ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 1 - 2026

CER: Learning Causally Sufficient Evidence Sets for Verifiable and Trustworthy RAG

Jizhang Tan[1], *Yonghui Xu*[2,*] *and Lizhen Cui*[3]

[1] School of Software, Shandong University, Jinan, 25000,Shandong, China
[2] Joint SDU-NTU Centre for Artificial Intelligence Research (C-FAIR), Shandong University, Jinan, 25000, Shandong, China
[3] School of Software, Shandong University, Jinan, 25000, Shandong, China

SUMMARY: *Aiming at the problem that evidence selection in retrieval enhancement generation focuses on semantic relevance, and it is difficult to ensure causal necessity and minimum sufficiency, a retrieval enhancement generation method based on verifiable causal sufficient evidence set was proposed. Method A counterfactual delete-replace intervention is used to construct an evidence-level supervision signal. A pre-trained language model causal re-ranking, R-GCN evidence complementary modeling, quality proxy function calibration, and minimum sufficient set search are combined to jointly optimize the answer quality, evidence size, and reasoning cost. Experiments on HotpotQA, HOVER, QASPER and ALCE datasets show that the proposed method achieves F1 of 67.1 on HotpotQA and 82.3 on HOVER, with an average evidence size reduction of 18%-22%. The context overhead is reduced by up to 60%, and it maintains better robustness under the condition of conflict evidence and near-repeated interference. The results show that the proposed method can effectively improve the verifiability, compactness and stability of the retrieval enhancement generation system.*

KEYWORDS: *Retrieval enhancement generation; Counterfactual learning; Evidence graph modeling; Verifiable generation*

1 Introduction

Retrieve-augmented generation collaborates the external retrieval module and the large language model generation process, which has become an important technical path for open-domain question answering, fact checking and scientific question answering. Ram et al. (2023) proposed to dynamically access external evidence in context to enhance the knowledge utilization ability of language models [1]. Siriwardhana et al. (2023) studied the domain adaptation problem of RAG in open-domain question answering [2]. Katsis et al. (2025) further constructed RAG evaluation benchmarks in multi-turn dialogue scenarios [9]. Related studies have shown that external evidence access can improve the answer accuracy, reduce the risk of unfounded generation, and enhance the source traceability of results [1, 2, 9]. However, the existing mainstream methods still mainly rely on BM25, dense retrieval or cross-encoder to achieve semantic relevance ranking, and their optimization goal focuses on "retrieving relevant text" rather than "identifying the really necessary evidence for the answer". Li et al. (2025) studied the development path of generative information retrieval [12], and Huang et al. (2025) pointed out that large model illusion control is closely related to evidence quality [13].

*xu.yonghui@hotmail.com
https://doi.org/10.65102/is2026255
Published: 30 April, 2026

Therefore, how to move from relevance selection to causal adequacy selection under the framework of RAG has become a key issue to improve system trustworthiness and verifiability.

The prominent problem with existing methods is that highly correlated evidence does not necessarily constitute causal support for the answer to hold. Russo et al. (2023) studied the task of fact-checking explanation generation and pointed out that explanation quality depends not only on content relevance, but also on the pertinence and adequacy of the supporting chain [3]. Adlakha et al. (2024) proposed that correctness and faithfulness in question answering systems are not completely consistent [6]. Chaturvedi et al. (2024) analyzed the true dependence of language models on evidence through input intervention [18]. The examples of medical questions and answers in the original manuscript show that for a drug efficacy question, the relevance search will return multiple paragraphs such as trial results, trial protocols, approval history, disease statistics, and guideline text, but only a few of the experimental evidence are really determining the answer.

