frequency over the candidate document set $D = \{d_1, \ldots, d_N\}$:

$$\mathrm{df}(t_{i,j}) = \big|\{\, d_k \in D \mid t_{i,j} \in d_k \,\}\big|. \tag{4}$$

We then define a normalized inverse document frequency (IDF) weight:

$$w(t_{i,j}) = \frac{\log \frac{N+1}{\mathrm{df}(t_{i,j})+1}}{\log(N+1)} \tag{5}$$

where N is the total number of documents. Tokens that appear in many documents receive smaller $w(t_{i,j})$, while distinctive tokens retain larger weights. Then the token-level calibrated attention scores $s_{d_{i,j}}$ (from Section 3) are adjusted as:

$$\tilde{s}_{d_{i,j}} = \begin{cases} w(t_{i,j}) \cdot s_{d_{i,j}} & \text{if } t_{i,j} \in \mathcal{I}_Q \\ s_{d_{i,j}} & \text{otherwise.} \end{cases} \tag{6}$$

This re-weighting reduces the dominance of ubiquitous query tokens while preserving the influence of distinctive terms. The document-level base score after IDF re-weighting is obtained by summing token scores within the filtered token set $\mathcal{S}_{d_i}^*$:

$$B_i = \sum_{j \in \mathcal{S}_{d_i}^*} \tilde{s}_{d_{i,j}}. \tag{7}$$

B_i serves as the IDF-adjusted relevance score for document d_i, emphasizing tokens that provide stronger discriminative evidence of relevance.

4.2 Entropy-Based Regularization

Although IDF weighting alleviates lexical bias, it does not prevent signal concentration, where a few tokens within a few documents dominate the relevance estimation and most documents receive negligible scores. To improve the score dispersion across documents, we introduce an entropy-based regularization that quantifies how evenly attention is distributed within each document. Specifically, for each document d_i, we normalize its token-level scores into a probability distribution:

$$p_{i,j} = \frac{\tilde{s}_{d_i,j}}{B_i}, \quad \text{where} \sum_{j \in \mathcal{S}^*_{d_i}} p_{i,j} = 1. \tag{8}$$

The normalized Shannon entropy of this distribution measures the internal dispersion of attention:

$$E_i = -\frac{\sum_{j \in \mathcal{S}^*_{d_i}} p_{i,j} \log p_{i,j}}{\log |\mathcal{S}^*_{d_i}|} \tag{9}$$

where $E_i \in [0, 1]$. High-entropy documents have attention spread across multiple informative tokens, while low-entropy documents exhibit overly narrow attention focus. We compute the base-score-weighted mean entropy across all documents:

$$\bar{E}_B = \frac{\sum_{k=1}^{N} B_k E_k}{\sum_{k=1}^{N} B_k}. \tag{10}$$

Each document is then assigned a dispersion weight according to its deviation from the average entropy:

$$W_i = 1 + (E_i - \bar{E}_B). \tag{11}$$

Documents with broader attention coverage ($E_i > \bar{E}_B$) receive a small positive adjustment, whereas narrowly focused ones ($E_i < \bar{E}_B$) are slightly penalized. The adjusted document-level score is:

$$s'_i = B_i \cdot W_i. \tag{12}$$

Finally, to maintain comparability across documents, we normalize the scores so that they sum to one:

$$s^{\text{final}}_{d_i} = \frac{s'_i}{\sum_{k=1}^{N} s'_k}. \tag{13}$$

This entropy-based regularization ensures that attention-based re-ranking captures both strong token-level relevance and balanced document-level coverage. Combined with IDF re-weighting, ReAttn effectively mitigates lexical bias and enhances score discrimination, yielding more robust and fine-grained re-ranking performance.

5 Experiments

5.1 Experimental Setup

Datasets. Following (Zhang et al., 2025c), We evaluate our method on two tasks: re-ranking and long-context reasoning. For re-ranking task, we experiment on eleven public datasets in the BEIR benchmark (Thakur et al., 2021) consisting of diverse domains, including NQ (Kwiatkowski et al., 2019), COVID (Voorhees et al., 2021), NFCorpus (Boteva et al., 2016), FiQA (Maia et al., 2018), SciFact (Wadden et al., 2020), SciDocs (Cohan et al., 2020), FEVER (Thorne et al., 2018), Climate (Diggelmann et al., 2020), DBPedia (Hasibi et al., 2017), Robust04 (Jeronymo et al., 2022). For long-context reasoning task, we use LongMemEval (Wu et al., 2025a) and CLIPPER (Pham et al., 2025).

