Tri-Level Relevance Learning with Ordered Ranking Constraints for Compulsory Education Quality Evaluation

Hua Yang[a], Rong Cao[b], Yang Shen[c,*], Meijie Shi[d], Xujing Liu[e]

Collaborative Innovation Center of Assessment for Basic Education Quality, Beijing Normal University, Beijing, China

[a]987754124@qq.com, [b]18813192326@163.com, [c,*]shenyang@bnu.edu.cn, [d]smj_1999@126.com, [e]liuxujing221@163.com

Abstract

Recent advances in large language models (LLMs) have made retrieval-augmented generation (RAG) an effective means of enhancing LLMs by retrieving relevant documents from external knowledge bases, thereby improving retrieval accuracy. However, conventional binary retrieval cannot distinguish partially relevant evaluation texts-such as indicator-relevant but dimension-mismatched passages or policy fragments with only partial usability-leading to distorted evaluation conclusions and misaligned indicator matching. This paper proposes AAR-TRL, a tri-level relevance retrieval-augmented method for compulsory education quality evaluation, and validates it on a dedicated dataset covering indicator-definition and policy-basis queries. It leverages the cross-attention scores of the FiD fusion decoder to construct an automated tri-level labeling system that partitions documents into fully supportive, partially supportive, and non-supportive categories, and jointly optimizes the list-wise InfoNCE loss and ordered pairwise loss to sharpen fine-grained discrimination among easily confused evaluation documents. On our constructed compulsory education quality evaluation dataset, we evaluate AAR-TRL with Precision@10, Recall@10, and MRR. AAR-TRL attains 60.28%, 82.91%, and 0.6528 on the three metrics, respectively; compared with the vanilla AAR, it improves Precision@10 by 11.2 percentage points, Recall@10 by 6.82 percentage points, and MRR by 0.024; compared with ARL2, the improvements are 3.11 percentage points, 1.37 percentage points, and 0.012, respectively. In generation quality evaluation, AAR-TRL achieves an indicator matching accuracy of 87.0% and a policy citation validity of 82.0%, both superior to the baseline methods on the same test queries. These results, obtained within the compulsory education stage, provide preliminary evidence of AAR-TRL's effectiveness in retrieval and question-answering generation for compulsory education quality evaluation; its generalizability to other education stages remains to be validated.

Keywords

Retrieval-Augmented Generation, tri-level relevance, Adaptive Retriever, Large Language Model, compulsory education quality evaluation.

1. Introduction

Educational evaluation is vital to the direction of education development and the overall goal of building a education powerhouse. The implementation of the Overall Plan for Deepening the Reform of Educational Evaluation in the New Era and the release of the Plan Outline for Building an Education Powerhouse (2024–2035) have provided a clear policy framework for educational evaluation reform. Using modern information technologies such as artificial intelligence, big data, and large language models (LLMs) to innovate evaluation tools, so as to

enhance the scientific validity, professionalism, and objectivity of evaluation, has become a basic consensus[1]. However, educational quality evaluation in the compulsory education stage faces two specific challenges that constrain the effectiveness of intelligent question-answering (QA) systems. First, the concepts of evaluation, assessment, examination, and monitoring are ambiguously oriented, unclearly positioned, and associated with diverse indicators, which significantly affects indicator selection and interpretation, indicator-dimension misalignment, concept definition, and weight assignment. Second, national educational policy documents contain little content on educational quality evaluation, often only one paragraph, giving rise to the ambiguous granularity of partially usable policy provisions.

Against this backdrop, applying LLMs to education quality evaluation faces three distinctive challenges. First, insufficient depth of domain expertise: education quality evaluation relies heavily on subject-specific teaching standards, evaluation indicators, and expert experience, which general-purpose LLM training corpora hardly cover. Second, the dynamism of standards: education policies, curriculum standards, and evaluation indicators evolve continuously, whereas the static training data of LLMs cannot reflect the latest requirements in a timely manner. Third, the context dependency of evaluation: the same evaluative act may carry markedly different meanings across school stages, subjects, and regions, requiring the model to flexibly invoke appropriate reference information[2].