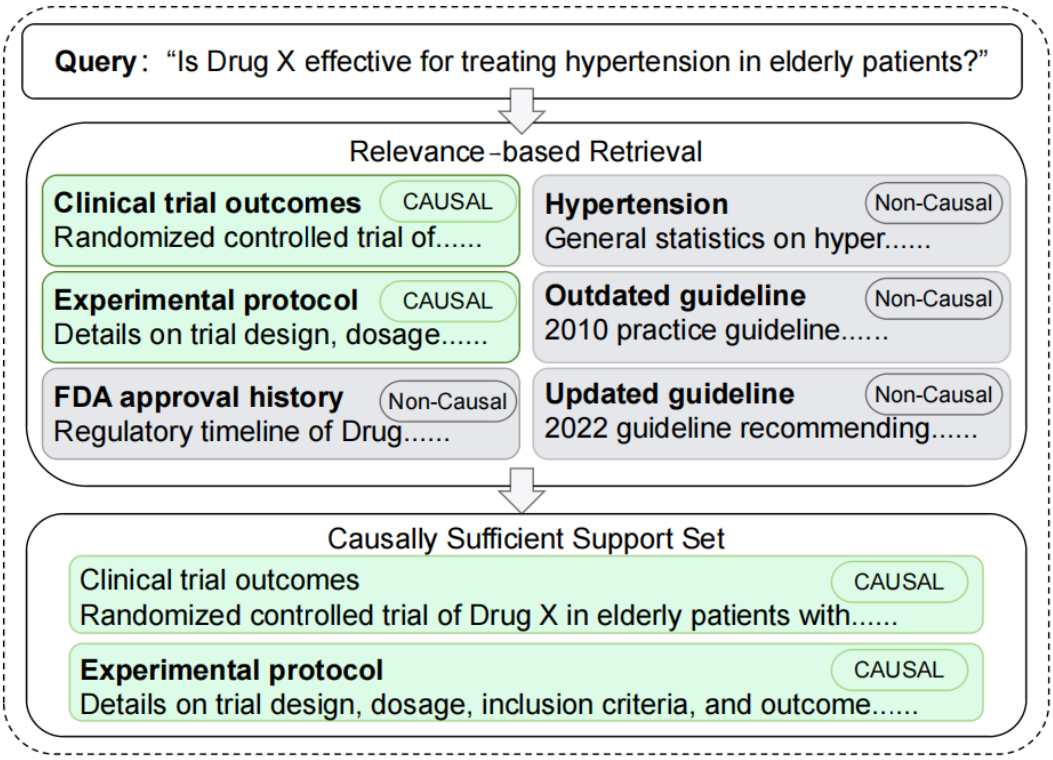

Figure 1: Figure: Relevance.based retrieval returns many related passages; causal filtering keeps only the few that actually determine the answer

As shown in Figure 1, relevance retrieval returns "a large number of relevant texts", while causal filtering retains only "a few key pieces of evidence that determine the answer". The manuscript further points out that about 60% of the retrieved passages in a typical scenario belong to causal lazy evidence, that is, deletion does not change the validity of the answer, which will directly cause context waste, increase the burden of manual verification, and enhance the interference of conflict sources.

Around the above issues, RAG evidence selection faces three technical challenges. First, evidence necessity judgment is essentially a counterfactual reasoning problem, which needs to answer "is the answer invalid if a piece of evidence is removed", while traditional relevance learning methods lack such counterfactual perception ability [4, 18]. Secondly, there is a complex relationship among complementary, redundant and conflicting evidences. Glockner et

al. (2024) studied the problem of supporting multiple evidences in ambiguous fact checking [7], Zeng and Gao (2024) studied the reason generation mechanism in interpretable fact checking [8], and Hosking et al. (2024) discussed the problem of multi-paragraph evidence organization from the perspective of hierarchical indexing [16]. Third, under the constraints of limited token budget and reasoning cost, evidence selection needs to simultaneously consider answer quality, evidence minimality and context efficiency. Liu et al. (2024) found that long context does not linearly improve model performance, and excessive evidence may weaken the utilization efficiency of key evidence [20]. Wang et al. (2024) proposed that "retrieving the really needed content" is more important than simply expanding the candidate set in open-domain question answering [19].

In view of the above shortcomings and challenges, this paper constructs a task of causal sufficient evidence set selection, which elevates the evidence selection from the problem of relevance ranking to the problem of causal contribution estimation. A counterfactual supervision driven evidence contribution learning method is proposed, which describes the marginal impact of candidate evidence on answer quality by deleting intervention and replacing intervention. A framework called RAG is designed, which integrates evidence graph modeling and minimal set search. It combines pre-trained language model encoding, graph neural network relationship propagation, quality proxy function calibration and cost constraint search algorithm to effectively distinguish complementary evidence, redundant evidence and conflict evidence. And the comprehensive advantages of this method in accuracy, minimality, robustness and cost control are verified on multiple types of knowledge-intensive data sets [11, 17]. This research provides a new technical path for building a retrieval enhanced generation system with interpretability, auditability and engineering deployability.