Baselines. We compare our approach with several existing re-ranking methods that leverage LLMs, including: (1) RankGPT (Sun et al., 2023b): a generation-based re-ranking approach that prompts LLMs to generate the ranking order list of a candidate document set conditioned on the query. We evaluate two variants of RankGPT: **RankGPT$^{\text{w/o}}$**, which feeds all candidate documents into the model prompt simultaneously without sliding window, and **RankGPT$^{\text{Bubble}}$**, which employs a bubble-sort strategy to iteratively rank smaller subsets of documents. (2) **ICR** (Chen et al., 2025): an attention-based re-ranking approach that aggregates the attention signals across all heads and layers in LLMs to compute document relevance scores. (3) **QRhead** (Zhang et al., 2025c): an attention-based re-ranking approach. Instead of aggregating attention weights from all heads and layers, QRhead first identify specialized attention heads that attend selectively to information most relevant to the query on a labeled dataset. Only the attention signals from these selected heads are used to compute relevance scores.

We also include several representative traditional retrievers as baselines, including Contriever, GTR-T5-base, BGE-Reranker-base, MSMARCO-MiniLM and Stella.

Base LLMs. As attention-based re-ranking method requires access to the attention weights across all layers and heads of an LLM, we conduct experiments using open-weight LLMs. Specifically, we select three widely used instruction-tuned LLMs from two model families of different scales, including Llama-3.2 (3B), Llama-3.1 (8B) from the

	NQ	COVID	NFCorpus	FiQA	Scifact	Scidocs	FEVER	Climate	DBPedia	Robust04	News	Avg
BM25	30.5	59.5	32.2	23.6	67.9	14.9	65.1	16.5	31.8	40.7	39.5	38.4
						Base LLM: Llama-3.2-3B-Instruct						
RankGPT$^{\text{w/o}}$	30.0	59.5	32.2	23.6	67.9	14.9	65.9	17.1	31.8	40.7	39.5	38.5
RankGPT$^{\text{Bubble}}$	33.2	61.8	32.0	22.4	66.1	14.8	65.8	17.1	34.8	40.5	40.2	39.0
ICR	49.2	72.3	33.8	31.8	73.3	17.4	82.6	24.2	34.7	47.2	44.7	46.5
ICR+ReAttn	49.8	72.5	34.4	32.5	74.1	17.5	83.0	24.5	36.8	47.7	45.4	47.3
QRhead	54.9	77.4	35.1	35.1	74.7	18.3	83.7	24.5	36	49.7	45.1	48.6
QRhead+ReAttn	**55.6**	**77.6**	**36.6**	**36.2**	**75.3**	**19.5**	**84.3**	**25.6**	**38**	**49.8**	**46**	**49.3**
						Base LLM: Llama-3.1-8B-Instruct						
RankGPT$^{\text{w/o}}$	30.0	59.5	32.2	23.6	67.9	14.9	65.9	16.8	31.8	40.7	39.5	38.4
RankGPT$^{\text{Bubble}}$	53.7	75.5	34.3	31.4	69.3	17.4	67.5	23.8	**42.9**	47.8	46.2	46.3
ICR	54.0	73.3	34.8	35.6	75.5	19.0	85.8	24.8	36.9	49.0	44.5	48.5
ICR+ReAttn	54.4	73.5	35	35.9	76.2	19.3	86.1	**25.7**	37.5	49.3	45.1	49.3
QRhead	58.6	77.5	35.3	39.1	76.2	19.4	85.3	23.9	37.2	51.4	46.2	50.0
QRhead+ReAttn	**59.3**	**78.2**	**36.0**	**40.4**	**76.8**	**20.2**	**86.4**	25.3	38.6	**52.5**	**47**	**50.9**
						Base LLM: Qwen-2.5-7B-Instruct						
RankGPT$^{\text{w/o}}$	30.0	59.5	32.2	23.6	67.9	14.9	65.9	16.8	31.8	40.7	39.5	38.4
RankGPT$^{\text{Bubble}}$	42.7	**70.5**	34.1	29.5	69.3	16.6	70.5	19.7	37.1	**46.4**	**43.6**	43.6
ICR	43.1	66.1	32.7	27.0	71.1	16.4	79.2	19.6	35.3	43.0	40.0	43.0
ICR+ReAttn	44.0	67.0	33.9	28.1	**72.2**	**17.8**	80.0	20.9	36.6	43.6	41.2	44.0
QRhead	49.9	67.7	33.1	29.2	71	15.3	80.7	20.1	35.7	43.7	39.8	44.2
QRhead+ReAttn	**50.7**	68.6	**34.2**	**30.3**	71.9	16.6	**81.5**	**21.3**	**37.4**	44.8	41.0	**45.3**
						Traditional retrievers						
Contriever	44.6	67.5	32.8	28.4	67.1	18.9	64.2	28.0	39.5	45.7	41.7	43.5
GTR-T5-base	51.4	74.8	32.5	34.7	62.1	15.8	72.9	26.8	37.1	46.1	42.8	45.2
BGE-Reranker-base	55.2	66.4	31.0	31.7	70.8	15.7	88.6	36.5	42.5	39.9	37.0	46.8
MSMARCO-MiniLM	55.8	74.3	35.2	35.1	68.5	17.5	80.4	25.5	45.3	47.9	43.0	48.0