Retrieval-Augmented Generation (RAG) offers a promising technical route to these challenges: during evaluative reasoning, the LLM retrieves relevant evaluation standards, historical cases, and professional knowledge from an external educational knowledge base to supplement and calibrate its generation. Yet intelligent question-answering systems remain limited in compulsory education quality evaluation. Traditional retrieval systems (e.g., keyword matching, BM25) can hardly capture subject semantics and indicator associations, yielding low retrieval precision; general-purpose LLMs possess certain reasoning capability but suffer from knowledge staleness and hallucination, and cannot cite authoritative evidence; dense retrievers (e.g., DPR, Contriever), trained on corpora detached from this domain, are misaligned with LLMs, and their retrieval results are underutilized. In practice, when a teacher asks "how to evaluate students' mathematics academic proficiency rate under the new curriculum standards", a conventional retriever returns documents such as a "mathematical literacy evaluation framework" (dimensional misalignment) and a "provincial curriculum arrangement policy" (only locally applicable), feeding them to the LLM as equally relevant and driving its conclusions away from compulsory education evaluation standards. The essence of this phenomenon lies in the inability of existing retrieval systems to identify educational texts that are partially relevant yet dimensionally misaligned.

To address these issues, this paper proposes AAR-TRL, a tri-level relevance adaptive retrieval-augmented method for compulsory education quality evaluation. The main contributions are threefold. (1) We analyze the multi-level usefulness of retrieved documents in compulsory education evaluation scenarios and design an automated tri-level labeling scheme based on the distribution of FiD (Fusion-in-Decoder) cross-attention feature scores, categorizing evaluation knowledge documents into fully supportive, partially supportive, and non-supportive levels, thereby alleviating the prohibitive cost of fine-grained annotation in this domain. (2) We introduce a combined training objective of list-wise InfoNCE loss and ordered pairwise loss, enabling the retriever to discriminate documents of graded relevance. (3) The method mitigates indicator-dimensional misalignment and the partial applicability of policy provisions in compulsory education evaluation retrieval, improving the stability and precision of question answering on our constructed dataset.

2. Related Work

2.1. Research on RAG

Existing research on RAG centers on retriever training paradigms, which fall into four categories. The first is the traditional term-frequency paradigm, exemplified by BM25. The second is the general dense retrieval paradigm, as in DocReLM[3]. The third is the LLM-preference-based alignment paradigm: Yu et al. proposed the augmentation-adapted retriever (AAR), which labels LLM-preferred documents via cross-attention scores of a small source language model, combines human-annotated documents to construct the training set, and mines hard negatives with ANCE, mitigating the tight coupling between conventional retrievers and LLMs[4]; Zhang et al. proposed ARL2, which innovatively employs the LLM itself as a direct labeler together with an adaptive self-training strategy to reduce annotation cost, overcoming the reliance on indirect supervision signals and the difficulty in distinguishing "similar yet useless" documents. The fourth is the black-box alignment paradigm[5]: Shi et al. proposed a black-box retrieval augmentation scheme that builds a KL loss from the generation probabilities of the target LLM; it aligns well but requires a large source LM for supervision, entails prohibitive API costs, and each retriever adapts to only a single LLM[6].

However, none of these methods designs a dedicated ranking constraint for dimensionally misaligned, partially supportive documents in educational evaluation. Although ARL2 introduces tri-level relevance labeling, it does not explicitly enforce an ordered ranking across the three levels; AAR relies on ANCE for negative mining and thus cannot guarantee the sampling of weak negatives with dimensional misalignment.

2.2. Research on RAG in Educational Evaluation

In educational evaluation, R. et al. built an RAG-based LLM chatbot configured with structured rubric instructions to automatically screen health-equity content in nursing course slides, validating the feasibility of AI-assisted curriculum evaluation[7]. Cai et al. proposed a knowledge-graph QA model improving entity detection and intent recognition for automatic control[8]. Yang et al. combined pruning, knowledge-distillation lightweighting, and distributed mixed-precision training with knowledge-graph enhancement to deploy large models in education[9]. Liu et al. adopted an OBE + LoRA-finetuned LLM + hybrid RAG framework to overcome single-source educational graphs and weak extraction, QA, and reasoning[10]. Lin optimized LightRAG with a vector–graph two-tier retrieval to reduce hallucination and improve knowledge extraction accuracy in textbook question answering for primary and secondary schools[11]. Zhang et al. proposed parallel text segmentation plus three-stage knowledge-graph retrieval to remedy inaccurate computer science course question answering and single-dimensional retrieval. Although these studies tackle educational question answering, their retrievers still follow generic semantic matching or knowledge-graph augmentation, lacking dedicated optimization for partially confusing documents[12].

