What Makes Attention Distillation Work? An Exploration of Attention Distillation in Retrieval-Based Language Model

Anonymous ACL submission

Abstract

Retrieval-based language models address the limitations of large language models by enabling real-time knowledge updates for more accurate answers. An efficient way in the training phase of retrieval-based models is attention distillation, which uses attention scores as a supervision signal instead of manually annotated query-document pairs. Despite its growing popularity, the detailed mechanisms behind the success of attention distillation remain unexplored, particularly the specific patterns it leverages to benefit training. In this paper, we address this gap by conducting a comprehensive review of attention distillation workflow and identifying key factors influencing the learning quality of retrieval-based language models. We further propose indicators for optimizing models' training methods and avoiding ineffective training.

1 Introduction

Large language models have showcased remarkable capabilities across various natural language processing tasks (Min et al., 2023; OpenAI, 2023; Ouyang et al., 2022). However, their fixed parameters limit their ability to update knowledge in real-time, making them prone to producing unreliable content (Zhang et al., 2023). Additionally, these models also lack protection for sensitive training data (Nasr et al., 2023; Lin et al., 2021). One promising method to overcome these limitations is using retrieval-based language models (Ram et al., 2023; Shi et al.; Izacard et al., 2022b; Guu et al., 2020; Karpukhin et al., 2020; Khandelwal et al., 2019). Retrieval-based language models typically comprise two main components: (1) *the retriever*, which selects relevant information, and (2) *the reader*, incorporates this information into the generation process. Combining these two components, retrieval-based language models not only improve accuracy and reliability by dynamically using external knowledge but also reduce training costs with fewer trainable parameters (Shi et al., 2023; Shuster et al., 2021).

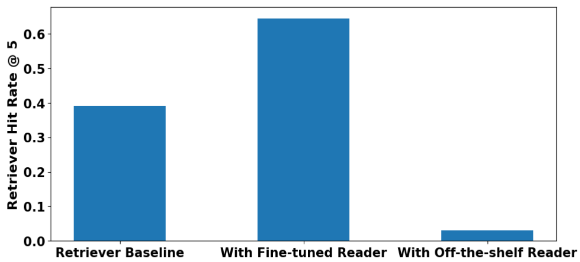

Figure 1: Training *Contriever* on *NaturalQuestions* for the QA task with attention distillation shows an improved Hit Rate @ 5 with a fine-tuned reader but a significant decline with an off-the-shelf reader.

Various methods have been proposed to improve the coordination between the retriever and the reader (Karpukhin et al., 2020; Jiang et al., 2023). Among these, attention score-based knowledge distillation has shown its effectiveness (Izacard and Grave, 2020a), outperforming other established methods (Karpukhin et al., 2020; Lewis et al., 2020; Izacard and Grave, 2020b) in QA tasks. In this process, the attention scores from the reader are captured and conveyed to the retriever as the supervisory signal, enabling the retrieval model to more effectively identify information candidates that can significantly improve the language model's responses. This efficient strategy reduces the need for manual annotation of the knowledge corpus, saving resources while achieving satisfactory results (Hu et al., 2023; Wang et al., 2023).

However, its efficiency heavily relies on the reader model's quality. As Figure 1 shows, low-quality reader models yield ineffective supervision signals, detrimentally impacting the retriever's performance. A fundamental hypothesis underpinning this mechanism is that more attention to certain tokens suggests greater relevance in answering questions (Izacard and Grave, 2020a), yet this correlation is not clearly defined. Our research seeks

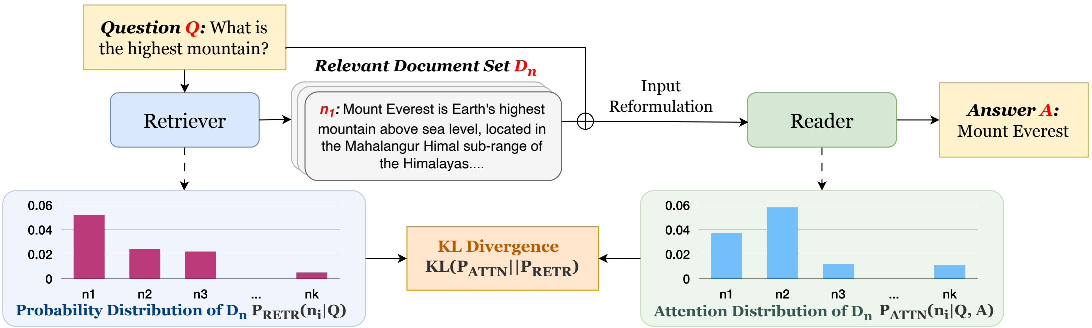

Figure 2: The framework of the Retrieval-Based Language Model of our experiment.

to understand which text segments gather more attention and how to assess attention quality. Given the unpredictable training outcomes due to these uncertainties, we aim to enhance the applicability and reliability of attention distillation training.

