YNU-HPCC at SemEval-2026 Task 12: Retrieval-Guided Reasoning with Teacher Distillation for Abductive Event Reasoning

Yuwei Sun, Jin Wang and Xuejie Zhang
School of Information Science and Engineering
Yunnan University
Kunming, China
Contact: sunyuwei@stu.ynu.edu.cn, {wangjin, xjzhang}@ynu.edu.cn

Abstract

This paper describes the YNU-HPCC system for SemEval-2026 Task 12, Abductive Event Reasoning (AER). Given multi-document retrieved evidence with distractors, the task requires selecting all direct-cause options for a target event and outputting an answer set. The main challenges are sparse and dispersed evidence in long documents and a boundary-sensitive set-level evaluation. This paper proposes a two-stage framework. Stage 1 trains a DeBERTa-v3-base student with retrieval-guided evidence modeling: documents are split into overlapping windows, BM25 ranks and filters candidate windows, and Top-K pooling aggregates window-level scores into option probabilities. Stage 2 distills soft targets from a Qwen-14B teacher with temperature scaling and high-confidence filtering to reduce pseudo-label noise and improve generalization. The system achieves an official dev score of 0.9712 (micro-F1 0.9746, macro-F1 0.9745) and improves the test score from 0.46 to 0.73, ranking 84th out of 221 submissions.

1 Introduction

Causal inference and event explanation are core capabilities of natural language understanding, with broad applications in event attribution, decision support, and automated knowledge discovery. SemEval-2026 Task 12, Abductive Event Reasoning (AER), requires selecting all candidate options that can explain a target event given a topic and a set of retrieved evidence documents, and outputting the answer as a set. The official evaluation adopts a tiered scoring scheme: an exact match with the gold set receives 1 point, a non-empty strict subset receives 0.5 points, and all other predictions receive 0 points. This metric highlights the importance of accurately modeling the boundary of the correct explanation set. This retrieval-based setting makes AER substantially harder than short-context abductive inference: evidence may be sparse and dispersed across long documents, distractor documents can introduce substantial noise, and the set-level partial-matching evaluation is sensitive to controlling the decision boundary of the predicted answer set.

Prior work on abductive reasoning provides useful modeling foundations. Discriminative hypothesis scoring, commonly used in αNLI-style tasks, concatenates each candidate explanation with context and uses pretrained cross-encoders, such as BERT or RoBERTa, trained with ranking or contrastive objectives (Devlin et al., 2019; Bhagavatula et al., 2020). Generation-based approaches construct missing events or explanations and improve consistency with the context through constrained decoding or reranking (Qin et al., 2020). Knowledge-enhanced and prompt-based methods leverage external commonsense resources such as ATOMIC or generative completion models such as COMET, and may further improve robustness via prompting strategies and self-consistency sampling (Sap et al., 2019; Bosselut et al., 2019; Chan et al., 2023). However, these paradigms typically assume compact contexts or single-best selection, and therefore require additional design choices for evidence aggregation and boundary-sensitive set prediction under noisy retrieval.

To address these challenges, the system adopts a two-stage AER system. Stage 1 segments long-topic documents into overlapping windows and applies retrieval-guided Top-K aggregation to improve evidence coverage and robustness against distractors. Stage 2 distills soft supervision from a large teacher model to refine the student's boundary between partially correct and fully correct answer sets, using high-confidence filtering to reduce negative transfer from noisy pseudo supervision (Hinton et al., 2015; Qwen et al., 2025). Overall, this two-stage pipeline improves set prediction stability and cross-domain generalization across the official dev set and successive test submissions.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 206–212

2 Related Work

2.1 Sparse Retrieval for Long Evidence

Long-document and multi-document evidence reasoning is often framed as a retrieve-then-read pipeline: a lightweight retriever first narrows the evidence space, and a neural model then performs fine-grained discrimination over the selected candidates. In retrieval-augmented generation and related evidence-based reasoning settings, sparse retrievers remain strong and stable baselines, particularly when training data are limited or domain shift is substantial (Gupta et al., 2024; Jiang et al., 2025). This study uses BM25 to rank and select evidence windows because it is training-free, efficient, and robust under distribution shift (Gupta et al., 2024).

2.2 Data Augmentation

Large language model–driven augmentation can improve expression diversity and coverage, but its effectiveness depends on generation constraints and sample selection; otherwise, semantic drift and hallucinations may introduce label noise (Ding et al., 2024). The system adopts controlled rewriting that paraphrases only the target event and candidate options while keeping the gold labels unchanged, thereby expanding surface forms without altering the evidence distribution. The system further applies hard negative mining to expose the model to confusing distractors and sharpen the decision boundary; prior work suggests that harder negatives often provide stronger training signals than random negatives (Meghwani et al., 2025).