In summary, prior work has contributed substantially to RAG-based retrieval in educational evaluation, yet gaps remain. First, existing LLMs are poorly adapted to educational evaluation question answering, yielding unstable and imprecise answers that fail to meet the subjectivity of evaluative judgment. Second, general RAG optimization schemes are tightly coupled and lack cross-scenario reusability, causing content redundancy, context deficiency, and hallucination; some are costly and model-specific, limiting generality. Third, vertical-domain solutions remain fragmented, with retrievers unaligned with LLM preferences, so that retrieved results, though semantically relevant, suit generation poorly and support educational decision-making inefficiently. Accordingly, this paper proposes a tri-level relevance adaptive retrieval-augmented method driven by LLM preferences to address these gaps and improve answer precision in compulsory education quality evaluation.

3. Framework of the Tri-Level Relevance Enhanced Retriever

Building upon the augmentation-adapted retriever, our retrieval augmentation module extends the native binary attention scoring to a tri-level relevance scheme, in which documents are labeled as fully supportive (1), partially supportive (0.5), or non-supportive (0). Fine-tuning with these labels yields a plug-and-play retrieval adapter compatible with the DeepSeek model API.

The overall pipeline comprises an offline retriever training stage and an online inference stage. Offline, a compact encoder–decoder source model, Flan-T5, extracts FiD cross-attention features to partition domain documents into three grades and generate soft labels; these are merged with manually annotated tri-level domain samples into a hybrid training set, and the retriever is jointly optimized with the list-wise InfoNCE loss and the ordered pairwise loss to strengthen its discrimination of partially relevant weak negatives. Online, the trained retriever serves as a standalone plugin: the LLM requires no fine-tuning, and only the retrieved layered evidence is concatenated into the user prompt for retrieval-augmented generation. For the underlying retrieval unit, we employ a dual-encoder dense architecture that encodes queries and documents separately, and construct the global vector index upon the FAISS vector database.

Table 1. Sample examples for the indicator-query scenario

Query	Relevance level	Document example	Rationale
Definition of the academic proficiency rate	Fully supportive	The academic proficiency rate, i.e., the proportion of students meeting the curriculum-standard baseline within a given school stage, is a core indicator for monitoring compulsory education quality.	Provides the complete and precise definition of the queried indicator; directly usable as an evaluation basis
	Partially supportive	The academic level evaluation system comprises three statistical dimensions-proficiency rate, excellence rate, and average score-reflecting students' overall learning status.	Mentions concepts related to the proficiency rate without giving the complete definition; of local reference value only
	Non-supportive	Curriculum arrangement in compulsory education shall balance foundational and comprehensive goals and allocate arts and physical education hours evenly.	Concerns curriculum arrangement and is entirely unrelated to the queried indicator definition

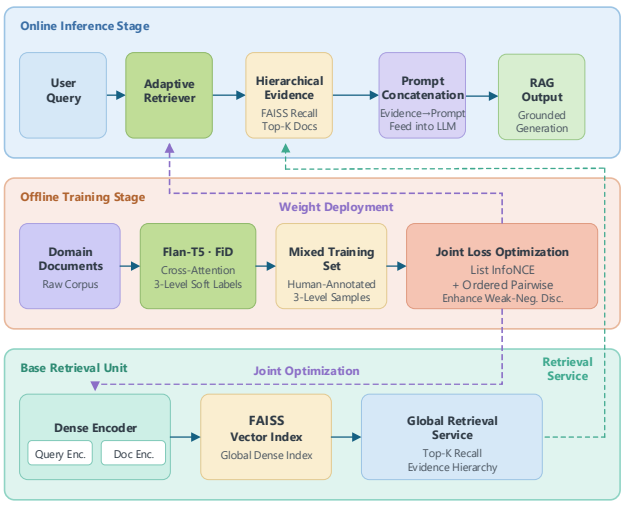


Fig. 1 Pipeline of retriever training and inference.

As shown in 0, in the offline stage, Flan-T5 extracts FiD cross-attention features for query–document pairs, which dual thresholds partition into tri-level soft labels. Combined with human-annotated samples, these form a hybrid training set that jointly optimizes the bge-m3 retriever via listwise InfoNCE and ordered pairwise losses. In the online stage, the query retrieves Top-k documents from the FAISS index via the trained retriever; the documents, arranged by supportiveness, are appended to the prompt for DeepSeek to generate answers.