This paper conducts a detailed analysis of attention distillation training methods in question-answering (QA) tasks, exploring various settings to determine their effects on retrieval-based language model performance. We aim to identify the characteristics of high-quality attention scores and establish criteria for evaluating them in retrieval-based language model training. Specifically, our main contributions are as follows:

- We conduct an extensive analysis of attention scores in language models, mainly focusing on the prevalent decoder-only structure, to understand their impact on retriever model training and the overall performance of retrieval-based language models, thereby identifying key factors that significantly influence the model's performance.
- We introduce novel metrics to evaluate the reader model's proficiency in attention distillation, aiming to improve training performance by leaning on effective training sessions.

2 Method

In our experiment, we adapt the ATLAS architecture (Izacard et al., 2022b) but use a decoder-only structure for our empirical analysis, focusing on question-answering tasks to study attention mechanisms in the reader models. Specifically, for a given question Q, we supply models with a knowledge base $D = \{d_1, d_2, ..., d_m\}$, where each d_i is a unique document. The objective of the models is to find the question-relevant documents $D_n = \{n_1, n_2, ..., n_k\} \subseteq D$ using the retriever, and then generate the answer A using the reader.

To accommodate the change in reader structure, we modify the original attention distillation method. Instead of using *cross-attention scores* between the input document and output as an indicator of document relevance, we utilize *self-attention scores* concerning the output tokens. Notice that the contribution of a token t is not only evaluated from the attention score α_t but also the norm of the value should be taken v_t into account (Izacard et al., 2022b). The attention score distribution over D_n can be calculated as

$$p_{ATTN}(n_i|Q, A) = \sum_{t=1}^{T} \alpha_t v_t \quad (1)$$

where T represents the total number of tokens in n_i. During training, the attention scores are distilled into the retriever by minimizing KL-divergence with the retriever's probability distribution p_{RETR}. p_{RETR} over D_n can be defined as

$$p_{RETR}(n_i|Q) = \frac{exp(s(n_i, Q)/\theta)}{\sum_{k=1}^{K} exp(s(n_k, Q)/\theta)} \quad (2)$$

where s denotes the dot-product of query and document vectors, and θ is the temperature hyper-parameter. Figure 2 visually illustrates the retrieval process and the utilization of attention scores during training.

3 Experiments

We chose *Falcon-1b* (Penedo et al., 2023a) as our primary decoder-only reader model for its performance and flexibility, and we follow ATLAS (Izacard et al., 2022b) in using *Contriver* as the retriever model. During the retrieval process, we fix the retrieved documents D_n's size to $k = 5$ to balance training costs with the amount of information retrieved, avoiding inefficiencies of either extreme.

3.1 Experiment Setup

Dataset We assess the model's performance using the *NaturalQuestions* (Kwiatkowski et al., 2019) and the *TriviaQA* (Joshi et al., 2017) benchmarks. For the knowledge base, we utilize data from Wikipedia as of December 20, 2018.

Experimental Settings Specifically, we use the following settings for our experiments.

1) Off-the-shelf Distillation Training: We synchronously train the model using the initial *Falcon-1b* (Penedo et al., 2023b) as the reader and *Contriever* (Izacard et al., 2022a) as the retriever.

2) Fine-tuned Distillation Training: This experiment involves two steps:

Step1. We start with the initial *Falcon-1b* as reader and *Contriever* as retriever, only fine-tuning reader while keeping retriever's parameters fixed.

Step2. We continue training the retriever using the fine-tuned reader from Step1, updating the knowledge base index periodically.

Evaluation Metrics: We assess the model performance in terms of retrieval quality and question-answering correctness, given the involvement of both retriever and reader models. We use the *top-5* retrieval Hit Rate (HR@5), which is the proportion of retrieved documents D_n containing at least one answer A, to measure the retriever's effectiveness. For the reader's QA performance, we employ the standard Exact Match (EM) metric and F1-Score.

