CONTRASTIVE RETRIEVAL HEADS IMPROVE ATTENTION-BASED RE-RANKING

Linh Tran[1] **Yulong Li**[2] **Radu Florian**[2] **Wei Sun**[2]

[1]Rensselaer Polytechnic Institute [2] IBM Research

ABSTRACT

The strong zero-shot and long-context capabilities of recent Large Language Models (LLMs) have paved the way for highly effective re-ranking systems. Attention-based re-rankers leverage attention weights from transformer heads to produce relevance scores, but not all heads are created equally: many contribute noise and redundancy, thus limiting performance. To address this, we introduce *CoRe heads*, a small set of retrieval heads identified via a contrastive scoring metric that explicitly rewards high attention heads that correlate with relevant documents, while downplaying nodes with higher attention that correlate with irrelevant documents. This relative ranking criterion isolates the most discriminative heads for re-ranking and yields a state-of-the-art list-wise re-ranker. Extensive experiments with three LLMs show that aggregated signals from CoRe heads, constituting less than 1% of all heads, substantially improve re-ranking accuracy over strong baselines. We further find that CoRe heads are concentrated in middle layers, and pruning the computation of final 50% of model layers preserves accuracy while significantly reducing inference time and memory usage.

1 INTRODUCTION

Information retrieval systems form the backbone of search engines and retrieval-augmented generation (RAG). The order of retrieved passages not only determines search relevance but also shapes the context provided to downstream generation models (Lewis et al., 2020; Gao et al., 2023). Modern retrieval systems typically adopt a two-stage pipeline: a lightweight retriever, such as BM25 (Robertson et al., 2009) or dense retrievers (Karpukhin et al., 2020), first selects a candidate set of documents, which is then refined by a re-ranker using more powerful models. Recent advances in Large Language Models (LLMs) have transformed re-ranking, delivering strong zero-shot performance (Sachan et al., 2022; Sun et al., 2023; Qin et al., 2024; Chen et al., 2025).

Recent work has extended this trend by investigating *list-wise* re-rankers, which consider the entire candidate set simultaneously and jointly produce an ordering. By exploiting cross-document attention, such models capture relative preferences across candidates, leading to rankings that align more closely with global evaluation metrics. For example, Sun et al. (2023) introduce RankGPT, which casts re-ranking as a text generation problem: the LLM is prompted to output an ordered list of document identifiers. However, this generative formulation introduces several drawbacks. Autoregressive decoding adds linear computational overhead; outputs often contain incomplete or duplicate rankings; performance is unstable due to prompt sensitivity; and the approach underuses the relevance signals already present in attention patterns. To address these challenges, Chen et al. (2025) propose an attention-based re-ranker that directly aggregates attention scores across all heads to estimate document relevance. This approach eliminates the need for text generation, reduces runtime to a single forward pass, and achieves superior accuracy compared to RankGPT. However, aggregating uniformly across all attention heads may still be suboptimal, since many heads are either redundant or capture non-informative patterns (Michel et al., 2019; Voita et al., 2019).

To better characterize the functional role of attention heads, Wu et al. (2025) identify a subset of *retrieval heads* in transformers that specialize in extracting relevant information from long contexts. Their approach tracks copy–paste operations in Needle-in-a-Haystack tasks, which constrains its applicability to narrow question-answering scenarios. Building on this, Zhang et al. (2025) propose

QR heads (query-focused retrieval heads), selected according to the absolute attention allocated to the correct answer. While aggregated QR-head scores improve performance on downstream retrieval tasks, the selection criterion overlooks relative ranking. For instance, a head may assign high attention to the gold document yet allocate even greater weight to irrelevant documents, undermining its discriminative capacity. Since the QR-head method does not penalize such cases, it may fail to identify the heads that are most effective at distinguishing relevant from irrelevant information.

To this end, we introduce *CoRe heads* (Contrastive Retrieval heads), a subset of attention heads specialized for document re-ranking. Leveraging CoRe heads yields a state-of-the-art list-wise re-ranker that improves accuracy while simultaneously reducing computational and memory overhead, making the approach practical for real-world retrieval systems. Our contributions are threefold:

- We propose a contrastive scoring metric that identifies CoRe heads by rewarding attention directed towards relevant documents while penalizing attention to irrelevant ones. In contrast to prior retrieval-head methods that consider only absolute attention to a single document, our approach explicitly models relative ranking, resulting in stronger re-ranking accuracy.
- We demonstrate that CoRe heads identified using a small subset of the Natural Questions (NQ) dataset generalize across the BEIR benchmark (Thakur et al., 2021) and even across languages in the MLDR benchmark (Chen et al., 2024). Attention-based re-rankers using CoRe heads consistently outperform strong baselines, achieving state-of-the-art list-wise re-ranker performance.
- We observe that top-scoring CoRe heads are primarily concentrated in the middle transformers layers. Exploiting this insight, we show that pruning most final layers incurs negligible loss in re-ranking accuracy, while reducing memory usage by 40% and inference latency by 20%. This highlights the efficiency of CoRe heads for real-world retrieval systems.

2 Related Work

Zero-Shot Re-Ranking. Zero-shot re-ranking methods with LLMs can be broadly grouped into three main approaches: point-wise, pair-wise, and list-wise. Point-wise methods (Sachan et al., 2022; Liang et al., 2022) score each document independently, typically through logits or direct generation. These approaches are computationally efficient but often underperform due to the absence of cross-document comparison. Pair-wise methods (Qin et al., 2024) compare two documents at a time and aggregate over all pairs to form a ranking. While improving accuracy, the computational cost grows quadratically with the number of documents. List-wise methods (Ma et al., 2023; Sun et al., 2023; Chen et al., 2025) instead consider all documents jointly to produce a single ranked list. Although list-wise methods require long-context modeling, advances in LLMs with extended context windows (Chen et al., 2023; Jin et al., 2024a; Fu et al., 2024) have made this increasingly practical. Consequently, list-wise approaches now achieve both scalability and strong effectiveness. Our work follows this trajectory, focusing on attention-based list-wise re-ranking (Chen et al., 2025), which offers more efficiency and stability compared to generation-based methods (Ma et al., 2023; Sun et al., 2023).