2 Related Work

2.1 Retrieval and reranking methods in Retrieval Enhancement Generation

Retrieval enhanced generation alleviates the knowledge gap and illusion problems of large language models in open domain question answering, fact checking and scientific question answering by introducing external knowledge sources. Ram et al. (2023) proposed to access retrieval results in context to enhance the knowledge utilization ability of language models [1]. Siriwardhana et al. (2023) studied the domain adaptation mechanism of RAG in open-domain question answering [2]. Li et al. (2025) reviewed the technical evolution of generative information retrieval from matching-driven to generation-driven [12]. In the existing technology links, BM25 relies on term frequency and inverse document frequency to achieve efficient sparse recall, DPR uses dual coding structure to complete dense vector matching between questions and documents, ColBERT retains fine-grained semantic alignment information with token-level late interaction. Cross-Encoder improves the reranking accuracy by jointly encoding the query and the candidate paragraph. Hosking et al. (2024) further discussed the hierarchical indexing problem in multi-segment evidence organization [16].

These methods significantly improve the semantic matching ability of candidate evidences, but their core optimization goal is still relevance ranking, rather than causal necessity identification. Wang et al. (2024) show that the performance improvement of open-domain question answering depends on whether the desired content is retrieved, not only on expanding the size of the candidate set [19]. Park and Lee (2024) pointed out that imperfect retrieval can significantly weaken the stability of retrieval augmented language models [4]. Therefore,

BM25, DPR, ColBERT and Cross-Encoder can well answer "which evidence is relevant to the question", but it is difficult to judge "which evidence has an irreplaceable supporting role for the answer". This objective bias makes the existing retrieval and re-ranking methods, although constituting the key foundation of RAG systems, still insufficient to directly support the construction of sufficient evidence sets for verifiable causality.

2.2 Attribution, Interpretable, and verifiable Generation Methods in RAG

The research on attribution, explainable and verifiable generation in RAG mainly focuses on whether the generated content can be adequately supported by external evidence, and whether the model output maintains factual fidelity. Russo et al. (2023) studied explanation generation in a fact-checking scenario and pointed out that high-quality explanations need not only correct conclusions, but also be consistent with supporting evidence and verifiable [3]. Adlakha et al. (2024) further distinguish the evaluation of question answering system into two dimensions: correctness and faithfulness, emphasizing that a correct answer does not mean that the answer is truly supported by evidence [6]. On this basis, Zeng and Gao (2024) studied justification generation around interpretable fact checking, indicating that there is a close connection between justification generation and evidence support [8]. Glockner et al. (2024) introduced multiple evidence analysis into the verification of ambiguity claims and showed that evidence support relations are not always single, certain and independent [7]. At the same time, Lyu et al. (2024) systematically summarized the main technical paths of faithful model explanation in NLP, and pointed out that the key of interpretability research is whether the explanation truly reflects the decision basis of the model, rather than only providing a superficial reasonable explanation [17]. Chaturvedi et al. (2024) analyzed the semantic faithfulness of language models through input intervention, and further revealed the consistency problem between model output and input evidence [18].

This kind of research has significantly improved the interpretability and auditability of RAG systems. However, the focus of existing methods is still on "whether the answer can be supported" rather than "whether the supporting evidence has reached minimum and sufficient". Rosenthal et al. (2025) studied the task of long answer generation based on passages, emphasizing that answer organization and evidence correspondence play an important role in RAG evaluation [10]. Katsis et al. (2025) constructed multiple rounds of RAG evaluation benchmarks, which promoted the systematic analysis of the verifiability of retrieval enhancement generation systems [9]. In addition, Huang et al. (2025) and Ji et al. (2023) pointed out that insufficient external support, inaccurate evidence citation and context redundancy are the core factors affecting the credibility of generation from the perspective of large language model illusion and natural language generation illusion respectively [13, 14]. Therefore, although the existing imputation, explainable and verifiable generation methods can improve the ability of evidence support analysis, they rarely further describe the minimality, complementarity and causal necessity of evidence, and are not enough to directly support the construction of verifiable causal sufficient evidence sets.