Table 1: Performance comparison across different base LLMs and re-ranking methods on BEIR benchmarks evaluated by nDCG@10. **Bold** indicates the best results.

Llama family, and Qwen2.5 (7B) from the Qwen family.

Setting. Our setting largely follows prior works (Chen et al., 2025; Zhang et al., 2025c), we re-rank the top 200 documents returned by BM25(Robertson and Zaragoza, 2009) and report nDCG@10. We sub-sampled 512 random questions for each domain for evaluation. We apply a sliding window of size 20 and stride 10 for RankGPT. For ICR and QRhead, we directly re-rank all documents.

5.2 Experiment results

Re-ranking Task. Table 1 reports the results on the BEIR benchmark. It demonstrate that ReAttn consistently improves over the baseline attention-based re-rankers, confirming the effectiveness of refining intrinsic attention signals through IDF and entropy regularization. Across all settings, adding ReAttn to ICR or QRhead leads to steady performance gains. Specifically, under Llama-3.2-3B-Instruct, ICR+ReAttn improves the average nDCG@10 from 46.5 to 47.3, and QRhead+ReAttn further lifts it from 48.6 to 49.3. A similar trend holds for Llama-3.1-8B-Instruct, where ICR+ReAttn improves by +0.8 points (48.5 → 49.3), and QRhead+ReAttn achieves the best average score of 50.9. For Qwen-2.5-7B-Instruct, improvements are smaller but consistent (+1.0 for ICR and +1.1 for QRhead). These consistent gains across distinct model families and parameter scales indicate that ReAttn generalizes well, even when underlying LLM architectures differ in tokenizer or attention mechanisms. Notably, ReAttn brings slightly larger relative improvements on smaller model (e.g. 3B). This is because smaller LLMs often exhibit sharper attention distributions and are therefore more susceptible to attention collapse. The entropy-based regularization in ReAttn helps redistribute attention weights, allowing mid-ranked documents to receive more informative gradients and improving ranking metrics such as nDCG@10, which depend on accurate ordering beyond the top few results.

The improvements are most prominent on datasets with high lexical overlap or entity-centric