Role of Attention Heads. A growing body of work in mechanistic interpretability examines how the activation of attention heads shapes model behavior. Michel et al. (2019) and Voita et al. (2019) show that only a small fraction of heads are necessary for translation tasks. Olsson et al. (2022) identify induction heads that capture repeated input patterns, and later extended the work to heads supporting in-context learning (Yin & Steinhardt, 2025; Ren et al., 2024). Other studies reveal specialized heads responsible for knowledge conflict (Shi et al., 2024; Jin et al., 2024b) and context distraction (Zhu et al., 2025). More recently, Wu et al. (2025) identify *retrieval heads*, which copy answer tokens from long contexts into the output, while Zhang et al. (2025) propose *query-focused retrieval heads* (QR heads) that better capture query–context interactions beyond direct copying. Collectively, these studies suggest that different subsets of heads specialize in distinct computational roles, and isolating the ones relevant to retrieval remains an open challenge. In this work, we build on the notion of retrieval heads (Wu et al., 2025; Zhang et al., 2025) as the mechanisms responsible for retrieving relevant information from long-context inputs – directly aligning with the re-ranking objective.

3 BACKGROUND

Attention-Based Re-Ranker. Given a query q and a set of k candidate documents $D = \{d_1, d_2, \ldots, d_k\}$ returned by a base retriever, the goal of the re-ranker is to reorder D so that the documents are sorted in the order of relevance to q . To achieve this, Chen et al. (2025) propose In-Context Re-ranking (ICR), an attention-based approach that leverages the attention signal from all layers and heads of a LLM to compute fine-grained relevance scores.

Given a LLM with L layers and H heads, ICR constructs an input prompt consisting of the list of k documents followed by the query q. For each document d_i, ICR computes the relevance score s_i by aggregating the attention score that each token in d_i gets from q across all heads and layers. The relevance score $s_{d_{i,j};q}$ of the j-th token in document d_i with respect to the query q is:

$$s_{d_{i,j};q} = \frac{1}{|\mathcal{T}_q|} \sum_{l=1}^{L} \sum_{h=1}^{H} \sum_{t \in \mathcal{T}_q} a_{j,t}^{l,h}, \tag{1}$$

where $\mathcal{T}_q$ denotes the set of the query tokens, and $a_{j,t}^{l,h}$ denotes the attention score from the t-th token in query q to the j-th token in document d_i by the h-th attention head in the l-th layer.

To reduce bias relative to document length and low-information tokens (e.g., punctuation), ICR applies *contextual calibration* on the relevance score with a content-free query q_{cf} (e.g. 'N/A') following Zhao et al. (2021). Specifically, the calibrated relevance score $s_{d_{i,j}}$ of the j-th token in document d_i is then calculated as:

$$s_{d_{i,j}} = s_{d_{i,j};q} - s_{d_{i,j};q_{cf}}. \tag{2}$$

After eliminating tokens with abnormally negative calibrated scores, the relevance score of document d_i is computed as the sum of remaining token scores.

Re-Ranker with QR Heads. As described above, ICR computes relevance scores by aggregating attention signals from all heads and layers. While comprehensive, this aggregation can introduce redundancy and noise, lowering re-ranking performance. Recent work on QR heads (Zhang et al., 2025) demonstrates that attention from as few as 1% of heads can improve retrieval effectiveness, indicating that many heads are unnecessary. A head h is considered QR head if its attention to the gold document, quantified as s_{gold}^h, is the highest among all heads. Let $\mathcal{T}_{gold}$ be the set of tokens in the gold document, then s_{gold}^h is computed as:

$$s_{gold}^h = \frac{1}{|\mathcal{T}_q|} \sum_{t \in \mathcal{T}_q} \sum_{j \in \mathcal{T}_{gold}} a_{j,t}^h. \tag{3}$$

However, the QR scoring criterion considers only absolute attention to relevant documents, without accounting for their relative ranking within each head's distribution. This omission can result in the selection of misleading heads. As a consequence, QR-based re-ranking shows inconsistently performance across datasets: as shown in Figure 1, selecting the top 8 QR heads reduces accuracy on the Quora duplicate-question retrieval task with Mistral 7B and Llama-3.1 8B, underperforming the ICR baseline that aggregates signals from all heads.

4 CONTRASTIVE RETRIEVAL HEADS

Addressing the limitations of prior approaches, we propose a contrastive head detection framework that identifies a small subset of *Contrastive Retrieval heads* (CoRe heads) specialized for document re-ranking. For clarity, we denote the default attention-based re-ranker as ICR, the QR-head re-ranker as QR-R, and the CoRe-head re-ranker as CoRe-R. Central to our method is a new scoring metric, S_{CoRe}, which measures a head's ability to highlight the gold document while simultaneously suppressing attention to irrelevant ones. By selecting high-quality heads using this contrastive metric, CoRe-R achieves more consistent and robust improvements in re-ranking, outperforming ICR and QR-R on Quora (Figure 1) and across the BEIR benchmark, as demonstrated in Section 5.2.

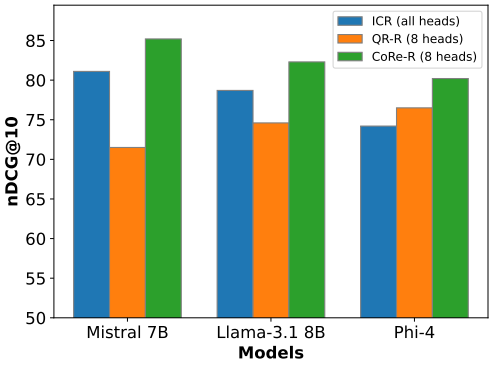


Figure 1: nDCG@10 on Quora top-40. Re-ranker with top 8 QR heads (QR-R) degrades the re-ranking task compared to ICR which uses all heads, while re-ranker with top 8 CoRe heads (CoRe-R) outperforms both ICR and QR-R.

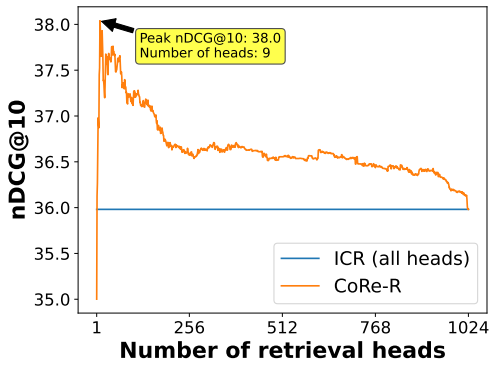


Figure 2: nDCG@10 on DBPedia top-40 with CoRe-R for Mistral 7B. Aggregated attention signal from fewer heads results in higher score. Re-ranking score peaks with top 9 CoRe heads.