3.1. Tri-Level Label Construction

As noted in the Introduction, retrieval for compulsory education quality evaluation faces two problems. First, ambiguously delineated concepts-evaluation, assessment, examination, and monitoring-and heterogeneous indicators leave many documents lexically relevant to a query yet misaligned in evaluation dimension. Second, national policy documents often contain only locally usable fragments on education quality evaluation, so that entire documents are only partially relevant to the query topic. These two document types constitute the identification blind spots of conventional binary relevance labeling and blur retrieval boundaries. This section therefore introduces a tri-level labeling scheme (fully supportive/partially supportive/non-supportive) that explicitly models this intermediate state at the label level. We consolidate corpora on compulsory education quality evaluation and use LLMs to generate diverse evaluation queries covering two scenarios-indicator queries (e.g., "what is the definition of the academic proficiency rate") and policy queries (e.g., "what dimensions does education quality evaluation comprise"), with 3-5 pseudo-queries per document to enrich training diversity. Table 1 presents sample examples for the indicator-query scenario.

For easily confounded educational indicators and concepts (e.g., "academic proficiency rate" vs. "literacy proficiency rate", "reliability" vs. "validity", "assessment" vs. "evaluation"), query–document pairs that are similar yet semantically distinct are manually selected as hard negatives to sharpen the retriever's discrimination; data quality is controlled by combining LLM automatic annotation with manual spot checks.

A lightweight Flan-T5 Base serves as the offline source model. Under the FiD (Fusion-in-Decoder) architecture, for each retrieved document d_i, the cross-attention is averaged over all layers, all attention heads, and all input tokens to obtain the document contribution score:

$$S_i^a = \frac{1}{ln \cdot hn \cdot tn} \sum_{layers} \sum_{heads} \sum_{t \in d_i \oplus q} FiDAtt(FiD(q)) \quad (1)$$

where ln denotes the number of encoder layers, hn the number of attention heads, and tn the total number of tokens of the concatenated document input; a larger value of (S_i^a) indicates stronger support of the document for answer generation by the LLM.

Two thresholds τ_1, τ_2 partition the attention score S_i^a into segments, yielding tri-level relevance soft labels $s \in \{1, 0.5, 0\}$. Specifically, when $S_i^a \geq \tau_1$ the label is $s = 1$, indicating full support: the document contains all core professional evidence required to answer the query and serves as a positive sample; when $\tau_2 < S_i^a < \tau_1$ the label is $s = 0.5$, indicating partial support: the document matches only domain keywords and lacks key inferential information, serving as a weak negative; when $S_i^a \leq \tau_2$ the label is $s = 0$, indicating no support: the document is topically irrelevant and serves as a hard negative. Table 2 summarizes the tri-level relevance labels.

3.2. Loss Function Design

An ordered pairwise loss is introduced to form a joint dual-loss training objective that explicitly constrains the similarity ranking of the three document levels and precisely separates partially relevant weak negatives. The total loss is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{list} + \alpha \cdot \mathcal{L}_{pair} \quad (2)$$

α is the weight hyperparameter of the pairwise loss, modulating the strength of the hierarchical ranking constraint.

Table 2. Tri-level relevance labels

Level	Condition	Labels	Description
Fully supportive	$S_i^a \geq \tau_1$	1.0	Contains all core professional evidence required to answer the query; positive sample
Partially supportive	$\tau_2 < S_i^a < \tau_1$	0.5	Matches only domain keywords and lacks key inferential information; weak negative
Non-supportive	$S_i^a \leq \tau_2$	0.0	Topically irrelevant; hard negative

The list-wise InfoNCE loss $\mathcal{L}_{\text{list}}$ treats all irrelevant and partially relevant documents as negatives and maximizes the vector distance between the fully supportive positive and the remaining documents, retaining the standard contrastive learning framework:

$$\mathcal{L}_{list} = -\sum_q log \frac{exp(sim(q,d^+)/\tau)}{exp(sim(q,d^+)/\tau)+\sum_{d^0,d^{0.5}} exp(sim(q,d)/\tau)} \quad (3)$$

sim(·) denotes the cosine similarity between query and document embeddings, and τ is the temperature coefficient; this loss keeps fully relevant documents ranked toward the top overall. The ordered pairwise loss $\mathcal{L}_{\text{pair}}$ (the core modification of this work) adds a hierarchical constraint that forces the retriever to satisfy the similarity ordering $\text{sim}(q, d^1) > \text{sim}(q, d^{0.5}) > \text{sim}(q, d^0)$.