2.3 Counterfactual Learning and Causal Inference Methods

Counterfactual learning and causal inference methods provide important tools for model decision basis analysis. Chaturvedi et al. (2024) tested the semantic fidelity of the QA model by deleting intervention and negating intervention, and pointed out that high task performance does not mean that the model really depends on the semantic content of the input [18]. Lyu et al. (2024) systematically sorted out the technical paths in faithful explanation research, such as similarity analysis, model internal structure analysis, backpropagation attribution,

counterfactual intervention and self-explanatory model [17]. In a generative system, Asai et al. (2024) proposed Self-RAG, which controls the output factuality and citation quality through retrieval, generation, and self-reflection. However, its focus is still on self-checking and on-demand retrieval, rather than the explicit estimation of evidence necessity. The above studies show that the deletion intervention, substitution intervention, robustness analysis and posterior attribution methods can characterize the sensitivity of the model to input changes and provide experimental evidence for the explanatory validity.

The existing methods still mainly stay at the input-level or output-level fidelity analysis level, and lack of unified modeling of marginal contribution, complementary relationship and redundant relationship under the condition of multiple evidence. The posterior attribution method can identify high-impact features, but it is difficult to distinguish relevant contributions from causally necessary contributions. Robustness analysis can reveal the vulnerability of the model to perturbations, but usually does not directly serve the minimum sufficient evidence set search. Although deletion or substitution interventions have causal analysis characteristics, they are mostly used for overall input sensitivity assessment rather than evidence-level selection optimization. Therefore, it is necessary to advance counterfactual interventions from a general explanation framework to evidence-level causal contribution modeling, which provides a methodological foundation for verifiable causally sufficient evidence set construction by explicitly estimating the marginal impact of candidate evidence on answer quality.

3 Retrieval enhanced Generation of verifiable causally sufficient Evidence sets

3.1 Task definition and formal modeling

Let the query be q and the set of candidate evidence be $D = \{d_1, d_2, \dots, d_n\}$, the generator is G, and the answer quality function is F. Instead of selecting a number of highly relevant passages from a candidate set, the problem of concern in this paper is to search for the minimal sufficient evidence subset $S \subseteq D$ from D that can support answer generation, given an answer quality threshold θ. Therefore, the task goal can be defined as: on the premise of satisfying $F(q,S) \geq \theta$, make the evidence set size $|S|$ as small as possible, and further quantify the necessity and contribution of each selected evidence to the final answer. In this modeling approach, the evidence selection is transformed from relevance ranking to combinatorial optimization problem under quality constraints, so that the evidence sufficiency, set minimization and result verifiability can be described in a unified framework.

To describe the marginal role of candidate evidence in the current context, let $C \subseteq D$ be the current set of selected evidence and $y \in [0,1]$ denote the answer quality. Then the conditional causal effect of evidence d is defined as follows.

$$f(d \mid q, C) \approx \tau(d \mid q, C) = \mathbb{E}[y \mid C \cup \{d\}] - \mathbb{E}[y \mid C] \tag{1}$$

where $\tau(d \mid q,C)$ represents the gain of answer quality brought by adding evidence d to the existing evidence set C. When $\tau(d \mid q,C)$ is large, the evidence provides irreplaceable new support. When its value is close to zero, it indicates that there is strong redundancy between the evidence and the current evidence set, or it only has superficial relevance but lacks substantial support. Accordingly, the minimum sufficient evidence set search objective can be written as follows.

$$S^* = \arg\min_{S \subseteq D} |S|, \quad \text{s.t.} \quad F(q, S) \geq \theta \tag{2}$$

This optimization objective takes the answer quality as the constraint and the minimum evidence scale as the goal, which helps to suppress redundant evidence splicing and reduce context overhead.

In order to make the answer quality assessment decomposable and auditable, the quality function is defined as follows.

$$F(q, S) = w_1 \, p_{\text{entail}}(q, S) + w_2 \, p_{\text{cov}}(q, S) - w_3 \, p_{\text{contra}}(q, S) - w_4 \, p_{\text{len}}(S) \tag{3}$$

Here, $p_{\text{entail}}(q, S)$ represents the implication probability between the generated answer and the evidence set, $p_{\text{cov}}(q, S)$ represents the proportion of key entities, numerical and temporal information in the answer covered by evidence, $p_{\text{contra}}(q, S)$ represents the degree of contradiction between the answer statement and the evidence, $p_{\text{len}}(S)$ represents the context length penalty term. This can be written as |S|/Kmax. By jointly modeling support, coverage, contradiction and length cost, the answer quality no longer depends on a single correct rate index, but can reflect the completeness and compactness of the evidence support chain.