Contrastive Scoring Metric. Our retrieval scoring metric follows the InfoNCE loss function (Oord et al., 2018), but requires no additional training. We consider the input prompt consisting of instructions (ignored in attention score computation), k documents and a query q. Suppose the attention score from the t-th token in query q to the j-th token in document d by the an attention head h is $a^h_{j,t}$, then the aggregated attention score head h pays toward document d is computed as:

$$s^h_d = \frac{1}{|\mathcal{T}_q|} \sum_{t \in \mathcal{T}_q} \sum_{j \in \mathcal{T}_d} a^h_{j,t}. \tag{4}$$

Let s^h_{pos} be the attention score head h pays toward the positive document, and $s^h_{neg,i}$ be the attention score head h pays toward the i-th negative document. Then, the retrieval score of head h is:

$$S_{CoRe}(h) = \frac{\exp(s^h_{pos}/t)}{\exp(s^h_{pos}/t) + \sum_i \exp(s^h_{neg,i}/t)}, \tag{5}$$

where t is a contrastive temperature hyperparameter.

The proposed contrastive scoring metric quantifies the degree to which an attention head prioritizes the gold document for a given query relative to competing irrelevant documents. A higher value of S_{CoRe} indicates that the corresponding attention head more effectively discriminates the positive document from hard negative documents.

Within the standard contrastive formulation, the temperature t acts as a scaling factor that controls the sharpness of the attention score distribution between the positive and hard negative documents. A lower temperature produces a sharper distribution, imposing greater penalties on hard negatives and therefore identifying the most effective attention heads. However, excessively low temperatures may become overly selective, favoring only heads that perform exceptionally well at distinguishing the correct document while discarding heads that are broadly useful. Conversely, higher temperatures soften the distribution, reducing sensitivity to hard negatives and allowing for greater tolerance to documents that are semantically similar to the gold document.

To identify CoRe heads, we compute the S_{CoRe} of each attention head averaged over a set of data samples, and select the top-scoring few attention heads within each model as CoRe heads. By aggregating attention signals from these CoRe heads, the re-ranker achieves improved performance compared to aggregating over all heads. Figure 2 shows the affect of aggregating over different number of retrieval heads on the re-ranking results on the DBPedia dataset. We observe that using fewer – but higher quality – attention heads yields superior accuracy, reaching peak re-ranking accuracy with only 9 CoRe heads.

Hard Negative Data. We compute the retrieval score S_{CoRe} for all attention heads using a subset of 1000 samples from the Natural Questions (NQ) training set (Kwiatkowski et al., 2019). Each

sample contains a query, a positive answer passage (gold document), and 49 hard negatives mined with the *ibm-granite/granite-embedding-30m-english* model (Awasthy et al., 2025). Hard negatives are constructed by sampling from the top 100 candidates returned by the retriever system. To further reduce false negatives, we follow de Souza P. Moreira et al. (2025) and discard any passage whose similarity to the query exceeds that of the gold passage. This procedure ensures that S_{CoRe} is computed under challenging retrieval conditions, ensuring that the detected heads are more discriminative and robust for re-ranking tasks.

Head Detection Process. For each sample, we construct a LLM prompt consisting of the 50 documents, followed by the instruction and query (see Appendix A for details). The positive document is inserted among the first five different positions, totaling to 5000 samples per language model. We compute the S_{CoRe} of each attention head on each sample, and compute the final retrieval head score by averaging across all samples. Less than top 1% of all heads with the highest averaged score are selected as CoRe heads (8 heads in our experiment). This setup yields stable average retrieval scores, with the top CoRe heads consistently converging to the same subset for each model. Moreover, CoRe heads identified using a single dataset NQ generalize effectively across multiple datasets, enhancing re-ranking performance as shown later in Section 5.2.

Re-Ranker with CoRe Heads. Once CoRe heads are identified for a given language model, re-ranking is performed by aggregating attention signals exclusively from this subset. Specifically, the new relevance score $s^*_{d_{i,j};q}$ in Equation 1 becomes:

$$s^*_{d_{i,j};q} = \frac{1}{|\mathcal{T}_q|} \sum_{h \in H^*} \sum_{t \in \mathcal{T}_q} a^h_{j,t}, \tag{6}$$

where H^* denotes the set of detected CoRe heads. The relevance score with respect to the content-free query $s^*_{d_{i,j};q_{cf}}$ is computed similarly. The calibrated relevance score $s^*_{d_{i,j}}$ of the j-th token in document d_i is calculated as:

$$s^*_{d_{i,j}} = s^*_{d_{i,j};q} - s^*_{d_{i,j};q_{cf}}. \tag{7}$$

Finally, the calibrated relevance scores are aggregated to produce the document level re-ranking.

5 Experiments

In this section, we evaluate the effectiveness of CoRe-R across a diverse set of datasets and four different open-weight decoder models, comparing its performance against other zero-shot attention-based re-rankers. We also analyze the effect of layer pruning, showing CoRe-R achieves strong re-ranking performance without requiring the computation of all layers, significantly reducing computational overhead. Our code is available at `https://github.com/linhhtran/CoRe-Reranking`.

5.1 Setup

Baselines. We compare CoRe-R with the following zero-shot attention-based re-rankers: ICR (Chen et al., 2025) and QR-R (Zhang et al., 2025). For QR head detection, we adopt the same detection data as described in Section 4 and exclude the contextual calibration step for a fair comparison with our head detection algorithm. Contextual calibration is applied only during the re-ranking stage, not the head detection step. For completeness, we report the re-ranking performance of QR-R with the original QR heads (detected with contextual calibration) in Section 5.4.

Datasets. We evaluate our approach on the BEIR benchmark (Thakur et al., 2021) which consists of fifteen diverse datasets: TREC-COVID (Voorhees et al., 2021), NFCorpus (Boteva et al., 2016), DBPedia-entity (Hasibi et al., 2017), SciFact (Wadden et al., 2020), SciDocs (Cohan et al., 2020), FiQA (Maia et al., 2018), NQ (Kwiatkowski et al., 2019), FEVER (Thorne et al., 2018), Climate-FEVER (Diggelmann et al., 2020), HotpotQA (Yang et al., 2018), Touche (Bondarenko et al., 2020), MSMARCO (Nguyen et al., 2016), Quora, ArguAna (Wachsmuth et al., 2018) and the CQADupStack series (Hoogeveen et al., 2015). Among them, NQ, HotpotQA, FEVER, Climate-FEVER,

Table 1: nDCG@10 on the BEIR benchmark.