3.2 Results and Discussion

In this section, we empirically analyze the effectiveness of attention distillation training by answering the following research questions:

RQ1: When does the attention distillation work? As shown in Table 1, the *Fine-tuned Distillation Training* after Step2 shows the best performance in both EM and HR@5. In contrast, *Off-the-shelf Distillation Training* performs the worst, with its retriever even underperforming the initial Contriever model (i.e., the retriever model of *Fine-tuned Distillation Training* Step1). Notice that the critical difference lies in the quality of the reader models: *Off-the-shelf Distillation Training* uses the initial Falcon-1b model, whereas *Fine-tuned Distillation Training* employs a well-tuned Falcon-1b. These experimental results strongly suggest that the quality of attention scores is pivotal: **attention scores from the high-quality readers enhance training, whereas low-quality ones lead to poor interaction between the retriever and the reader.**

RQ2: Are there any commonalities in attention scores from the high-quality readers?

We sample 1000 data instances from each experiment to obtain reliable analysis results. We focus on the attention score characteristics **at token level** to identify which tokens receive more attention from high-quality signals. Our analysis firstly finds that in the high-quality readers, the tokens most related to *answer* and *nouns in question* receive the most attention. Based on our initial observations, we secondly focus on studying the distribution of attention scores for *answer-related* and *question-related* [1] tokens. We use token embedding's *cosine similarity* to measure its proximity to targets (i.e., answer or nouns in question), selecting the top 5% and top 10% of closest tokens and analyzing their average *attention scores* and *Spearman correlation with similarity to target tokens*, as shown in Table 2[2]. We also include the *Off-the-shelf Checkpoint* as a baseline to observe attention score evolution in different settings. This analysis identifies the key commonalities in high-quality attention scores.

Commonality1. Higher attention to answer tokens in higher-quality models. In all training settings, tokens closer to answer tokens (i.e., from a similarity higher than 90^{th} percentile to a similarity higher than 95^{th} percentile) receive increasingly higher attention scores. It can be observed that for both two measure metrics, the *Off-the-shelf Distillation Training* results are lower compared to the *Off-the-shelf Checkpoint*, while *Fine-tuned Distillation Training* shows improvement in both Step1 and Step2. The results suggest that in *Off-the-shelf Distillation*, the reader's attention does not effectively "highlight" key information, leading to suboptimal training. In contrast, *Fine-tuned Distillation* after Step1 and Step2 both indicate that high-quality readers focus more on relevant answer tokens, thereby enhancing both the retriever's performance and the relevance of attention allocated to these tokens.

Commonality 2. Tokens similar to question

Table 1: Model's Performance of Different Experimental Settings

Method	Dataset	Evaluation Metrics		
		EM↑	F1↑	HR@5↑
Off-the-shelf Distillation	NQ	27.24	33.62	0.030
	TriviaQA	30.55	35.24	0.022
Fine-tuned Distillation (Step1)	NQ	31.76	38.72	0.391
	TriviaQA	44.62	50.79	0.516
Fine-tuned Distillation (Step2)	NQ	35.22	43.44	0.645
	TriviaQA	54.59	61.04	0.643

[1] We only focus on the nouns in the question in selecting *question-related* tokens.

[2] The highest values in the table are highlighted in bold on the NQ Dataset and underlined on the TriviaQA Dataset.

Table 2: Average values of attention scores and Spearman correlation in *answer-related* and *question-related* tokens

Experiment	Dataset	Answer-related				Question-related			
		90^{th} percentile		95^{th} percentile		90^{th} percentile		95^{th} percentile	
		Attn.	Corr.	Attn.	Corr.	Attn.	Corr.	Attn.	Corr.
Off-the-shelf Checkpoint	NQ	0.033	0.227	0.039	0.196	0.023	0.103	0.024	0.092
	TriviaQA	0.027	0.218	0.032	0.206	0.021	0.103	0.023	0.067
Off-the-shelf Attention Distillation	NQ	0.017	0.145	0.017	0.076	0.027	0.139	0.039	0.153
	TriviaQA	0.031	0.160	0.035	0.172	0.047	0.144	0.063	0.260
Fine-tuned Attention Distillation (Step1)	NQ	0.039	0.308	0.052	0.282	**0.035**	**0.343**	**0.045**	**0.333**
	TriviaQA	0.058	0.259	0.074	0.258	0.058	0.349	0.078	0.372
Fine-tuned Attention Distillation (Step2)	NQ	**0.049**	**0.316**	**0.066**	**0.350**	0.032	0.310	0.039	0.225
	TriviaQA	0.069	0.290	0.089	0.320	0.060	0.367	0.078	0.326