In evidence-level verifiability analysis, the necessity score is defined as follows.

$$\text{Nec}(d) = \frac{\max(0, \Delta y_{\text{del}}(d))}{\sum_{d' \in S} \max(0, \Delta y_{\text{del}}(d'))} \tag{4}$$

Here, $y_{\text{del}}(d)$ represents the decrease in answer quality after removing evidence d. This index is used to measure the relative contribution of a single piece of evidence to the overall support chain. When Nec(d) is higher, it means that the evidence has stronger irreplaceability for the answer. When its value is low, it means that the evidence is more likely to belong to complementary or redundant information. Based on the above definitions, conditional causal effects, answer quality functions and necessity scores can be used in counterfactual supervision, evidence re-ranking and minimal set search to jointly model the verifiable causal sufficient evidence set.

3.2 Evidence Supervised Signal Construction with counterfactual interventions

The core of evidence supervision signal construction is to transform the relevance judgment of candidate evidence into a computable measure of answer quality change. The existing research on fidelity analysis shows that input intervention can more effectively reveal the semantic basis of the true dependence of the model. Chaturvedi et al. (2024) studied the relationship between input intervention and semantic fidelity in question answering scenarios [18], and Lyu et al. (2024) summarized the role of counterfactual intervention in model interpretation and causal analysis [17]. On this basis, we learn to construct two types of counterfactual supervision signals for the minimum sufficient evidence set: deletion intervention and replacement intervention. Suppose that the training sample consists of a query q, a standard answer $a*$ and a minimum sufficient evidence set $E*$. For any evidence $d \in E*$, we first perform a deletion intervention to remove d from the evidence set and calculate the decline of the answer quality:

$$\Delta y_{\text{del}}(d) = F(q, E^*) - F(q, E^* \backslash \{d\}) \tag{5}$$

If the quality of the answer decreases significantly after deletion, it means that the evidence

has a strong necessity for the answer. A smaller decrease indicates that the evidence is more likely to be complementary or redundant information. This mechanism uses the answer quality function F as a bridge to convert whether evidence is necessary into a supervised numerical signal. The whole counterfactual supervised construction flow is shown in Figure 2.

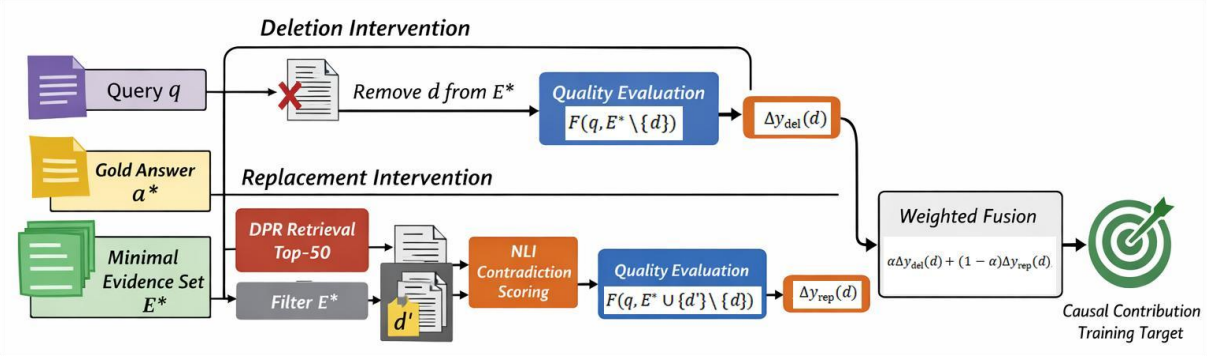

Figure 2: Counterfactual supervised construction process

Relying on deletion interventions alone may still suffer from retrieval bias, where certain pieces of evidence exhibit high correlation due to long-term co-occurrence with the correct answer, but do not necessarily provide irreplaceable causal support. To reduce this bias, substitution intervention is further introduced in this paper. For each piece of target evidence d, we construct hard negative samples d' that are semantically similar to it but have conflicting facts, and calculate the quality change after replacement:

$$\Delta y_{rep}(d) = F(q, E^*) - F(q, E^* \cup \{d'\} \setminus \{d\}) \tag{6}$$