Dataset	Retriever	Mistral 7B			Llama-3.1 8B			Phi-4			Granite-3.2 8B		
	Baseline	ICR	QR-R	CoRe-R	ICR	QR-R	CoRe-R	ICR	QR-R	CoRe-R	ICR	QR-R	CoRe-R
TREC-COVID	63.1	72.1	73.1	**73.8**	**76.8**	75.7	75.7	75.2	**76.8**	76.0	70.4	70.3	**71.8**
NFCorpus	33.7	33.1	35.0	**35.2**	35.0	36.4	**36.7**	31.6	35.5	**35.6**	31.8	32.5	**33.9**
DBPedia	36.0	36.0	36.8	**37.5**	38.7	39.0	**39.4**	36.9	39.4	**39.6**	35.6	35.9	**36.8**
SciFact	71.3	71.3	71.9	**74.7**	74.7	**75.9**	75.1	71.7	74.2	**74.7**	74.6	74.7	**75.2**
SciDocs	22.5	16.6	17.3	**19.5**	18.5	19.7	**20.2**	17.8	21.0	**21.1**	18.4	**20.3**	20.0
FiQA	36.9	35.8	40.3	**41.4**	41.3	44.0	**44.1**	37.2	43.8	**44.1**	39.7	41.6	**42.4**
NQ	51.6	55.4	56.5	**58.1**	60.8	63.0	**63.2**	57.4	63.3	**63.6**	57.4	57.9	**60.7**
FEVER	85.5	87.5	87.1	**88.7**	**89.2**	87.4	88.4	89.1	**89.6**	**89.6**	**88.0**	87.5	86.2
Climate-FEVER	30.3	21.6	21.1	**23.3**	22.5	**22.9**	22.7	22.2	**24.5**	24.3	20.5	**21.2**	21.0
HotpotQA	62.9	71.7	71.6	**73.4**	73.7	73.6	**73.8**	73.3	74.2	**74.5**	**73.6**	72.8	72.6
Touche	24.0	23.3	26.4	**26.7**	26.1	26.3	**26.4**	21.8	25.4	**25.7**	20.5	24.6	**24.8**
MSMARCO	30.7	30.1	**32.1**	31.7	32.9	33.9	**34.9**	30.6	34.6	**34.8**	29.6	31.2	**32.5**
Quora	86.7	81.1	71.5	**85.2**	78.7	74.6	**82.3**	74.2	76.5	**80.2**	**83.1**	79.8	75.1
ArguAna	56.4	45.0	51.1	**52.7**	42.5	54.2	**55.0**	41.7	52.2	**52.7**	52.3	56.1	**57.3**
CQADupstack	44.3	38.7	40.8	**41.6**	41.7	43.2	**43.9**	39.4	44.0	**44.3**	41.4	42.3	**43.4**
Average	49.1	47.9	48.9	**50.9**	50.2	51.3	**52.1**	48.0	51.7	**52.1**	49.1	50.0	**50.3**

and DBPedia-entity are in-domain datasets and are derived from Wikipedia articles. Other datasets are out-of-domain. In addition, we evaluate on the Multilingual Long-Document Retrieval (MLDR) dataset (Chen et al., 2024) to assess cross-language generalization.

Re-ranking Details. For each dataset in the BEIR benchmark, we re-rank the top-40 documents retrieved using the *ibm-granite/granite-embedding-30m-english* model (Awasthy et al., 2025). For the multilingual dataset MLDR, we re-rank the top-40 documents retrieved using the *ibm-granite/granite-embedding-107m-multilingual* model (Awasthy et al., 2025). Both QR-R and CoRe-R use the top 8 retrieval heads, and report performance using the nDCG@10 scores. To ensure consistency, we adopt a uniform prompt structure for all datasets; details of the prompt design are provided in Appendix A.

Language Models. Since attention-based re-ranking requires access to all attention scores of all layers and heads, we focus on open-source LLMs. Specifically, we evaluate Mistral 7B (Jiang et al., 2023), LLama-3.1 8B (Dubey et al., 2024) and Phi-4 (Abdin et al., 2024). We additionally conduct re-ranking experiment with Granite-3.2 8B (IBM Research, 2025) which belongs to the same LLM family as the chosen retriever embedding model. Main re-ranking results for Granite-3.2 8B are shown in Table 1 and Table 2, with further results and analyses provided in Appendix C.1.

Hyperparameters. The temperature hyperparameter t is tuned through grid search on a separate subset of the NQ train dataset. We select $t = 0.001$ for Mistral 7B and Granite-3.2 8B, and $t = 0.1$ for Llama-3.1 8B and Phi-4. These values are fixed for CoRe-R head detection prior to running the actual re-ranking experiments.

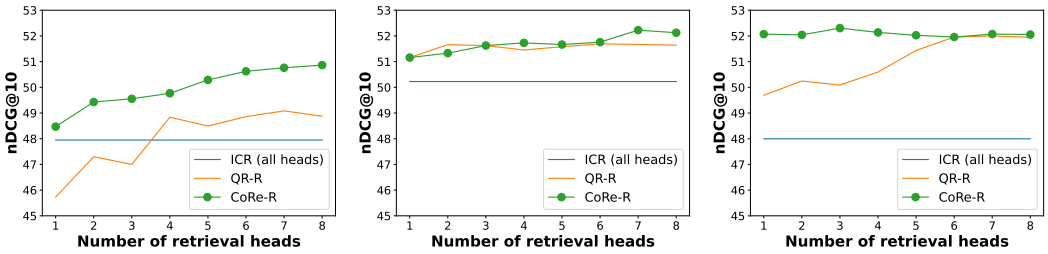
5.2 Results

Re-ranking performance. Table 1 presents the re-ranking results on the BEIR benchmark using the granite-embedding-30m-english retriever baseline. On average, CoRe-R achieves the best performance across all four models, consistently outperforming both ICR and QR-R. Aggregating attention from all heads in ICR yields the weakest results, in some cases falling below the retriever baseline for Mistral 7B and Phi-4. In contrast, CoRe-R delivers substantial improvements over the retriever baseline and ICR on every model. Compared to QR-R, CoRe-R shows the largest gains on Mistral 7B with average improvements of $+2.0$ points over QR-R and $+3.0$ points over ICR, and on Llama-3.1 8B with $+0.8$ and $+1.9$ points respectively. While QR-R remains competitive on the stronger Phi-4 model, its performance fluctuates across datasets, whereas CoRe-R provides more consistent and robust improvements. These findings demonstrate that CoRe-R effectively isolates meaningful attention heads across models and datasets, establishing it as a reliable state-of-the-art list-wise re-ranker.