	LongMemEval				**Clipper**			
RETRIEVER	RETRIEVAL RECALL@K		END-TO-END PERFORMANCE		RETRIEVAL RECALL@K		END-TO-END PERFORMANCE	
	k = 5	k = 10	Top-5	Top-10	k = 3	k = 5	Top-3	Top-5
Base LLM: Llama-3.2-3B-Instruct								
Full context	-	-	28.1		-	-	25.2	
BM25	57.5	67.5	46.1	44.9	74.6	83.7	20.0	22.8
Contriever	62.7	79.2	**48.6**	46.5	60.2	78.9	12.6	18.4
Stella	63.9	77.6	44.9	47.7	83.3	90.0	21.3	25.1
RankGPT$^{\text{w/o}}$	1.8	3.4	23.5	23.3	16.8	27.3	3.6	8.8
RankGPT$^{\text{Bubble}}$	2.1	3.8	24.0	24.4	17.0	27.4	3.8	8.8
ICR	68.7	78.8	46.5	45.1	72.8	83.6	19.4	23.6
ICR+ReAttn	68.8	79.2	46.6	45.5	72.9	83.8	19.4	23.8
QRhead	77.6	86.6	47.4	47.7	85.5	93.4	23.4	26.9
QRhead+ReAttn	**77.9**	**87.3**	47.6	**48**	**85.7**	**93.5**	**23.4**	**27.2**
Base LLM: Llama-3.1-8B-Instruct								
Full context	-	-	46.5		-	-	31.3	
BM25	57.5	67.5	48.8	50.9	74.6	83.7	37.9	37.9
Contriever	62.7	79.2	52.6	55.4	60.2	78.9	28.2	31.1
Stella	63.9	77.6	50.9	58.4	83.3	90.0	38.8	39.6
RankGPT$^{\text{w/o}}$	2.1	4.0	26.7	24.2	30.0	39.4	15.9	19.4
RankGPT$^{\text{Bubble}}$	8.3	9.0	28.1	27.0	36.7	44.3	19.7	20.4
ICR	77.0	84.4	59.3	56.1	89.3	94.7	43.8	42.5
ICR+ReAttn	77.1	85.0	59.4	56.3	89.4	94.9	43.9	**42.8**
QRhead	85.5	91.7	59.8	60.2	93.8	96.9	47.6	41.9
QRhead+ReAttn	**85.8**	**92.4**	**59.9**	**60.5**	**94**	**97.3**	**47.7**	42

Table 2: Performance comparison across different base LLMs and re-ranking methods on LongMemEval and Clipper. The base LLM denotes the LLM used for both the retriever and end-to-end generation.

structure, such as DBPedia, FiQA, and FEVER. For example, on DBPedia, ICR+ReAttn improves by +2.1 points under Llama-3B and +0.6 points under Llama-8B, indicating that IDF weighting effectively suppresses distractors sharing repeated entity tokens with the query. On FiQA, QRhead+ReAttn achieves +1.1 (Llama-3B) and +1.3 (Llama-8B), suggesting that entropy regularization improves ranking discrimination when relevant evidence is distributed across multiple tokens.

Although ReAttn operates without fine-tuning, its re-ranked results are competitive with supervised cross-encoders. For instance, QRhead+ReAttn (Llama-8B) achieves 50.9 average nDCG@10, surpassing MSMARCO-MiniLM (48.0) and GTR-T5-base (45.2). These results indicate that careful post-hoc refinement of intrinsic attention patterns can yield retrieval performance comparable to trained models, while remaining fully parameter-free. Overall, ReAttn consistently enhances the discrimination of attention-based re-rankers. The IDF weighting component mitigates the dominance of ubiquitous lexical overlaps, while entropy regularization improves ranking calibration among mid-ranked documents. Together, these mechanisms address the two main weaknesses of ICR—attention concentration and lexical bias—resulting in more balanced and semantically aligned re-ranking behavior across diverse retrieval tasks.

Long-Context Reasoning Task Table 2 reports results on the **LongMemEval** and **Clipper** benchmarks, which evaluate retrieval and reasoning over extended contexts. The results consistently demonstrate positive gains across both retrieval and end-to-end settings. Under Llama-3.2-3B and Llama-3.1-8B, ReAttn improves recall@k and slightly raises downstream reasoning accuracy for both ICR and QRhead. The trend is stable across datasets and model sizes, indicating that ReAttn's refinements to the attention signal generalize well to longer input sequences, where attention saturation and token redundancy become more severe. In long-context scenarios, re-rankers must identify sparse but semantically crucial evidence distributed across a large number of retrieved documents. This setting magnifies the limitations of unadjusted attention scores, which often overemphasize lexical overlaps near query tokens and underweight relevant evidence