nouns receive more attention in high-quality models. Table 2 also indicates that tokens closer to the nouns in question tokens receive higher attention scores. The *Fine-tuned Distillation* experiments exhibit much higher values in both metrics compared to *Off-the-shelf Checkpoint* and *Off-the-shelf Attention Distillation*, aligning with their superior performance. However, unlike Commonality 1, the Spearman correlation between attention to question-related tokens and model performance isn't consistent: while *Fine-tuned Attention Distillation* Step2 surpasses Step1, its metric values do not consistently align with this improvement, suggesting a more complex relationship.

RQ3: How do we evaluate the quality of attention distillation on decoder-only readers based on the analysis results?

Indicator1. Focusing on the attention scores of the nearest tokens to answer A, denoted as $M_A = \{ma_1, ..., ma_k\}$. Higher average $P_{ATTN}(ma_i)$ values indicate better attention distillation quality. Additionally, a higher average Spearman correlation between the $P_{ATTN}(ma_i)$ and their semantic similarity to A also signifies better quality.

Indicator2. Examining the attention scores of tokens closest to nouns in question Q, denoted as $M_Q = \{mq_1, ..., mq_k\}$. An increase in average $P_{ATTN}(mq_i)$ suggests better quality. Moreover, if the average Spearman correlation between the attention scores of M_Q and their similarity to Q is above the threshold for a weak monotonic relationship (i.e., value > 0.3), the attention distillation quality is considered good.

RQ4: Can we extend the proposed indicators to encoder-to-decoder structure readers?

An analysis with the fine-tuned encoder-to-decoder structure *Atlas-large* model is presented in Figure 3. The results show that the performance of *Atlas-large* surpasses *Fine-tuned Distillation Training (Step2)*. However, only the average $P_{ATTN}(ma_i)$ trend from Indicator1 applies to this encoder-to-decoder structure model, while *Atlas-large* exhibits a polarized distribution for the Spearman correlation values. (see Appendix A).

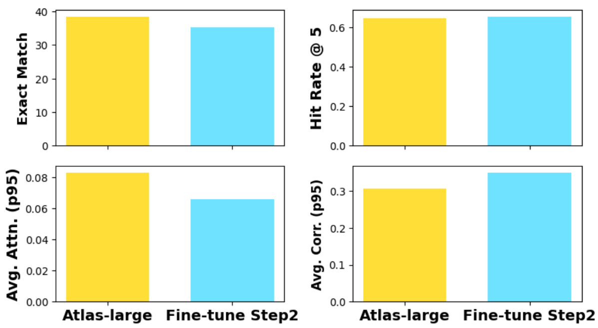

Figure 3: Model performance (top) and their attention distillation analysis (bottom) of *Atlas-large* model (yellow) for the answer-related tokens, comparing with *Fine-tuned Distillation Training (Step2)* (blue).

RQ5: Can we extend the proposed indicators to perplexity distillation training?

Finally, we want to determine if our indicators can apply to perplexity distillation, another popular knowledge distillation method used in training the retriever model. We fine-tune *Atlas-large* model with the perplexity distillation method and find that the perplexity distribution does not align with either Commonality 1 or Commonality 2, saying that our indicators are not suitable for perplexity distillation (details in Appendix A and B).

4 Conclusion

In this paper, we comprehensively evaluate attention distillation for training retrieval-based language models, emphasizing the importance of attention to answer and question-related tokens. We further introduce novel metrics for assessing language models' attention distillation ability to optimize the training process.

5 Limitation

This paper analyzes the attention score-based knowledge distillation quality in training retrieval-based language models under various experimental settings in QA tasks. Furthermore, based on our findings, we have developed two indicators to assess the quality of attention score supervision. However, our exploration is conducted based on lightweight language models (i.e., language models with about one billion parameters) due to their flexibility and have yet to extend to larger-scale language models. In future work, we will focus on validating the accuracy of our methods on more extensive language models to enhance the generalizability and applicability of our results.