2.3 Knowledge Distillation and Pseudo-Labeling

When labeled data are limited or the evaluation domain shifts, knowledge distillation is commonly used to transfer information from a stronger teacher to a smaller student via soft targets, improving data efficiency and generalization (Hinton et al., 2015; Xu et al., 2024). In practice, distillation is often combined with pseudo-labeling or self-training, but its gains depend critically on pseudo-label quality; confidence-based filtering is therefore widely used to reduce noise and confirmation bias and improve stability (Li et al., 2025). Multi-stage distillation has also been explored to progressively improve pseudo-label consistency and mitigate error accumulation across rounds of pseudo supervision (Zhao et al., 2024).

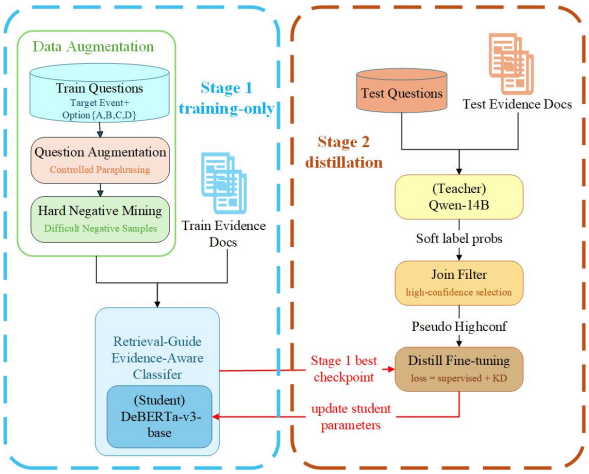

Figure 1: Retrieval-guided Two-stage Distillation Architecture

3 System Overview

The system is an evidence-aware classifier for set-valued abductive reasoning. This work formulates AER as multi-label prediction over four options and trains a two-stage pipeline: (i) supervised learning with retrieval-guided evidence aggregation, and (ii) distillation fine-tuning with a large teacher model and confidence filtering to improve cross-domain generalization (Li et al., 2025). Figure 1 summarizes the overall architecture.

3.1 Evidence Windowing and Selection

To avoid missing evidence in long documents, each topic document is segmented into overlapping windows. Candidate windows are ranked with BM25 using a query that concatenates the target event and all options, $q = [e; o_1; o_2; o_3; o_4]$, and only the Top-M windows are kept for encoding. To increase coverage under a fixed budget, near-duplicate windows are suppressed to encourage diversity (Wang et al., 2024).

3.2 Stage 1: Retrieval-guided Evidence-Aware Classifier

The student encoder is a DeBERTa-v3-base model (He et al., 2023). For each selected window ω and option o, the model produces a window-level logit $s_{\omega,o}$. Evidence is then aggregated across windows with Top-K mean pooling:

$$\hat{s}_o = \frac{1}{K} \sum_{\omega \in \mathrm{TopK}(\{s_{\omega,o}\})} s_{\omega,o} \quad (1)$$

$$p_o = \sigma(\hat{s}_o) \quad (2)$$

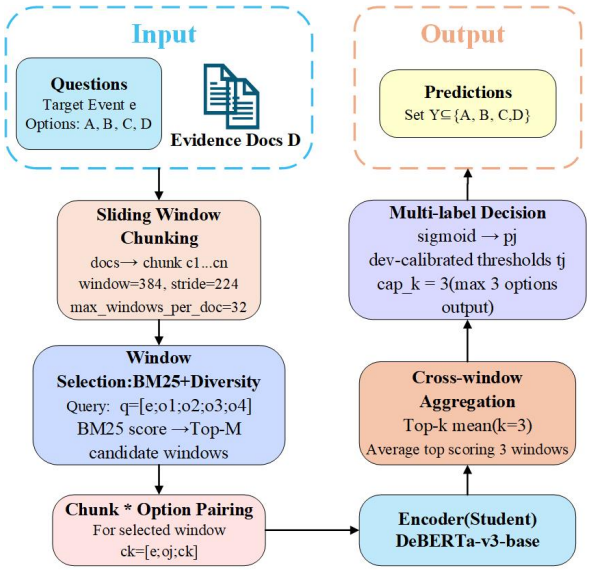

Figure 2: Retrieval-guided Evidence-Aware Classifier

where p_o is the probability that the option o is correct. The student is trained with multi-label supervision and uses Focal Loss to emphasize hard cases. In addition, controlled paraphrasing of events/options and hard negative mining are applied to strengthen boundary learning under distractor-like conditions. The specific details of the Retrieval-guided Evidence-Aware Classifier are shown in Figure 2.