For any triple $(d^1, d^{0.5}, d^0)$, two pairwise losses are constructed:

$$\mathcal{L}_{pair1} = max(0, sim(q, d^{0.5}) - sim(q, d^1) + \gamma) \quad (4)$$

$$\mathcal{L}_{pair2} = max(0, sim(q, d^0) - sim(q, d^{0.5}) + \gamma) \quad (5)$$

$$\mathcal{L}_{pair} = \mathcal{L}_{pair1} + \mathcal{L}_{pair2} \quad (6)$$

γ is the margin that lower-bounds the similarity gap between adjacent levels. Whenever a weak negative attains higher similarity than the positive, or an irrelevant document higher than a partially relevant one, the loss yields gradients that update the retriever, progressively suppressing the premature ranking of weakly interfering documents in the vertical domain.

4. Experiments and Analysis

4.1. Experimental Setup

No public benchmark dataset currently targets the vertical scenario of education quality evaluation question answering. We therefore construct a dedicated dataset from publicly available materials on compulsory education quality evaluation in China (indicator definitions and policy bases), including: (1) educational evaluation reform policy documents issued by the Ministry of Education and provincial education departments (e.g., the Overall Plan for Deepening Educational Evaluation Reform in the New Era and the Guidelines for Compulsory

Education Quality Evaluation); (2) public indicator specifications and evaluation standards (e.g., the Compulsory Education Quality Evaluation Indicators and the Statistical Indicator System for China Education Monitoring and Evaluation); and (3) academic literature on compulsory education quality evaluation in core education journals. Following the tri-level labeling scheme in Section 3.1, initial tri-level relevance labels (fully supportive/partially supportive/non-supportive) are generated. The dataset contains about 12,000 query–document pairs, covering indicator-definition queries (58%) and policy-basis queries (42%). After anonymization, the data are split into training, test, and validation sets at an 8:1:1 ratio. To avoid evaluation bias arising from a shared labeling mechanism between the training and test sets, we adopt differentiated labeling: the training set (~9,600 pairs) is labeled by FiD cross-attention automatic annotation combined with manual spot checks to reduce cost, whereas the test (~1,200 pairs) and validation (~1,200 pairs) sets are independently annotated by three domain experts, ensuring that the test set faithfully reflects human judgment.

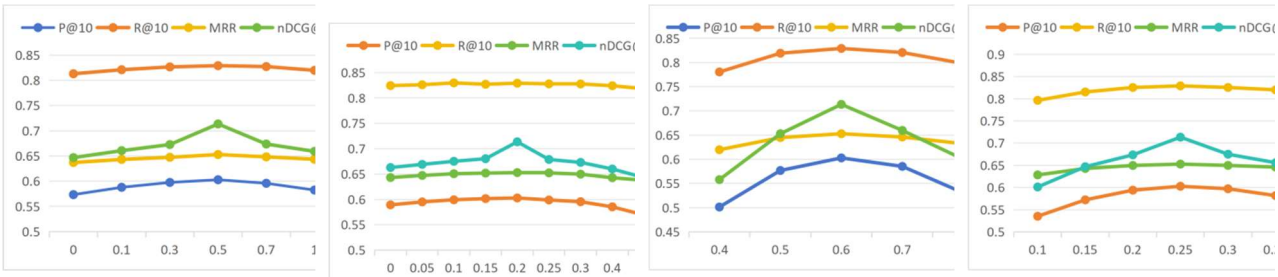
We design baseline comparison experiments, ablation studies, and a human evaluation of generation quality. Precision@10, Recall@10, nDCG@10, and MRR serve as evaluation metrics. Precision@k (P@k; k = 10 throughout) is the top-k precision: for a single query, the number of fully supportive documents (s = 1) among the top-k retrieved divided by k. Recall@k (R@k) is the top-k recall: the number of fully supportive documents among the top-10 results divided by the total number of fully supportive documents for the query in the entire knowledge base. MRR is the mean reciprocal rank: for each query, the reciprocal of the rank r of the first fully supportive document, taken as 1/r, or 0 if none exists. nDCG@k is a graded metric that treats the tri-level labels as relevance grades (fully supportive = 2, partially supportive = 1, non-supportive = 0), accumulates the discounted gains, and normalizes by the ideal DCG; it reflects the model's ability to discriminate partially supportive documents and avoids the information loss incurred by treating them all as negatives. All metrics are averaged over all test queries.