Table 2: nDCG@10 on the MLDR datasets.

Dataset	Retriever Baseline	Mistral 7B			Llama-3.1 8B			Granite-3.2 8B		
		ICR	QR-R	CoRe-R	ICR	QR-R	CoRe-R	ICR	QR-R	CoRe-R
German	19.9	23.7	27.2	**28.2**	24.9	27.8	**28.6**	**28.1**	27.1	27.0
English	29.5	23.7	24.6	**28.1**	29.1	29.7	**30.2**	28.6	28.3	**29.2**
Spanish	43.3	39.1	43.3	**45.8**	43.3	46.1	**48.0**	46.4	46.9	**47.6**
French	49.1	47.3	46.9	**53.0**	50.1	52.3	**53.6**	53.1	53.7	**53.9**
Italian	41.9	36.0	35.9	**42.5**	41.6	43.1	**43.7**	42.4	**43.8**	**43.8**
Portuguese	52.1	46.1	49.1	**54.7**	52.3	55.2	**57.1**	55.6	56.0	**57.5**
Average	39.3	36.0	37.8	**42.1**	40.2	42.4	**43.5**	42.4	42.6	**43.2**

Cross-lingual Generalization. Table 2 reports the nDCG@10 on the MLDR datasets across three language models. We exclude Phi-4 as the model is not intended to support multilingual use (Microsoft Research, 2024). To ensure compatibility, we evaluate only languages that are supported by all three models. With the same set of CoRe heads detected using a subset of the NQ dataset, CoRe-R shows major improvement in the re-ranking accuracy over all baselines. For Mistral 7B, both ICR and QR-R underperform the retriever baseline on average, while CoRe-R achieves an average improvement of 2.8 points. For Llama-3.1 8B and Granite-3.2 8B, CoRe-R attains the best nDCG@10 scores on all languages with an exception of German for Granite. Overall, CoRe-R demonstrates consistent cross-lingual gains beyond English, underscoring the strong generalization ability of CoRe heads.



(a) Mistral 7B (b) Llama-3.1 8B (c) Phi-4

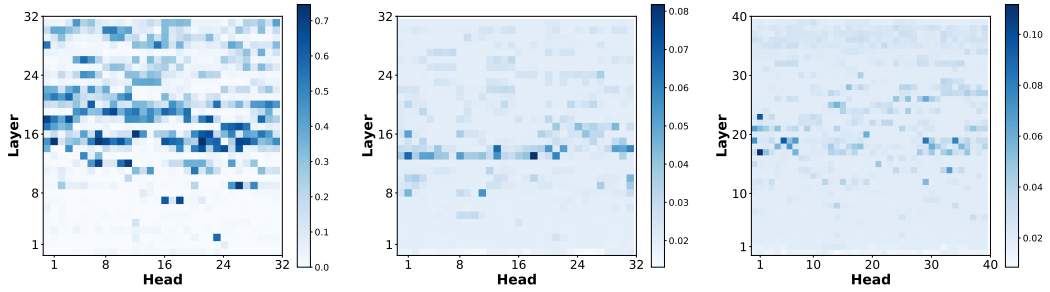
Figure 3: Average nDCG@10 on BEIR benchmark using the attention signal from different number of top retrieval heads.

Less Is More: Re-Ranking with Few CoRe Heads. Figure 3 illustrates the re-ranking performance as a function of the number of retrieval heads across three models. Across all settings, CoRe-R consistently outperforms ICR, demonstrating the effectiveness of CoRe heads for relevance ranking across all models. For Mistral 7B, CoRe-R substantially outperforms QR-R in performance for any number of retrieval heads. With Llama-3.1 8B (Figure 3b), QR-R and CoRe-R are comparable in the re-ranking performance for six or fewer number of retrieval heads, but CoRe-R at 8 heads surpasses QR-R. Interestingly, with Phi-4 (Figure 3c), although QR-R and CoRe-R deliver close average nDCG@10 scores with top 8 retrieval heads (51.7 vs. 52.1), CoRe-R has a very strong re-ranking performance even with one single attention head (52.1), outperforming QR-R at all head levels.

5.3 Re-ranking with Layer Pruning

Our empirical study shows that CoRe-R outperforms existing state-of-the-art list-wise re-rankers. Nevertheless, like other attention-based re-rankers, CoRe-R is memory-intensive, as it requires storing full attention score matrices that scale with the number of tokens. Moreover, computing the attention score matrices introduces latency, which can be particularly costly for list-wise re-rankers.

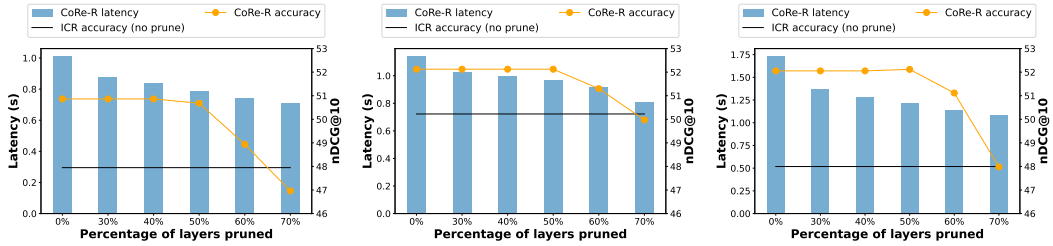
As illustrated in Figure 4, we observe that the highest scoring CoRe heads are concentrated in the middle layers across all evaluated models. Since attention-based re-ranking does not involve text



(a) Mistral 7B (b) Llama-3.1 8B (c) Phi-4

Figure 4: Distribution of S_{CoRe} for all heads in each model. We choose the temperature $t = 0.001$ for Mistral 7B and $t = 0.1$ for Llama-3.1 8B and Phi-4. The top CoRe heads are concentrated mostly in the middle layers of every model.

generation, the final layers, primarily responsible for token generation, are not essential for our task. This observation motivates the use of layer pruning to improve inference efficiency and reduce memory usage. We hypothesize that pruning the majority of the late layers can substantially reduce both memory consumption and computational cost while preserving re-ranking accuracy.