3.3 Stage 2: Distillation Fine-tuning with a Large Teacher Model

While Stage 1 achieves near-saturated performance on the dev set, performance drops substantially on the test domain due to distribution shift. To improve generalization, Stage 2 introduces a Qwen-14B teacher model to produce soft targets and performs distillation fine-tuning on a high-confidence subset.

A KL-based distillation objective is minimized to match the student distribution to the teacher distribution:

$$L_{\mathrm{KD}} = T^2 \,\mathrm{KL}\left(p_T^{(t)} \,\middle\|\, p_T^{(s)}\right) \tag{3}$$

where $p_T^{(t)}$ is $\mathrm{softmax}(\mathbf{z}^{(t)}/T)$ and $p_T^{(s)}$ is $\mathrm{softmax}(\mathbf{z}^{(s)}/T)$. Since distillation quality depends on pseudo-label noise, High-confidence filtering is applied based on criteria such as maximum probability, the Top-1/Top-2 margin, and entropy (Li et al., 2025).

3.4 Inference and Decision Rule

At inference time, we follow the same pipeline (BM25 window selection, window–option scoring, and Top-K aggregation) to obtain option probabilities $\{p_o\}$. We calibrate a decision threshold on the dev set and select all options that exceed it as the predicted set. To prevent over-prediction, we apply a cap-K constraint on the maximum set size. If no option passes the threshold, we use $\arg\max_j p_j$ to ensure a non-empty output. This decision rule matches the set-based scoring mechanism and is convenient for stable parameter tuning on the development set.

4 Experiment Details

4.1 Datasets

The dataset used in this paper was created and released by the competition organizers. Each topic is associated with a set of retrieved external evidence documents and contains multiple-choice reasoning questions. Each question includes a target event, four candidate options, and a multi-label gold answer set.

Train and dev share the same topics and evidence documents, meaning they use the same evidence pool. In contrast, the test split has no overlap with dev in either topics or documents (topic overlap = 0, document overlap = 0), which introduces substantial cross-topic and cross-evidence-pool distribution shift. In addition, train/dev documents exhibit a strong long-tail length distribution, with the longest document reaching approximately 280,000 characters, further increasing the difficulty of evidence localization under multi-document long-context conditions.

4.2 Model Selection

DeBERTa-v3-base is used as the discriminative student model due to its strong encoding capacity and efficiency (He et al., 2023). For Stage-2, Qwen-14B is adopted as the teacher to provide soft targets for distillation under domain shift (Qwen et al., 2025). Qwen2.5-3B-Instruct is additionally used for controlled paraphrasing of target events and options for data augmentation.

4.3 Experimental Setup

All experiments use window-level encoding and cross-window aggregation after sliding-window segmentation to handle long-document evidence. We set max_len = 512, window = 384, and stride =

Table 1: Ablation Results on the Dev Set.

	Bert	DeBERTa	Sliding Window	Top-K	BCE	Focal	pos-weight	amp	R-drop	Official Score	micro-F1	macro-F1
A	✓									0.5437	0.6076	0.5200
B	✓		✓		✓					0.6069	0.6741	0.6525
C	✓		✓	✓	✓					0.7287	0.7534	0.7526
D	✓		✓	✓	✓			✓		0.7500	0.7876	0.7861
E	✓		✓	✓	✓		✓			0.7613	0.7992	0.7984
F		✓								0.5013	0.5404	0.4732
G		✓	✓		✓					0.6725	0.6731	0.6668
H		✓	✓	✓	✓					0.8258	0.8554	0.8535
I		✓	✓	✓	✓			✓		0.8759	0.8945	0.8926
J		✓	✓	✓	✓		✓			0.8576	0.8786	0.8755
K		✓	✓	✓		✓		✓		**0.9712**	**0.9746**	**0.9745**
L		✓	✓	✓		✓		✓	✓	0.9025	0.9381	0.9346

224, and cap the number of windows per question at max_windows = 32. Cross-window aggregation uses Top-K mean pooling with K=3. The predicted set size is capped at 3 to prevent over-prediction.

The training objective focuses on multi-label discrimination. The final system uses Focal Loss to emphasize hard cases and long-tail label patterns. (Lin et al., 2017).

4.4 Evaluation Metrics

The official evaluation adopts set-level partial-match scoring. Let G denote the gold set and P the predicted set. The per-instance score is defined as:

$$s(P,G) = \begin{cases} 1, & P = G, \\ 0.5, & \emptyset \neq P \subsetneq G, \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

The final official score is the average over all instances. In addition to the official score, micro-F1 and macro-F1 are also reported on the dev set to better characterize multi-label behavior. The distribution of predicted set sizes is also computed to diagnose overly conservative or overly aggressive set-output preferences.