4.2. Comparison with Baseline Retrieval Methods

Six representative retrieval methods serve as baselines: the term-frequency method BM25, the pre-trained dense retriever Contriever, the black-box-LLM retrieval alignment framework RePlug, the LLM-preference-based augmented retriever AAR, the LLM-self-labeled ARL2, and the domain-finetuned retriever DocReLM. For our AAR-TRL, the retrieval backbone is BAAI/bge-m3, the offline labeling source model is Flan-T5 Base, and the target LLM is DeepSeek-V4. The tri-level thresholds are set to $\tau_1 = 0.6, \tau_2 = 0.25$, the joint loss weight to $\alpha = 0.5$, and the margin to $\gamma = 0.2$; threshold sensitivity is analyzed in Section 4.4. Training uses a single RTX 6000 D GPU (84 GB) with batch size 16 and learning rate $1e-5$, for 10 epochs, retaining the weights optimal on the validation set.

As Table 3 shows, AAR-TRL attains the best results on all metrics. Relative to ARL2, P@10 improves by 3.11 percentage points, R@10 by 1.37, nDCG@10 by 4.79, and MRR by 0.012. Five independent runs confirm that the improvements over all baselines are stable. Overall, AAR-TRL achieves the best retrieval purity, key-indicator recall, and front-loading of high-quality documents, effectively filtering partially relevant interfering documents that cross school stages and evaluation dimensions in educational scenarios.

4.3. Ablation Study



(a) Joint loss weight α (b) Margin γ (c) Tri-level threshold τ_1 (d) Tri-level threshold τ_2

Fig. 2 Results of threshold sensitivity analysis

Table 3. Benchmark comparison results on education quality evaluation retrieval

Retrieval method	P@10	R@10	nDCG@10	MRR
BM25	0.3003	0.5235	0.4012	0.3194
Contriever	0.3671	0.6041	0.4683	0.3891
RePlug	0.4230	0.6743	0.5127	0.4443
AAR	0.4908	0.7609	0.5982	0.6286
DocReLM	0.5278	0.7946	0.5456	0.5936
ARL2	0.5717	0.8154	0.6614	0.6409
AAR-TRL (ours)	0.6028	0.8291	0.7135	0.6528

To verify the contribution of each component, we design a set of ablation variants as follows: (1) w/o Pair Loss removes the ordered pairwise loss and trains with the list-wise InfoNCE loss only; (2) w/o Multi-Level degrades tri-level labeling to binary labeling by merging "partially supportive" and "fully supportive" as positives; (3) w/o FiDAtt replaces the FiD cross-attention feature scores with random signals. For statistical reliability, each configuration is run five times independently, and the means with standard deviations are reported. The results are given in Table 4.

Removing any core module degrades retrieval precision, recall, and ranking to varying degrees. Degrading tri-level to binary labeling costs 9.43 percentage points in P@10, showing that tri-level labeling effectively distinguishes samples of different support levels and remedies the coarse supervision and information confusion of binary labels. Removing the ordered pairwise loss or the FiD cross-attention module also causes noticeable drops, confirming that the three modules jointly underpin the overall retrieval performance of AAR-TRL.

4.4. Threshold Sensitivity Analysis

The dual thresholds τ_1, τ_2 partition the FiD attention scores into tri-level labels, and their values directly affect training-sample partitioning and final retrieval accuracy. To examine the influence of threshold combinations on P@10, R@10, nDCG@10, and MRR under the constraint $\tau_1 > \tau_2$, we traverse multiple combinations within reasonable ranges and record the results after training; the trends of τ_1, τ_2 are shown in 0(a) and 0(b). In the total loss $\mathscr{L}_{\text{total}} = \mathscr{L}_{\text{list}} + \alpha \cdot \mathscr{L}_{\text{pair}}$s , α controls the weight of the ordered pairwise loss and thus the emphasis on the hierarchical ranking of the three document levels. With the optimal thresholds fixed, we traverse $\alpha \in$ [0.1,0.2,0.4,0.6,0.8,1.0] and report the metric variations in 0(c). In the ordered pairwise loss, γ is the minimum similarity margin between positives and hard negatives; we traverse $\gamma \in$ [0.05,0.1,0.15,0.2,0.25], with the results in 0(d).