The construction of hard negative samples adopts a two-stage strategy of "DPR recall + NLI contradiction screening": Firstly, DPR is used to recall the Top-50 candidate paragraphs that are semantically closest to d from the corpus, and then the evidence of belonging to $E*$ is eliminated. Then, the remaining candidates are sorted according to the product of semantic similarity and NLI contradiction score, and the paragraph with the highest score is selected as d'. This design ensures that the replacement evidence is close to the original evidence at the topic level, but there is conflict with the evidence set at the time, fact or conclusion level, so as to more accurately test whether the target evidence has real support. Finally, the deletion effect and substitution effect are weighted and fused into a unified causal supervision signal:

$$\Delta y(d) = \alpha \Delta y_{del}(d) + (1 - \alpha)\Delta y_{rep}(d), \quad \alpha = 0.7 \tag{7}$$

Among them, a higher α is used to strengthen the necessity judgment, and a lower proportion of substitution effects is used to suppress the spurious contribution caused by correlation co-occurrence. The obtained (q,C,d,Δy(d)) training samples can be directly used for the supervised learning of the subsequent evidence contribution prediction network, so that the model can learn the marginal causal effect of evidence under different context conditions.

3.3 Causal Evidence Re-ranking Network with Pre-trained Language Models

After obtaining the counterfactual supervision signal Δy(d), we construct a causal evidence re-ranking network based on the pre-trained language model to estimate the marginal causal contribution of the candidate evidence under the current context. If the query is q, the currently

selected evidence set is C, and the candidate evidence is d, then the model prediction target is $\hat{\tau}(d \mid q,C)$, which means the incremental contribution of adding evidence d to the existing evidence set C to the answer quality. The main body of the network adopts BERT-large or DeBERTa-v3 as the coding backbone to make full use of the ability of large-scale pre-trained language models in cross-sentence semantic modeling, long-range dependency representation and fine-grained semantic matching. The training sample consists of quadruples (q,C,d,Δy(d)), where C represents the partial evidence context and Δy(d) is obtained by fusing the deletion intervention and replacement intervention described above. Thus, instead of relying solely on query-evidence relevance, the reranking process shifts to a direct regression on the marginal causal effect of candidate evidence.

The input layer adopts the concatenated sequence construction method:

$$X = [q;\ \mathrm{SUMMARY}(C);\ d] \tag{8}$$

Here, SUMMARY(C) represents the context summary of the current evidence set C, which can be obtained by sequentially concatenating the first 100 tokens of each evidence in the set. When C=∅, this part is set to an empty string. Such input design enables the model to simultaneously perceive the semantics of the question, the state of the current support chain and the content of the evidence to be evaluated, so as to learn the conditional judgment mechanism of "whether the evidence still provides new support in the current context". Let the global semantic representation of the encoder output be:

$$\mathrm{h_{cls}} = \mathrm{PLM(X)}_{[CLS]} \tag{9}$$

Then the marginal causal contribution regression is performed by using two layers of multi-layer perceptron:

$$\hat{\tau}(\mathrm{d} \mid \mathrm{q}, \mathrm{C}) = \mathrm{W}_2\, \sigma(\mathrm{W}_1 \mathrm{h_{cls}} + \mathrm{b}_1) + \mathrm{b}_2 \tag{10}$$

Here, σ(·) represents the nonlinear activation function, and W1,W2,b1,b2 are the learnable parameters. The final output is a single scalar, which characterizes the causal contribution strength of the candidate evidence under the current evidence condition.

Our network is structurally a conditional cross-encoder: the query, context summary, and candidate evidence are jointly modeled in the same encoding space, which is able to explicitly capture cross-segment interaction features without relying on static similarity estimates after independent vector recall. Because SUMMARY(C) encodes the main information of the selected evidence, the model can recognize two important phenomena in the inference process. One is that the candidate evidence is highly repeated with the existing evidence, and its $\hat{\tau}(d \mid q,C)$ will be depressed. The other is the candidate evidence, which may not be the most similar to the query, but can supplement key entities, methods, values, or constraints, and its predictive contribution will be significantly improved. The obtained causal reran score can be used as the direct input for subsequent evidence graph modeling and minimum sufficient evidence set search, thus providing a fine-grained, conditional, and regressible computational basis for evidence selection.