References

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Linmei Hu, Zeyi Liu, Ziwang Zhao, Lei Hou, Liqiang Nie, and Juanzi Li. 2023. A survey of knowledge enhanced pre-trained language models. *IEEE Transactions on Knowledge and Data Engineering*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022a. Unsupervised dense information retrieval with contrastive learning.

Gautier Izacard and Edouard Grave. 2020a. Distilling knowledge from reader to retriever for question answering. *arXiv preprint arXiv:2012.04584*.

Gautier Izacard and Edouard Grave. 2020b. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022b. Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. *arXiv preprint arXiv:2305.06983*.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

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*.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2019. Generalization through memorization: Nearest neighbor language models. *arXiv preprint arXiv:1911.00172*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

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

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40.

Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. 2023. Scalable extraction of training data from (production) language models.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023a. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023b. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *arXiv preprint arXiv:2302.00083*.

Weijia Shi, Julian Michael, Suchin Gururangan, and Luke Zettlemoyer. knn-prompt: Nearest neighbor zero-shot inference, 2022b. *URL https://arxiv.org/abs/2205.13792*.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2023. A survey on large language model based autonomous agents. *arXiv preprint arXiv:2308.11432*.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A Smith. 2023. How language model hallucinations can snowball. *arXiv preprint arXiv:2305.13534*.

A Quantitative Analysis of Answer-Related Tokens

We present detailed analysis of *answer-related* tokens' attention score distribution (or perplexity distribution of *Perplexity Distillation Training*) shown in Table 3, Figure 4, and Figure 5.

B Quantitative Analysis of Question-Related Tokens

We present detailed analysis of *question-related* tokens' attention score distribution (or perplexity distribution of *Perplexity Distillation Training*) shown in Table 4, Figure 6, and Figure 7.

C Dataset Statistics

For the *NaturalQuestions* dataset, we split it according to the number of 79168/8757/3610 to form the train/validation/test dataset; for the *TriviaQA* dataset, we split it according to the number of 78785/8837/11313 to form the train/validation/test dataset.

D Implementation Details

We conducted all computations on a Nvidia A100 GPU. For the *Off-the-shelf Distillation Training* and the *Fine-tuned Distillation Training*, we use *Falcon-1b* as the initial reader model and *Contriever* as the initial retriever model, which have about 1 billion and 110 millions training parameters respectively. For the *Atlas-large Distillation Training* and *Perplexity Distillation Training*, we use *T5-large* as the initial reader model and *Contriever* as the initial retriever model, which have about 770 millions and 110 millions training parameters respectively.

Off-the-shelf Distillation Training We set the batch size to 1, the maximum length of the input prompt to 128 and limit the generation max length to 32. We set the learning rate to 1e-5 and use Adam optimizer. For *NaturalQuestions* dataset, we set the total training steps to 160,000 with approximately 2000 warmup steps, training for about 40 hours. For *TriviaQA* dataset, we set the total training steps to 320,000 with approximately 4000 warmup steps, training for about 60 hours.

Fine-tuned Distillation Training For Step 1, we set the batch size to 1, the maximum length of the input prompt to 128 and limit the generation max length to 32. We set the learning rate to 1e-5 and use Adam optimizer. For *NaturalQuestions* dataset, we set the total training steps to 160,000 with approximately 2000 warmup steps, training for about 30 hours. For *TriviaQA* dataset, we set the total training steps to 320,000 with approximately 4000 warmup steps, training for about 45 hours.

For Step 2, we set the batch size to 1, the maximum length of the input prompt to 128 and limit the generation max length to 32. We set the learning rate to 5e-7 and use Adam optimizer. For *NaturalQuestions* dataset, we set the total training steps to 6,000 with approximately 300 warmup steps, training for about 2 hours. For *TriviaQA* dataset, we set the total training steps to 32,000 with approximately 600 warmup steps, training for about 3 hours.

Atlas-large Distillation Training We set the batch size to 1, the maximum length of the input prompt to 128 and limit the generation max length to 32. We set the learning rate to 4e-5 and use Adam optimizer. For *NaturalQuestions* dataset, we set the total training steps to 10,000 with approximately 500 warmup steps, training for about 20 hours. For *TriviaQA* dataset, we set the total training steps to 30,000 with approximately 600 warmup steps, training for about 40 hours.