Experiments show that the optimum is attained with the tri-level thresholds $\tau_1 = 0.6, \tau_2 = 0.25$, the joint loss weight $\alpha = 0.5$, and the margin $\gamma = 0.2$, which constitute the optimal parameter combination adopted in this paper.

Table 4. Tablation Results

Model variant	P@10	R@10	nDCG@10	MRR
AAR-TRL (ours)	0.6028±0.005	0.8291±0.006	0.7135±0.005	0.6528±0.004
w/o pairwise loss	0.5694±0.006	0.8112±0.007	0.6465±0.007	0.6352±0.005
w/o Multi-Level	0.5085±0.007	0.7762±0.008	0.6102±0.008	0.6351±0.006
w/o FiDAtt	0.5341±0.006	0.7935±0.007	0.6285±0.007	0.6318±0.005

4.5. Generation Quality Evaluation

To assess answer precision in compulsory education quality evaluation, a human evaluation of generation quality is conducted. One hundred queries are randomly sampled from the test set (58 indicator-definition queries and 42 policy-basis queries). Four representative baselines, namely BM25, Contriever, AAR, and ARL2, together with our AAR-TRL, provide retrieval-augmented inputs to DeepSeek-V4 to generate the corresponding answers. Three domain experts jointly formulate a unified reference-answer framework specifying, for each query, the core definitional elements to be included and the policy provisions to be cited, against which the generated answers are scored.

Two metrics are designed. (1) Indicator matching accuracy: the consistency of the indicator definitions and computation methods cited in the answer with the reference, where a complete match scores 1, a partial match with deviation scores 0.5, and no match or no citation scores 0; the scores are converted to a percentage. (2) Policy citation validity: the fit between the cited policy documents and provisions and the query demand, where an accurately cited policy with fully applicable provisions scores 1, a correct policy with only locally applicable provisions scores 0.5, and an erroneous or missing citation scores 0; the scores are converted to a percentage.

Table 5 reports the results. For indicator matching accuracy, AAR-TRL reaches 87.0%, exceeding the vanilla AAR by 11.5 percentage points, ARL2 by 4.7, and BM25 by 25.8. For policy citation validity, AAR-TRL reaches 82.0%, surpassing AAR by 11.2, ARL2 by 4.5, and BM25 by 25.2. AAR-TRL significantly outperforms all baselines on both metrics, and the gains in retrieval quality translate effectively into gains in generation quality, verifying the end-to-end effectiveness of AAR-TRL in question answering for compulsory education quality evaluation.

5. Conclusion and Future Work

To address the inability of binary retrieval in educational evaluation RAG scenarios to distinguish partially relevant hard negatives with dimensional misalignment, this paper evaluation RAG scenarios to distinguish partially relevant hard negatives with dimensional misalignment, this paper proposed AAR-TRL, a tri-level relevance adaptive retrieval method. It constructs automated tri-level soft labels via the FiD cross-attention of Flan-T5, designs a joint training objective combining the list-wise InfoNCE loss and the ordered pairwise loss, and introduces a multi-level hard negative sampling strategy centered on partially supportive texts. On our constructed compulsory education quality evaluation dataset, AAR-TRL achieves the best P@10, R@10, and MRR against mainstream baselines including BM25, AAR, and ARL2. It should be noted that all findings are based on a single dataset covering the compulsory education stage; conclusions drawn here are therefore scenario-specific rather than general claims about education quality evaluation.

Table 5. Human evaluation results of generation quality (%)

Method	Indicator matching accuracy	Policy citation validity
BM25	61.2	56.8
Contriever	65.8	61.2
AAR	75.5	70.8
ARL2	82.3	77.5
AAR-TRL (ours)	87.0	82.0

The method responds to the three challenges raised in the Introduction. For the insufficient depth of domain expertise, domain-specific tri-level labeling and joint-loss training improve the retriever's recognition of professional documents on compulsory education evaluation, raising P@10 by 11.2 percentage points over the vanilla AAR and attaining an indicator matching accuracy of 87.0%. For the dynamism of standards, the retriever works as a plug-and-play adapter: when policies are updated, re-indexing the document base suffices, without retraining the LLM. For the context dependency of evaluation, the tri-level labeling distinguishes the support of the same evaluative act across dimensions, and the ordered pairwise loss heightens the model's sensitivity to contextual differences. These improvements, obtained on a dataset confined to the compulsory education stage, suggest that AAR-TRL is effective for retrieval and question answering in compulsory education quality evaluation; whether it generalizes to preschool, vocational, or higher education awaits further validation. Nevertheless, they rest on specific premises-e.g., that the attention distribution of the proxy model approximates the generation preference of the target LLM, and that the data distribution is confined to the compulsory education stage-which delimit the boundary and limitations of this study.