3.4 Complementary Evidence Modeling Method fusing Graph Neural Networks

Candidate evidence does not play an independent role in supporting answer generation, but has multiple relationships such as complementarity, redundancy and conflict. Depending on the

local scoring of a single piece of evidence, it is difficult to identify the multi-hop support structure that is "individually insufficient but jointly effective". To this end, we further construct an undirected evidence graph G=(V,E): based on causal reranking, where the node set V represents the candidate evidence paragraphs, and the edge set E represents the semantic association and fact relationship between the evidence. Let the set of named entities of evidence di,dj be Ei,Ej respectively, and the release time or time label be yi,yj respectively. Then the edge weight is jointly defined by four types of relations: Entity overlap relation:

$$w_{ij}^{ent} = \text{Jaccard}(E_i, E_j) \tag{11}$$

Temporal consistency relation:

$$w_{ij}^{time} = \mathbb{I}(|y_i - y_j| \leq 2) \tag{12}$$

Source consistency relationship:

$$w_{ij}^{src} = \begin{cases} 0.8, & d_i, d_j \text{From the same source domain or author} \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

And the contradiction relations calculated based on the natural language inference model:

$$w_{ij}^{contra} = \max_{s_p \in d_i, s_q \in d_j} \text{NLI}_{contra}(s_p, s_q) \tag{14}$$

An undirected edge is established between nodes vi and vj when the entity overlap weight is greater than 0.3, or the contradiction score is greater than 0.5, or the time consistency and source consistency conditions are satisfied. The evidence graph constructed in this way can simultaneously preserve the shared entities in the supporting chain, the temporal consistency, the source reliability and the fact conflict information. Figure 3 shows the process of evidence graph construction and relation propagation.

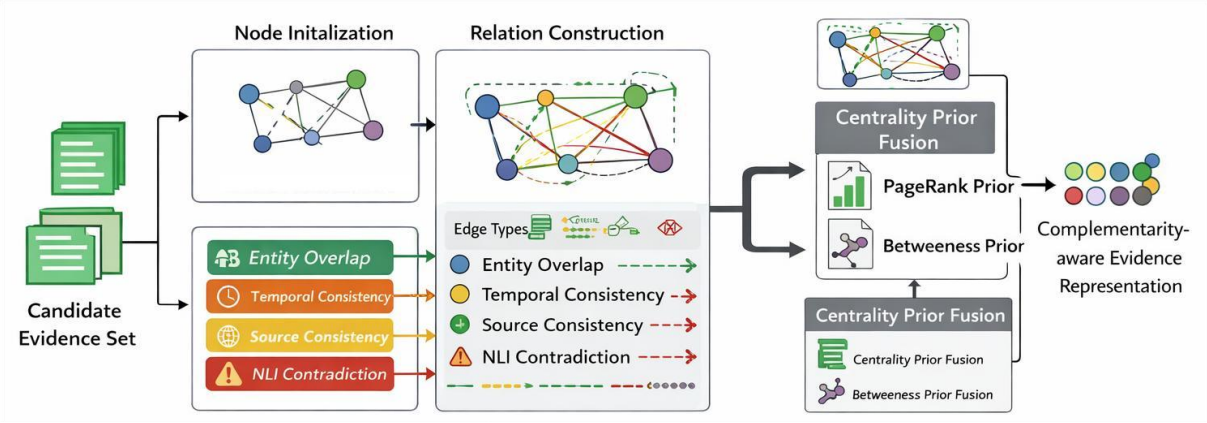

Figure 3: Evidence graph relationship construction and R-GCN complementary modeling framework

In the graph representation learning stage, this paper uses a three-layer relational graph convolutional network R-GCN to contextualize the evidence graph. Let the l-th layer node be denoted as $h_i^{(l)}$, then its update process is written as follows.

$$h_i^{(l+1)} = \sigma\left(\sum_{r\in\mathcal{R}}\sum_{j\in\mathcal{N}_i^r}\frac{1}{c_{i,r}}W_r^{(l)}h_j^{(l)} + W_0^{(l)}h_i^{(l)}\right) \tag{15}$$