(a) Mistral 7B (b) Llama-3.1 8B (c) Phi-4

Figure 5: Average latency and re-ranking accuracy on BEIR benchmark. Pruning 50% of the model's layers does not impact the re-ranking performance while saving 20% inference time.

Figure 5 reports the average re-ranking accuracy and latency on BEIR under different levels of layer pruning, while Figure 6 shows the corresponding peak GPU memory usage. The experiment is done on a single GPU H100 96GB. Results show that pruning up to 50% of the late layers preserves re-ranking performance while reducing GPU memory usage by approximately 40% and latency by 20%. Beyond this point, performance begins to degrade: pruning 60% of the layers leads to a noticeable decline in the re-ranking performance, underscoring the central role of the CoRe heads in the middle layers. Nevertheless, even with 60% layers pruned, CoRe-R continues to outperform the ICR baseline across all models. This indicates that the remaining CoRe heads in earlier layers[1] carry sufficiently strong and discriminating signals, surpassing the noisy aggregation of all heads in ICR.

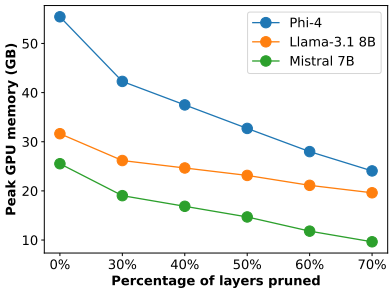


Figure 6: Peak GPU memory usage on BEIR datasets. Pruning 50% of the model's layers saves 40% memory usage.

[1]Specifically layers 8-12 for Mistral 7B and Llama-3.1 8B, and layers 12-16 for Phi-4

Table 3: nDCG@10 on BEIR benchmark for Llama-3.1 8B.

Dataset	Retriever Baseline	ICR	NIAH-R	QR-R (w/ calib)	QR-R (w/o calib)	CoRe-R
Covid	63.1	**76.8**	73.7	75.2	75.7	75.7
NFCorpus	33.7	35.0	34.2	35.7	36.4	**36.7**
DBPedia	36.0	38.7	37.1	39.2	39.0	**39.4**
SciFact	71.3	74.7	74.2	74.4	**75.9**	75.1
SciDocs	22.5	18.5	19.2	19.2	19.7	**20.2**
FiQA	36.9	41.3	40.8	43.4	44.0	**44.1**
NQ	51.6	60.8	57.5	61.4	63.0	**63.2**
FEVER	85.5	**89.2**	88.1	88.5	87.4	88.4
Climate-FEVER	30.3	22.5	**23.1**	22.8	22.9	22.7
HotpotQA	62.9	73.7	72.0	73.5	73.6	**73.8**
Touche	24.0	26.1	25.5	26.3	26.3	**26.4**
MSMARCO	30.7	32.9	31.6	34.7	33.9	**34.9**
Quora	86.7	78.7	80.1	74.5	74.6	**82.3**
ArguAna	56.4	42.5	51.7	51.2	54.2	**55.0**
CQADupstack	44.3	41.7	41.4	43.2	43.2	**43.9**
Average	49.1	50.2	50.0	50.9	51.3	**52.1**

5.4 Impact of Head Detection Strategies

In this subsection, we experiment with different set of retrieval heads, including the original retrieval heads (Wu et al., 2025) identified via copy-paste mechanism on Needle-in-a-Haystack task. We refer to these heads as *NIAH heads*. For Llama-3.1 8B, none of the top 8 CoRe heads overlaps with the top 8 NIAH heads, and only 3 overlap with the top 8 QR heads. For Mistral 7B, top 8 CoRe heads overlap with 2 QR heads and 2 NIAH heads. In addition, CoRe heads concentrate in lower layers than both QR heads and NIAH heads. We provide details of the head location in Appendix B.

Table 3 compares CoRe-R with the attention-based re-rankers using the NIAH heads (NIAH-R), QR heads (detected without calibration) and the original QR heads (detected with calibration), for Llama-3.1 8B. Results for ICR, QR-R (without calibration) and CoRe-R are duplicated from Table 1 for reference. As NIAH retrieval heads are limited to copy-paste operations, NIAH-R underperforms ICR on average, showing gains only on a few datasets such as Climate-FEVER and ArguAna, while showing worse performance on many other datasets. On average, the QR-R with calibration improves over ICR by 0.7 points and QR-R without calibration improves by 1.1 points, indicating that contextual calibration is not essential for the head detection process. In contrast, CoRe-R shows the largest gain of 1.9 points, yielding the highest overall nDCG@10 among all baselines.

Overall, attention-based re-rankers achieve strong re-ranking performance across different models, with CoRe-R consistently providing the best nDCG@10 scores. However, there remain cases where all attention-based re-rankers underperform the retrieval baseline (see Tables 1 and 3). Notably, on Climate-FEVER and ArguAna, which often require retrieving counter-evidence, attention-based approaches lab behind the encoder retrieval baseline. Similarly, on the duplicated retrieval dataset Quora, performance also declines. We attribute this to the use of a universal instruction prompt across all datasets, which may not align with dataset-specific task objectives. Future work could explore dataset specific prompt designs to further improve re-ranking performance (see Appendix C.2 for an example on Quora dataset).

6 Conclusion

In this paper, we introduced *Contrastive Retrieval heads* (CoRe heads), a subset of attention heads that capture the most discriminative signals for document re-ranking. We proposed a contrastive scoring metric that identifies these heads by rewarding attention to relevant documents while penalizing focus on irrelevant ones. Across extensive experiments, we showed that aggregating attention from CoRe heads produces a state-of-the-art re-ranker, consistently surpassing prior baselines across tasks and models. Our analysis further revealed that CoRe heads cluster in the middle transformer layers, enabling an effective layer-pruning strategy that cuts inference latency and memory usage without sacrificing accuracy. Together, these results establish CoRe heads as primary carriers of relevance information for re-ranking and underscore their promise for building fast, accurate, and efficient retrieval systems.

REFERENCES

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

Parul Awasthy, Aashka Trivedi, Yulong Li, Mihaela Bornea, David Cox, Abraham Daniels, Martin Franz, Gabe Goodhart, Bhavani Iyer, Vishwajeet Kumar, Luis Lastras, Scott McCarley, Rudra Murthy, Vignesh P, Sara Rosenthal, Salim Roukos, Jaydeep Sen, Sukriti Sharma, Avirup Sil, Kate Soule, Arafat Sultan, and Radu Florian. Granite embedding models, 2025. URL `https://arxiv.org/abs/2502.20204`.