	NQ	COVID	NFCorpus	FiQA	Scifact	Scidocs	FEVER	Climate	DBPedia	Robust04	News	Avg
Base LLM: Llama-3.2-3B-Instruct												
ICR	49.2	72.3	33.8	31.8	73.3	17.4	82.6	24.2	34.7	47.2	44.7	46.5
ICR+IDF only	49.6	72.3	34.2	32.3	73.8	17.3	82.7	24.2	36.0	47.3	44.9	46.8
ICR+Entropy only	49.4	72.4	33.9	32.0	73.6	17.4	82.9	24.4	35.5	47.6	45.3	46.8
ICR+ReAttn	49.8	72.5	34.4	32.5	74.1	17.5	83.0	24.5	36.8	47.7	45.4	47.3
QRhead	54.9	77.4	35.1	35.1	74.7	18.3	83.7	24.5	36.0	49.7	45.1	48.6
QRhead+IDF only	55.3	77.5	36.0	35.7	75.0	18.9	84.0	25.2	37.0	49.7	45.5	49.1
QRhead+Entropy only	55.2	77.5	35.6	35.5	75.1	19.0	84.1	25.0	36.8	49.8	45.7	49.0
QRhead+ReAttn	**55.6**	**77.6**	**36.6**	**36.2**	**75.3**	**19.5**	**84.3**	**25.6**	**38.0**	**49.8**	**46.0**	**49.3**
Base LLM: Qwen-2.5-7B-Instruct												
ICR	43.1	66.1	32.7	27.0	71.1	16.4	79.2	19.6	35.3	43.0	40.0	43.0
ICR+IDF only	43.8	66.4	33.5	27.4	71.7	17.2	79.5	20.1	36.2	43.1	40.4	43.6
ICR+Entropy only	43.5	66.7	33.1	27.7	71.5	16.9	79.7	20.4	35.8	43.4	40.7	43.6
ICR+ReAttn	44.0	67.0	33.9	28.1	**72.2**	**17.8**	80.0	20.9	36.6	43.6	41.2	44.0
QRhead	49.9	67.7	33.1	29.2	71.0	15.3	80.7	20.1	35.7	43.7	39.8	44.2
QRhead+IDF only	50.4	67.9	33.9	29.7	71.5	16.2	81.0	20.8	36.7	44.1	40.4	44.8
QRhead+Entropy only	50.2	68.3	33.5	29.9	71.3	15.8	81.3	20.9	36.3	44.4	40.6	44.8
QRhead+ReAttn	**50.7**	68.6	**34.2**	**30.3**	71.9	16.6	**81.5**	**21.3**	**37.4**	44.8	41.0	**45.3**

Table 3: Performance comparison across different base LLMs and re-ranking methods on BEIR benchmarks evaluated by nDCG@10. **Bold** indicates the best results.

occurring later in the sequence. The IDF-based re-weighting in ReAttn reduces this redundancy by down-weighting query tokens that repeatedly appear across candidate documents, preventing these common tokens from dominating the attention budget. Meanwhile, the entropy regularization ensures that attention is more evenly distributed across informative spans, rather than collapsing onto a few locally similar segments. Together, these refinements improve document-level coverage and facilitate the retrieval of complementary evidence required for multi-hop reasoning.

5.3 Ablation Study

We further conduct ablations experiments to disentangle the respective contributions of the cross-document IDF re-weighting and the intra-document entropy regularization in ReAttn. Table 3 demonstrates a systematic dependence on dataset characteristics. On lexically dominated benchmarks (e.g., NQ, NFCorpus, FiQA, DBPedia), the +IDF only variant recovers the majority of ReAttn's improvements. This is consistent with the fact that relevance in these collections is strongly associated with explicit keyword matching, where rerankers can be disproportionately influenced by surface-level overlap. By reweighting token contributions according to their distinctiveness across the retrieved set, IDF alone effectively suppresses spurious lexical matches and yields substantial ranking gains. In contrast, for corpora featuring longer documents and higher topical heterogeneity (e.g., Robust04, News, Climate), the +Entropy only variant provides comparatively larger benefits. In such settings, the primary failure mode extends beyond lexical bias: attention distributions within long passages often become excessively concentrated on a small number of tokens, which undermines evidence aggregation and degrades robustness. The entropy term counteracts this degeneracy by discouraging overly peaked attention and promoting broader allocation over informative regions, thereby improving coverage of salient content within each document. On Scidocs, applying IDF in isolation can slightly reduce effectiveness, suggesting that distinctiveness-based down-weighting may inadvertently penalize frequent yet domain-informative scientific terminology. Importantly, the full ReAttn (IDF + entropy) restores and surpasses performance, indicating that the two factors target non-overlapping failure modes and are synergistic rather than substitutable.