Several limitations remain. First, the tri-level soft labels are indirectly derived from Flan-T5 attention and prone to labeling noise for fine-grained professional indicators, with fixed thresholds that cannot adapt to the data distribution. Second, the training and test data cover only the compulsory education stage; validation on preschool, vocational, and higher education datasets is absent, leaving cross-stage transfer unproven. Third, retrieval relies solely on plain-text dense vector matching without educational entity knowledge, lacking multi-hop evidence retrieval capability.

Future work will proceed in three directions. First, an adaptive dynamic threshold strategy based on sample confidence distributions, combined with a small number of expert annotations to correct soft-label noise, will improve tri-level labeling precision. Second, given the hierarchical relations and stage applicability among educational evaluation indicators (e.g., the academic proficiency rate is a sub-indicator of the first-level indicator "academic level" and applies only to compulsory education), we plan to build a lightweight educational evaluation ontology that encodes indicator - stage - policy relations as structured knowledge and adds a graph-path reranking layer over the dense retrieval results, explicitly penalizing candidate documents mismatched in dimension or school stage. Finally, we will construct a larger multi-stage dataset incorporating multimodal texts to comprehensively verify the model's cross-stage, multi-scenario generalization.

Acknowledgments

This work was supported by the planning-level project Large-Model-Driven Data Processing and Its Visualization (Grant No. 2025-01-131-BZk01).

References

[1] Wang, Q., & Zhang, H. (2024). Regional application of national compulsory education quality monitoring results: Practical dilemmas and remedies. Educational Measurement and Evaluation, (06), 47–55. https://doi.org/10.16518/j.cnki.emae.2024.06.004. (in Chinese)

[2] Sun, H., Wang, Z., Wu, Y., et al. (2025). Enhancing educational LLM question answering via reranking and post-retrieval reflection. Big Data Research, 11(05), 4–17. (in Chinese)

[3] Wei, G., et al. (2024). DocReLM: Mastering document retrieval with language model. arXiv preprint arXiv:2405.11461.

[4] Yu, Z., et al. (2023). Augmentation-adapted retriever improves generalization of language models as generic plug-in. arXiv preprint arXiv:2305.17331.

[5] Zhang, L., et al. (2024). ARL2: Aligning retrievers for black-box large language models via self-guided adaptive relevance labeling. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (pp. 3708–3719).

[6] Shi, W., et al. (2024). RePlug: Retrieval-augmented black-box language models. In Proceedings of the North American Chapter of the Association for Computational Linguistics.

[7] Rahman, G. R. R., Rahman, M. M. K., Lia, N., et al. (2026). Artificial intelligence enhanced content review for curriculum evaluation in nursing education: A pilot study. Nurse Educator. https://doi.org/10.1097/NNE.0000000000002243.

[8] Cai, Z., Xu, N., Cai, L., et al. (2025). Educational knowledge graph based intelligent question answering for automatic control disciplines. Applied Intelligence, 55(13), 943. https://doi.org/10.1007/s10489-025-06847-x.

[9] Yang, Y., & Li, R. (2025). Research on the application of large models in intelligent question answering and learning recommendation. Industry Science and Engineering, 2(11). https://doi.org/10.62381/I255B05.

[10] Liu, J., & Xia, L. (2026). Construction of educational knowledge graphs and intelligent question answering based on large language models and hybrid retrieval augmentation. Technology Intelligence Engineering, 12(1), 29–39. (in Chinese)

[11] Lin, S. (2026). Research on course intelligent question answering systems integrating knowledge graphs and RAG. Fujian Computer, 42(5), 69–73. (in Chinese)

[12] Zhang, W., Gan, L., Du, M., et al. (2026). A precise course knowledge question answering system integrating knowledge graphs and large language models. Modern Information Technology, 10(8), 113–117. (in Chinese)