Perplexity Distillation Training We set the batch size to 1, the maximum length of the input prompt to 128 and limit the generation max length to 32. We set the learning rate to 4e-5 and use Adam optimizer. For *NaturalQuestions* dataset, we set the total training steps to 20,000 with approximately 1000 warmup steps, training for about 40 hours. For *TriviaQA* dataset, we set the total training steps to 10,000 with approximately 500 warmup steps, training for about 15 hours.

Table 3: Mean and std. of attention scores (or perplexity distribution in *Perplexity Distillation Training*) and the Spearman correlations of the answer-related tokens.

Experiment	Dataset	Avg. Attn. (p90)	Spearman Corr. (p90)	Avg. Attn. (p95)	Spearman Corr. (p95)
Off-the-shelf Model Checkpoint	NQ	0.033 ±0.016	0.227 ±0.259	0.039 ±0.023	0.196 ±0.349
	TriviaQA	0.027 ±0.013	0.218 ±0.252	0.032 ±0.019	0.206 ±0.331
Off-the-shelf Attention Distillation	NQ	0.017 ±0.008	0.145 ±0.193	0.017 ±0.010	0.076 ±0.254
	TriviaQA	0.031 ±0.012	0.160 ±0.174	0.035 ±0.017	0.172 ±0.236
Fine-tuned Distillation Training (Step1)	NQ	0.039 ±0.023	0.308 ±0.276	0.052 ±0.036	0.282 ±0.336
	TriviaQA	0.058 ±0.031	0.259 ±0.261	0.074 ±0.050	0.258 ±0.331
Fine-tuned Distillation Training (Step2)	NQ	0.049 ±0.023	0.316 ±0.280	0.066 ±0.036	0.350 ±0.336
	TriviaQA	0.069 ±0.036	0.290 ±0.267	0.089 ±0.061	0.320 ±0.323
Atlas-large Distillation Training	NQ	0.062 ±0.036	0.171 ±0.462	0.083 ±0.058	0.307 ±0.471
	TriviaQA	0.072 ±0.045	0.141 ±0.379	0.091 ±0.067	0.217 ±0.438
Perplexity Distillation Training	TriviaQA	0.072 ±0.039	0.029 ±0.142	0.071 ±0.042	0.013 ±0.202

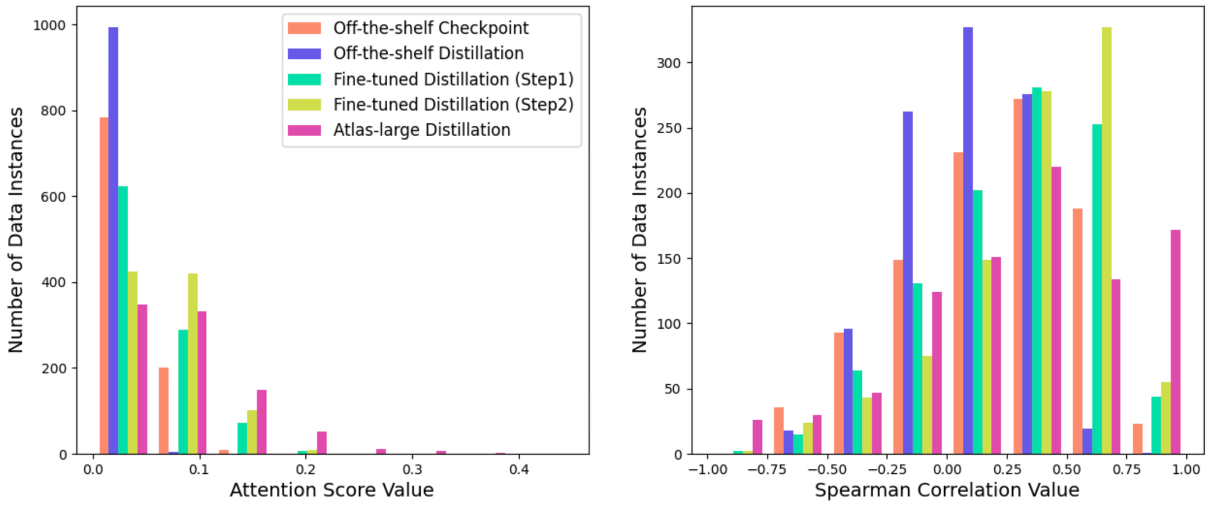

Figure 4: The attention score distribution histogram (left) and Spearman correlation distribution histogram of 95^{th} percentile *answer-related* tokens under NQ dataset.