To further substantiate these findings, we perform a qualitative analysis on FEVER using ICR as the base re-ranker and visualize token-level attention maps (Table 4, 5, 6). Removing the IDF component leads to a pronounced shift in scoring behavior: for distractor documents that share surface tokens with the query, ReAttn w/o IDF assigns inflated attention mass to such overlapping terms, which increases document scores despite weak semantic support. By contrast, the full ReAttn attenuates these generic matches via cross-document IDF re-weighting, resulting in attention patterns that concentrate on contextually diagnostic phrases and evidence-bearing spans rather than repeated or high-frequency query tokens. Collectively, these results establish that IDF re-weighting is critical for decoupling lexical overlap from semantic relevance in token-level reranking. By modulating token contributions based on distinctiveness across candidates, ReAttn preserves sharper discrimination between relevant and irrelevant documents, particularly under hard-negative scenarios dominated by verbatim overlap. When combined with entropy regularization that mitigates attention collapse in long passages, the full ReAttn achieves consistently stronger and more stable ranking behavior than either component in isolation.

6 Conclusion

We presented ReAttn, a post-hoc re-weighting strategy for attention-based re-ranking with large language models. ReAttn mitigates signal concentration and lexical bias through cross-document IDF weighting and entropy-based regularization, leading to more reliable and balanced attention-based relevance estimation. Extensive experiments across different datasets and base LLMs confirm that ReAttn consistently improves performance of existing attention-based re-ranking methods.

7 Limitations

Our work has several limitations that suggest promising directions for future research. First, the evaluation of ReAttn is restricted to general-purpose large language models. Although such models provide a representative basis for zero-shot retrieval, a growing number of LLMs have been fine-tuned specifically for information retrieval. These models often exhibit distinct attention dynamics due to task-oriented optimization and domain-adaptive pretraining. Understanding how ReAttn interacts with such IR-specialized models would provide deeper insight into the generality of its attention refinement mechanism.

Second, ReAttn focuses on post-hoc reweighting of attention scores without explicitly examining the contribution of individual attention heads to relevance estimation. Prior work such as QR-head has initiated exploration in this direction, yet its approach to identifying retrieval-relevant heads remains heuristic and coarse-grained. A more systematic analysis of head-level relevance attribution could enhance both the interpretability and the theoretical grounding of attention-based reranking frameworks.

Third, the current experiments are limited to English-language datasets. Since language models may exhibit divergent attention behaviors across linguistic typologies, assessing the cross-lingual robustness of ReAttn is a critical next step. Extending the analysis to multilingual retrieval scenarios would not only test the stability of the proposed adjustments but also inform the design of more universal attention regularization strategies.

References

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

Vera Boteva, Demian Gholipour, Artem Sokolov, and Stefan Riezler. 2016. A full-text learning to rank dataset for medical information retrieval. In *Advances in Information Retrieval: 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20–23, 2016. Proceedings 38*, pages 716–722. Springer.

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*.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. 2023. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S Weld. 2020. Specter: Document-level representation learning using citation-informed transformers. *arXiv preprint arXiv:2004.07180*.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. Climate-fever: A dataset for verification of real-world climate claims. *arXiv preprint arXiv:2012.00614*.

Yao Fu, Rameswar Panda, Xinyao Niu, Xiang Yue, Hannaneh Hajishirzi, Yoon Kim, and Hao Peng. 2024. Data engineering for scaling language models to 128k context. *arXiv preprint arXiv:2402.10171*.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, pages 1–79.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1).

Faegheh Hasibi, Fedor Nikolaev, Chenyan Xiong, Krisztian Balog, Svein Erik Bratsberg, Alexander Kotov, and Jamie Callan. 2017. Dbpedia-entity v2: a test collection for entity search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1265–1268.

Gautier Izacard and Edouard Grave. 2021. Distilling knowledge from reader to retriever for question answering. In *International Conference on Learning Representations*.