Alexander Bondarenko, Maik Fröbe, Meriem Beloucif, Lukas Gienapp, Yamen Ajjour, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, et al. Overview of touché 2020: argument retrieval. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pp. 384–395. Springer, 2020.

Vera Boteva, Demian Gholipour, Artem Sokolov, and Stefan Riezler. A full-text learning to rank dataset for medical information retrieval. In *European Conference on Information Retrieval*, pp. 716–722. Springer, 2016.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 2024.

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

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

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

Gabriel de Souza P. Moreira, Radek Osmulski, Mengyao Xu, Ronay Ak, Benedikt Schifferer, and Even Oldridge. Nv-retriever: Improving text embedding models with effective hard-negative mining, 2025. URL `https://arxiv.org/abs/2407.15831`.

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

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pp. arXiv–2407, 2024.

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

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

Faegheh Hasibi, Fedor Nikolaev, Chenyan Xiong, Krisztian Balog, Svein Erik Bratsberg, Alexander Kotov, and Jamie Callan. 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*, pp. 1265–1268, 2017.

Doris Hoogeveen, Karin M Verspoor, and Timothy Baldwin. Cqadupstack: A benchmark data set for community question-answering research. In *Proceedings of the 20th Australasian document computing symposium*, pp. 1–8, 2015.

IBM Research. Granite-3.2-8b-instruct, 2025. URL https://huggingface.co/ibm-granite/granite-3.2-8b-instruct.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023.

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

Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. 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*, pp. 1193–1215, 2024b.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, 2020.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

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

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

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. Www'18 open challenge: financial opinion mining and question answering. In *Companion proceedings of the the web conference 2018*, pp. 1941–1942, 2018.

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

Microsoft Research. Phi-4, 2024. URL https://huggingface.co/microsoft/phi-4.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. Ms marco: A human-generated machine reading comprehension dataset. 2016.

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

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

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

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

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

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

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

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 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*, pp. 14918–14937, 2023.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 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*, 2021.

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

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 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*, pp. 5797–5808, 2019.

Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. Trec-covid: constructing a pandemic information retrieval test collection. In *ACM SIGIR Forum*, pp. 1–12. ACM New York, NY, USA, 2021.

Henning Wachsmuth, Shahbaz Syed, and Benno Stein. Retrieval of the best counterargument without prior topic knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 241–251, 2018.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. *arXiv preprint arXiv:2004.14974*, 2020.

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

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*, 2018.

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

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

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

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

A PROMPT STRUCTURE

We use the same prompt structure for both head detection and re-ranking steps, which is adopted from Chen et al. (2025). The prompt starts with the instruction `Here are some paragraphs:` and the list of retrieved documents separated by a newline. After the list of document, there is a second prefix prompt for the query `Please find information that are relevant to the following query in the paragraphs above. Query:`, and the query is appended at the end. The entire prompt string is wrapped by the special tokens `start token` and `end token` which vary based on the LLM used. See Table 4 for an example.

Table 4: An example prompt from NQ with Mistral 7B.

`[INST] Here are some paragraphs:`

`[document 1]` Can't Help Falling in Love "Can't Help Falling in Love" is a pop ballad originally recorded by American singer Elvis Presley and published by Gladys Music, Presley's publishing company. It was written by Hugo Peretti, Luigi Creatore, and George David Weiss. The song was featured in Presley's 1961 film, Blue Hawaii. During the following four decades, it was recorded by numerous other artists, including Tom Smothers, Swedish pop group A-Teens, and the British reggae group UB40, whose 1993 version topped the U.S. and UK charts.

`[document 2]` Can't Help Falling in Love In 2015, the song was included on the If I Can Dream album, on the occasion of the 80th anniversary of Presley's birth. The version uses archival voice recordings of Presley and his singers, backed by new orchestral arrangements performed by the Royal Philharmonic Orchestra.

...

`[document 40]` I Can't Help It (If I'm Still in Love with You) Williams sang the song with Anita Carter on the Kate Smith Evening Hour on April 23, 1952. The rare television appearance is one of the few film clips of Williams in performance.

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

`Query:` who recorded i can't help falling in love with you `[/INST]`

B DISTRIBUTION OF RETRIEVAL HEADS

B.1 RETRIEVAL HEAD LOCATION

We list below the exact location of the top 8 NIAH retrieval heads, QR heads (without calibration) and CoRe heads for Llama-3.1 8B and Mistral 7B. We format (L-H) as layer L head H, for example, (7-19) means layer 7 head 19.

Llama-3.1 8B:

- NIAH retrieval heads: (15-30), (27-7), (8-1), (16-1), (24-27), (16-20), (5-8), (16-23).
- QR heads: (13-18), (14-13), (13-1), (20-14), (14-29), (16-1), (14-22), (17-29).
- CoRe heads: (13-18), (13-1), (14-13), (13-21), (14-31), (13-13), (8-11), (14-20).

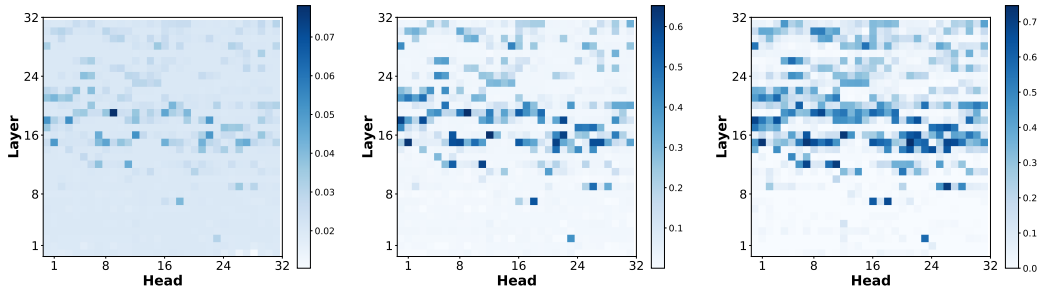
Mistral 7B:

- NIAH retrieval heads: (18-0), (12-7), (12-6), (18-2), (18-3), (18-1), (30-8), (28-0).
- QR heads: (18-22), (15-26), (20-17), (18-0), (19-9), (16-22), (16-12), (19-16).
- CoRe heads: (15-21), (15-1), (16-12), (15-7), (9-26), (12-11), (12-7), (18-0).