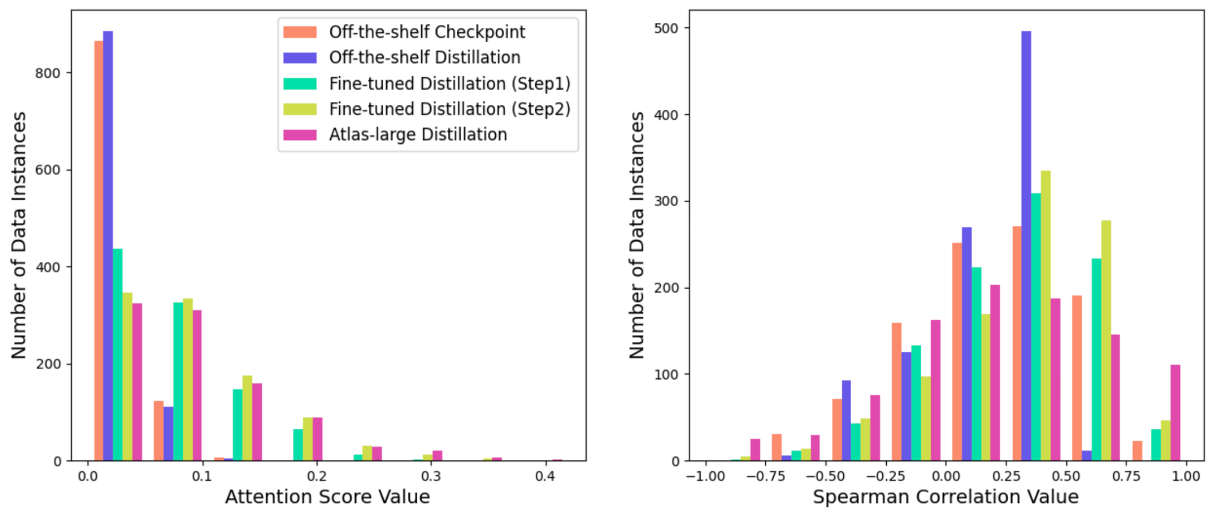

Figure 5: The attention score distribution histogram (left) and Spearman correlation distribution histogram of 95^{th} percentile *answer-related* tokens under TriviaQA dataset.

Table 4: Mean and std. of average attention scores (or perplexity distribution in *Perplexity Distillation Training*) and Spearman correlations of the question-related tokens

Experiment	Dataset	Avg. Attn. (p90)	Spearman Corr. (p90)	Avg. Attn. (p95)	Spearman Corr. (p95)
Off-the-shelf Model Checkpoint	NQ	0.023 ±0.011	0.103 ±0.253	0.024 ±0.014	0.092 ±0.309
	TriviaQA	0.021 ±0.010	0.103 ±0.245	0.023 ±0.013	0.067 ±0.304
Off-the-shelf Attention Distillation	NQ	0.027 ±0.010	0.139 ±0.237	0.039 ±0.017	0.153 ±0.341
	TriviaQA	0.047 ±0.016	0.144 ±0.220	0.063 ±0.025	0.260 ±0.280
Fine-tuned Distillation Training (Step1)	NQ	0.035 ±0.015	0.343 ±0.238	0.045 ±0.023	0.333 ±0.303
	TriviaQA	0.058 ±0.024	0.349 ±0.222	0.078 ±0.037	0.372 ±0.285
Fine-tuned Distillation Training (Step2)	NQ	0.032 ±0.014	0.310 ±0.256	0.039 ±0.021	0.225 ±0.340
	TriviaQA	0.060 ±0.025	0.367 ±0.227	0.078 ±0.037	0.326 ±0.311
Atlas-large Distillation Training	NQ	0.037 ±0.027	0.082 ±0.251	0.038 ±0.032	0.086 ±0.345
	TriviaQA	0.047 ±0.245	0.076 ±0.249	0.050 ±0.038	0.081 ±0.348
Perplexity Distillation Training	TriviaQA	0.063 ±0.038	-0.012 ±0.207	0.060 ±0.042	-0.036 ±0.297