Vitor Jeronymo, Mauricio Nascimento, Roberto Lotufo, and Rodrigo Nogueira. 2022. mrobust04: A multilingual version of the trec robust 2004 benchmark. *Preprint*, arXiv:2209.13738.

Hongye Jin, Xiaotian Han, Jingfeng Yang, Zhimeng Jiang, Zirui Liu, Chia-Yuan Chang, Huiyuan Chen, and Xia Hu. 2024a. Llm maybe longlm: Self-extend llm context window without tuning. *arXiv preprint arXiv:2401.01325*.

Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. 2024b. Cutting off the head ends the conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 1193–1215.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*.

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

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, and 1 others. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

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

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: Financial opinion mining and question answering. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 1941–1942, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? *Advances in neural information processing systems*, 32.

Fengran Mo, Yifan Gao, Chuan Meng, Xin Liu, Zhuofeng Wu, Kelong Mao, Zhengyang Wang, Pei Chen, Zheng Li, Xian Li, and 1 others. 2025. Uniconv: Unifying retrieval and response generation for large language models in conversations. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6936–6949.

Fengran Mo, Kelong Mao, Yutao Zhu, Yihong Wu, Kaiyu Huang, and Jian-Yun Nie. 2023. Convgqr: Generative query reformulation for conversational search. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4998–5012.

Fengran Mo, Zhan Su, Yuchen Hui, Jinghan Zhang, Jia Ao Sun, Zheyuan Liu, Chao Zhang, Tetsuya Sakai, and Jian-Yun Nie. 2026. Opendecoder: Open large language model decoding to incorporate document quality in rag. *arXiv preprint arXiv:2601.09028*.

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

Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy Lin. 2019. Multi-stage document ranking with bert. *arXiv preprint arXiv:1910.14424*.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, and 1 others. 2022. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*.

Alexander Peysakhovich and Adam Lerer. 2023. Attention sorting combats recency bias in long context language models. *ArXiv*, abs/2310.01427.

Chau Minh Pham, Yapei Chang, and Mohit Iyyer. 2025. Clipper: Compression enables long-context synthetic data generation. *Preprint*, arXiv:2502.14854.

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.

Jie Ren, Qipeng Guo, Hang Yan, Dongrui Liu, Quanshi Zhang, Xipeng Qiu, and Dahua Lin. 2024. Identifying semantic induction heads to understand in-context learning. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6916–6932.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Devendra Sachan, Mike Lewis, Mandar Joshi, Armen Aghajanyan, Wen-tau Yih, Joelle Pineau, and Luke Zettlemoyer. 2022. Improving passage retrieval with zero-shot question generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3781–3797.

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, Florence, Italy. Association for Computational Linguistics.

Dan Shi, Renren Jin, Tianhao Shen, Weilong Dong, Xinwei Wu, and Deyi Xiong. 2024. Ircan: Mitigating knowledge conflicts in llm generation via identifying and reweighting context-aware neurons. *Advances in Neural Information Processing Systems*, 37:4997–5024.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023a. 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.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023b. 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, Singapore. Association for Computational Linguistics.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808.

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. In *ACM SIGIR Forum*, volume 54, pages 1–12. ACM New York, NY, USA.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. 2025a. Longmemeval: Benchmarking chat assistants on long-term interactive memory. *Preprint*, arXiv:2410.10813.

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

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

Jinghan Zhang, Fengran Mo, Xiting Wang, and Kunpeng Liu. 2024. Blind spot navigation in llm reasoning with thought space explorer. *arXiv preprint arXiv:2410.24155*.

Jinghan Zhang, Xiting Wang, Fengran Mo, Yeyang Zhou, Wanfu Gao, and Kunpeng Liu. 2025a. Entropy-based exploration conduction for multi-step reasoning. *arXiv preprint arXiv:2503.15848*.

Jinghan Zhang, Xiting Wang, Weijieying Ren, Lu Jiang, Dongjie Wang, and Kunpeng Liu. 2025b. Ratt: A thought structure for coherent and correct llm reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26733–26741.

Wuwei Zhang, Fangcong Yin, Howard Yen, Danqi Chen, and Xi Ye. 2025c. 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.