We found that for Llama-3.1 8B, our top CoRe heads overlaps with QR heads in 3 heads, but does not overlap with NIAH retrieval heads. For Mistral 7B, CoRe heads overlap in 2 heads with both QR heads and NIAH retrieval heads. In both models, CoRe heads appear in earlier layers compared to other heads.

B.2 Effect of Contrastive Temperature

As discussed in Section 4, the contrastive temperature t controls the sharpness of the head distribution. Figure 7 demonstrates the effect of the temperature on the distribution of CoRe heads within Mistral 7B. Lower temperature heavily penalizes the non-CoRe heads which leads to larger gap in the S_{CoRe}. Nonetheless, all levels of temperatures result in similar distribution of the top $CoRe$ heads with difference in only 1 or 2 heads.



(a) $t = 0.1$ (b) $t = 0.01$ (c) t=0.001

Figure 7: Distribution of S_{CoRe} for all heads in Mistral 7B with different temperature t. All temperatures result in similar distribution of CoRe heads with the top few heads lie in the middle layers.

C Additional Experiment Results

C.1 Results on Granite-3.2 8B Model

In this subsection, we provide experimental results on Granite-3.2 8B model.

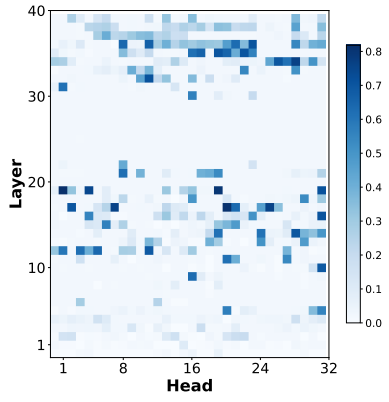


Figure 8: Distribution of S_{CoRe} for all heads in Granite-3.2 8B with temperature $t = 0.001$. The top CoRe heads are in middle layers or late layers.

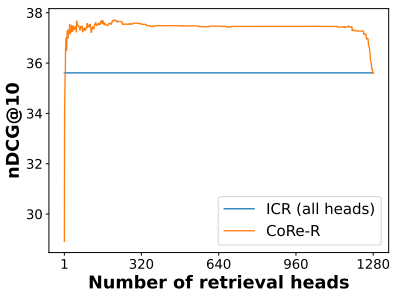


Figure 9: nDCG@10 on DBPedia top-40 with CoRe-R for Granite-3.2 8B. All attention heads within Granite-3.2 8B are equally important for re-ranking tasks.

Figure 8 demonstrates the distribution of CoRe heads within Granite-3.2 8B model, and the specific CoRe heads in Granite-3.2 8B are:

$$(19\text{-}1),\ (17\text{-}20),\ (19\text{-}19),\ (34\text{-}28),\ (17\text{-}25),\ (17\text{-}7),\ (19\text{-}4),\ (19\text{-}31).$$

Beside the CoRe heads concentrated in the middle layers, many heads in the late layers also exhibit high S_{CoRe} which can potentially be considered CoRe heads. Nevertheless, most of the top 8 CoRe heads still lie in the middle layers for Granite, with an exception for head (34-28).

Figure 9 shows the nDCG@10 score on DBPedia dataset for Granite-3.2 8B with different number of retrieval heads. Similar to other models, we observe that the attention signal from fewer number of heads achieves better re-ranking accuracy than the noisy aggregation over all heads (ICR), and the nDCG@10 score peaks with a small number of CoRe heads. Interestingly, there is negligible accuracy degradation across different number of retrieval heads. This trend indicates that most attention heads within Granite-3.2 8B do not contribute much noise to the aggregated attention signals. Hence, the activation of more attention heads has a minimal influence on the final re-ranking accuracy. Figure 10 provides a closer look at the top 8 CoRe heads with comparison to ICR and QR-R. Again, CoRe-R with more than 3 heads consistently outperforms ICR and QR-R.

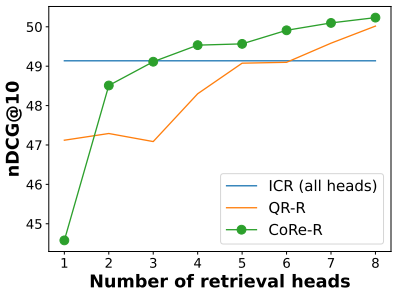


Figure 10: Average nDCG@10 on BEIR benchmark for Granite-3.2 8B using attention signal from different number of top retrieval heads.

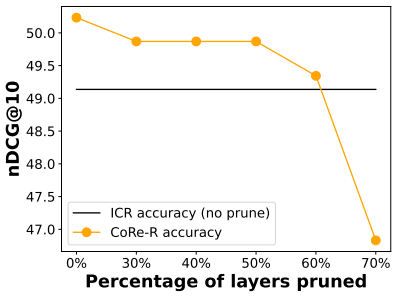


Figure 11: Average nDCG@10 on BEIR benchmark for Granite-3.2 8B with different levels of layer pruning.

Figure 11 demonstrates the affect of layer pruning on the re-ranking accuracy of CoRe-R with Granite-3.2 8B. We observe that pruning $30 - 50\%$ of the final layers results in nearly identical nDCG@10 with a slight drop of 0.4 points compared to no pruning. This result highlights the importance of the late CoRe head (34-28). Similar to other models, there is a decline in re-ranking performance with more than 50% layers pruned, underscoring the importance of the CoRe heads in the middle layers.

C.2 Results with Customized Prompts

As mentioned in Section 5.4, all attention-based re-rankers underperform the retriever baseline in some datasets that require more task-specific instruction prompts. For example, we show that a customized prompt for duplicated question dataset Quora, i.e. `Please identify question that has the exact same meaning with the following query. Query:` can improve the re-ranking performance of attention-based re-rankers.

Table 5: nDCG@10 on Quora dataset with the universal prompt and the customized prompt.

Method	Llama-3.1 8B		Phi-4		Granite-3.2 8B	
	universal	customized	universal	customized	universal	customized
ICR	78.7	83.8	74.2	76.8	**83.1**	83.8
QR-R	74.6	80.8	76.5	80.7	79.8	82.3
CoRe-R	**82.3**	**85.4**	**80.2**	**85.4**	75.1	**84.4**

We report the re-ranking results in Table 5. Overall, the task-specific prompt increases the nDCG@10 scores for all attention-based re-rankers by large margin. As expected, CoRe-R still stands out as the best-performing re-ranker, delivering the best re-ranking accuracy across all models. Future work could explore different ways to optimize the task-specific prompt design for Quora as well as other datasets, which further improve the final re-ranking performance.