5 Main Results and Analysis

5.1 Overall Results

The system follows a two-stage training scheme. In Stage 1, the evidence-aware DeBERTa-v3-base model achieves an official dev score of 0.9712 (micro-F1 = 0.9746, macro-F1 = 0.9745), compared to 0.5437 for the BERT single-choice baseline, yielding an absolute improvement of 0.3977. However, Stage 1 alone achieves only 0.46 on the test set. After introducing Stage-2 distillation with Qwen-14B soft labels and high-confidence filtering, the final test submission reaches 0.73, improving by 0.27 over the initial submission.

5.2 Ablation Study

Table 1 summarizes ablations on the dev set across two backbone routes, covering long-document processing, cross-window aggregation, training objectives, imbalance handling, and training techniques.

Comparing A and F, replacing BERT with DeBERTa without evidence modeling reduces the official score, from 0.5437 to 0.5013, suggesting that, under multi-document long-context settings, truncation-induced evidence loss can outweigh gains from a stronger encoder. DeBERTa's advantage becomes apparent only after evidence is sufficiently covered and organized (He et al., 2023).

Sliding-window segmentation substantially improves evidence coverage. From A to B and from F to G, adding sliding windows consistently increases the official score and macro-F1, indicating that improved coverage directly benefits multi-label discrimination. Top-K aggregation further reduces dilution from noisy windows: from B to C and from G to H, introducing Top-K pooling yields another large gain, consistent with the intuition that many windows are weakly relevant and aggregation should emphasize the most informative evidence.

Beyond structural evidence modeling, some training techniques have secondary or mixed effects. From C to D (and similarly from H to I), AMP provides modest but consistent gains, aligning with prior observations on mixed-precision training efficiency and numerical behavior (Micikevicius et al., 2018). From D to E, positive-class

Table 2: Test Submission Trajectory.

Submit	Crucial strategy	Test Score
1	Stage-1 supervised training	0.46
2	+ Augment-to-6000	0.57
3	+ Hard Negative-50	0.63
4	+ Stage-2 distillation	0.65
5	+ High-confidence filtering	**0.73**

reweighting offers limited improvements in the BERT route but can hurt in the DeBERTa route, reflecting a recall–overprediction trade-off that is penalized under set-level evaluation. From I to K, replacing BCE with focal loss produces the largest additional improvement on top of "DeBERTa + Sliding Window + Top-K," increasing the official score to 0.9712 and macro-F1 to 0.9745, indicating that focusing gradients on hard cases helps separate superficially plausible but unsupported distractors (Lin et al., 2017).

Finally, from K to L, adding R-Drop degrades performance in our setting. This may be due to the interaction between consistency regularization and window-level noise under small-batch training, which can suppress fitting to sparse key evidence (Liang et al., 2021).

5.3 Test Submission Trajectory

Because the number of official test submissions is limited, Table 2 reports a submission trajectory, where each submission adds one key strategy on top of the previous one. The Stage-1 model achieves 0.46 on test despite near-saturated dev performance, indicating a large distribution shift. From submission 1 to submission 2, controlled paraphrase augmentation increases the test score to 0.57, and from submission 2 to submission 3, adding hard negative training further improves it to 0.63, suggesting that richer surface forms and harder negatives improve robustness under domain shift (Wang et al., 2023).

Distillation benefits depend strongly on pseudo-label quality control. From submission 3 to submission 4, adding Stage-2 distillation yields only a small gain, from 0.63 to 0.65, implying that unfiltered pseudo supervision may contain noise that limits the benefit. In contrast, from submission 4 to submission 5, high-confidence filtering leads to the largest single-step improvement, from 0.65 to 0.73, indicating that filtering noisy pseudo labels is crucial to prevent negative transfer and stabilize student refinement (Hsieh et al., 2023; Qwen et al., 2025).

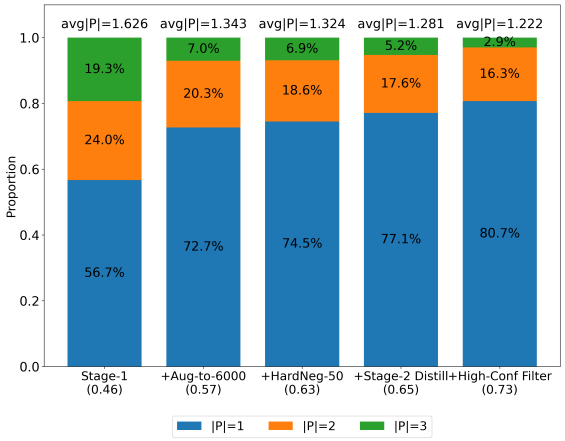

Figure 3: Prediction-set size distribution on test submissions.