Youxiang Zhu, Ruochen Li, Danqing Wang, Daniel Haehn, and Xiaohui Liang. 2025. Focus directions make your language models pay more attention to relevant contexts. *arXiv preprint arXiv:2503.23306*.

A Prompt Templates

We provide prompt templates used in our experiments for ICR,QRhead and our method in Figure 2 and rankGPT in Figure 3.

```
{prompt_prefix} Here are some paragraphs:

[1] {Title 1 (if available)}
{Paragraph text 1}

[2] {Title 2 (if available)}
{Paragraph text 2}

...

Please find information that is relevant
to the following query in the paragraphs
above.

Query: {query}{prompt_suffix}
```

Figure 2: Prompt used for ICR, QRhead and our method.

```
{prefix} This is an intelligent assistant
that can rank passages based on their
relevancy to the query.

The following are {N} passages, each
indicated by a numbered identifier [i]. I
can rank them based on their relevance to
the query: "{query}"

[1] {Title 1 (if available)}
{Paragraph text 1}

[2] {Title 2 (if available)}
{Paragraph text 2}

...

The search query is: "{query}". I will
rank the {N} passages above based on their
relevance to the search query. The passages
will be listed in descending order using
identifiers, the most relevant passages
should be listed first and the output format
should be [] > [] > etc, e.g., [1] > [2] >
etc. Be sure to list all {N} ranked passages
and do not explain your ranking until after
the list is done. {suffix} Ranked Passages:
[
```

Figure 3: Prompt used for rankGPT.

Table 4: Attention scores for all tokens in a FEVER example with the query “Night of the Living Dead was originally directed by Krzysztof Kieslowski.”

Method	Passage	Rank
REATTN W/O IDF	The Scar (Blizna) is a 1976 Polish film written and directed by Krzysztof Kieślowski and starring Franciszek Pieczka . Filmed on location in Olechów , Poland , the film is about a man put in charge of the construction ...	3
REATTN	The Scar (Blizna) is a 1976 Polish film written and directed by Krzysztof Kieślowski and starring Franciszek Pieczka . Filmed on location in Olechów , Poland , the film is about a man put in charge of the construction ...	6

Table 5: Attention scores for all tokens in a FEVER example with the query “Jiang Wen is exclusively a producer.”

Method	Passage	Rank
REATTN W/O IDF	Emperor Motion Pictures (known as EMP) is a film producer and distributor , part of the Emperor Group . Following the 2003 box-office hits The Twins Effect and The Medallion , EMP has produced ... The Sun Also Rises and Forever Enthralled , two works by renowned Chinese auteurs Jiang Wen and Chen Kaige. ...	2
REATTN	Emperor Motion Pictures (known as EMP) is a film producer and distributor , part of the Emperor Group . Following the 2003 box-office hits The Twins Effect and The Medallion , EMP has produced ... The Sun Also Rises and Forever Enthralled , two works by renowned Chinese auteurs Jiang Wen and Chen Kaige. ...	10

Table 6: Attention scores for all tokens in a DBPedia-Entity example with the query “Give me all launch pads operated by NASA.”

Method	Passage	Rank
REATTN W/O IDF	Cape Can av eral Air Force Station Launch Complex 19 Ċ Launch Complex 19 (LC - 19) is a deactivated launch site on Cape Can av eral Air Force Station , Florida used by NASA to launch all of the Gemini manned space fl ights . It was also used by unmanned Titan I and Titan II missiles .L C - 19 was in use from 195 9 to 196 6 , during which time it saw 27 launches , 10 of which were manned . The first use of LC - 19 was on August 14 , 195 9 . This was a Titan I and the mission was declared a failure after the rocket exploded while still on the pad .	12
REATTN	Cape Can av eral Air Force Station Launch Complex 19 Ċ Launch Complex 19 (LC - 19) is a deactivated launch site on Cape Can av eral Air Force Station , Florida used by NASA to launch all of the Gemini manned space fl ights . It was also used by unmanned Titan I and Titan II missiles .L C - 19 was in use from 195 9 to 196 6 , during which time it saw 27 launches , 10 of which were manned . The first use of LC - 19 was on August 14 , 195 9 . This was a Titan I and the mission was declared a failure after the rocket exploded while still on the pad .	18