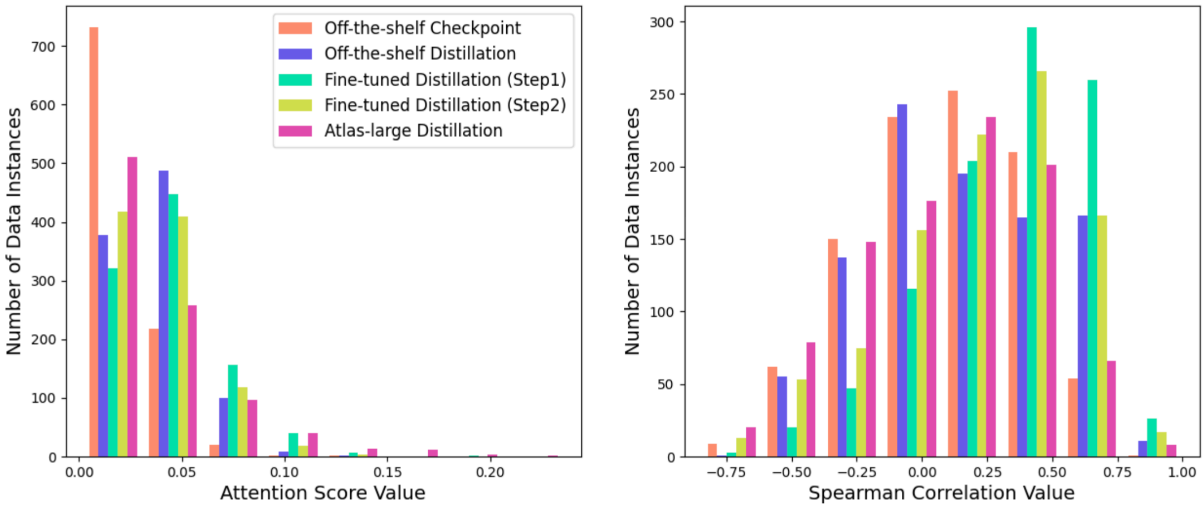

Figure 6: The attention score distribution histogram (left) and Spearman correlation distribution histogram of 95^{th} percentile *question-related* tokens under NQ dataset.

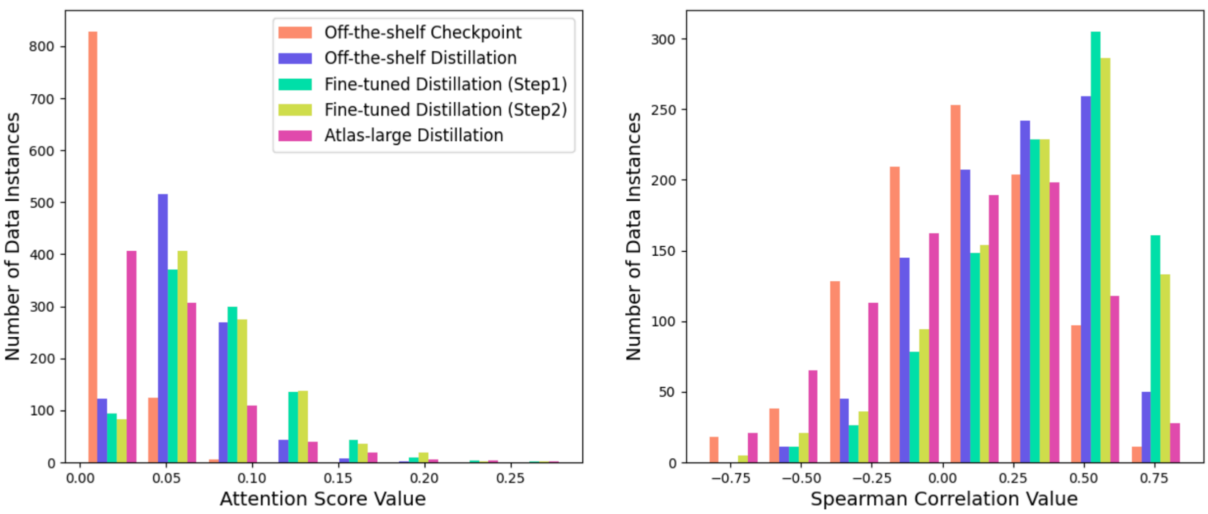

Figure 7: The attention score distribution histogram (left) and Spearman correlation distribution histogram of 95^{th} percentile *question-related* tokens under TriviaQA dataset.