Figure 3 shows the predicted set-size distribution across submissions. As training signals become stronger, predictions become more conservative: the proportion of single-option outputs increases while three-option outputs drop sharply, and the average set size decreases. This trend is consistent with the set-level metric, where over-prediction introduces extra incorrect options and reduces the chance of exact match; thus, suppressing over-prediction is beneficial for the final score.

6 Conclusion

This paper investigates long-document evidence aggregation and set-valued prediction for SemEval-2026 Task 12 (AER). This paper proposes a discriminative framework that combines sliding-window segmentation with Top-K aggregation, and the system further applies teacher distillation with confidence filtering to improve set-boundary calibration and cross-domain generalization. Future work will explore more precise evidence selection with cross-window consistency, incorporate commonsense knowledge to better handle implicit explanations, and improve pseudo-label quality control via uncertainty-aware filtering.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant Nos. 61966038 and 62266051. We would like to thank the anonymous reviewers for their constructive comments.

References

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. *International Conference on Learning Representations (ICLR)*.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779. Association for Computational Linguistics.

Chunkit Chan, Xin Liu, Tsz Ho Chan, Jiayang Cheng, Yangqiu Song, Ginny Wong, and Simon See. 2023. Self-consistent narrative prompts on abductive natural language inference. *IJCNLP-AACL 2023*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North*, pages 4171–4186. Association for Computational Linguistics.

Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Anh Tuan Luu, and Shafiq Joty. 2024. Data augmentation using LLMs: Data perspectives, learning paradigms and challenges. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 1679–1705. Association for Computational Linguistics.

Shailja Gupta, Rajesh Ranjan, and Surya Narayan Singh. 2024. A comprehensive survey of retrieval-augmented generation (RAG): Evolution, current landscape and future directions. *arXiv*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *ICLR 2023*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv*.

Cheng-Yu Hsieh, Chun-Liang Li, Chih kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017. Association for Computational Linguistics.

Pengcheng Jiang, Siru Ouyang, Yizhu Jiao, Ming Zhong, Runchu Tian, and Jiawei Han. 2025. Retrieval and structuring augmented generation with large language models. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, volume 2, pages 6032–6042. Association for Computing Machinery.

Juanhui Li, Sreyashi Nag, Hui Liu, Xianfeng Tang, Sheikh Muhammad Sarwar, Limeng Cui, Hansu Gu, Suhang Wang, Qi He, and Jiliang Tang. 2025. Learning with less: Knowledge distillation from large language models via unlabeled data. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2627–2641. Association for Computational Linguistics.

Xiaobo Liang, Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, and Tie-Yan Liu. 2021. R-drop: Regularized dropout for neural networks. *arXiv*.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. 2017. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007. IEEE.

Hansa Meghwani, Amit Agarwal, Priyaranjan Pattnayak, Hitesh Laxmichand Patel, and Srikant Panda. 2025. Hard negative mining for domain-specific retrieval in enterprise systems. *ACL 2025*.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. 2018. Mixed precision training. *ICLR 2018*.

Lianhui Qin, Vered Shwartz, Peter West, Chandra Bhagavatula, Jena D. Hwang, Ronan Le Bras, Antoine Bosselut, and Yejin Choi. 2020. Back to the future: Unsupervised backprop-based decoding for counterfactual and abductive commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 794–805. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *arXiv*.

Maarten Sap, Ronan LeBras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. ATOMIC: An atlas of machine commonsense for if-then reasoning. *Proceedings of AAAI 2019*.

Xiaohua Wang, Zhenghua Wang, Xuan Gao, Feiran Zhang, Yixin Wu, Zhibo Xu, Tianyuan Shi, Zhengyuan Wang, Shizheng Li, Qi Qian, Ruicheng Yin, Changze Lv, Xiaoqing Zheng, and Xuanjing Huang. 2024. Searching for best practices in retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17716–17736. Association for Computational Linguistics.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. *arXiv*.

Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. 2024. A survey on knowledge distillation of large language models. *arXiv*.

Jiachen Zhao, Wenlong Zhao, Andrew Drozdov, Benjamin Rozonoyer, Md Arafat Sultan, Jay-Yoon Lee, Mohit Iyyer, and Andrew McCallum. 2024. Multistage collaborative knowledge distillation from a large language model for semi-supervised sequence generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 14201–14214. Association for Computational Linguistics.