Here, R={ent,time,src,contra} represents the set of relationship types, $\mathcal{N}_i^r$ represents the set of neighbor nodes of node vi under relationship r, ci,r are the normalization coefficients, $W_r^{(l)}$ and $W_0^{(l)}$ are the relationship specific parameter matrices, and σ(·) is the nonlinear activation function. After three layers of propagation, the node representation not only encodes the current evidence content, but also aggregates the complementary constraints and conflict information provided by adjacent evidence, so as to identify multi-hop complementary structures such as "experimental results-experimental conditions", "conclusion statement-numerical evidence", and "claims-source of negative evidence". In order to enhance the ability of identifying the importance of cross-cluster connection evidence, we further introduce PageRank and betweenness as centrality priors to measure the importance of a node in the global propagation chain and its bridging ability in multi-hop support paths, respectively, and fuse them with the final graph representation:

$$\tilde{h}_i = \mathrm{MLP}([h_i^{(3)}; \mathrm{PR}(v_i); \mathrm{BC}(v_i)]) \tag{16}$$

Here, $\mathrm{PR}(v_i)$ represents PageRank score and $\mathrm{BC}(v_i)$ represents betweenness centrality. The introduction of centrality prior enables the model to preferentially retain the key nodes that connect different topic clusters and assume the role of bridging across evidence, reducing the interference of local highly relevant but global redundant evidence. Finally, the obtained $\tilde{h}_i$ will be used as the structured input of the subsequent causal evidence re-ranking and the minimum sufficient evidence set search, so as to realize the joint modeling of evidence complementarity, redundancy and conflict.

3.5 Causal score fusion calibrated with a quality proxy function

In order to avoid calling the generator G frequently for complete answer generation in the evidence selection stage, this paper constructs the quality surrogate function F^ as the answer quality evaluation sub-model to quickly estimate the supporting ability of the candidate evidence set. The sub-model takes four types of interpretable features as input, including the entailment score pentail, the entity coverage pcov, the contradiction detection score pcontra, and the length penalty term plen. Firstly, the features are concatenated to form a vector x, and then nonlinear fusion is completed by two layers of multi-layer perceptron to obtain the uncalibrated quality score:

$$z = w_1 p_{entail} + w_2 p_{cov} - w_3 p_{contra} - w_4 p_{len} \tag{17}$$

Among them, w1,w2,w3 and w4 are learnable parameters that control the contribution strength of support, coverage, contradiction and context cost to the overall quality, respectively. Different from the single relevance score, this fusion mechanism simultaneously describes four dimensions of "whether the evidence supports the answer", "whether the evidence covers the key information", "whether the evidence has conflict" and "whether the evidence cost is acceptable", which makes F^ have stronger interpretability and operability. The causal score fusion and quality proxy function calibration process is shown in Figure 4.

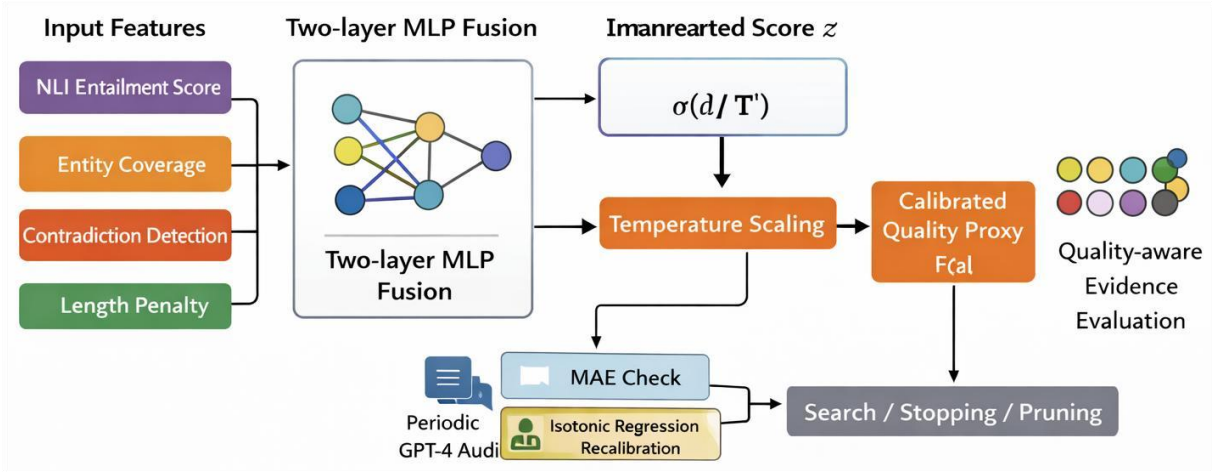

Figure 4: Causal score fusion with quality proxy function calibration framework

Since there is still a distribution bias between the uncalibrated score z and the true answer quality, we further introduce temperature scaling to achieve probabilistic calibration:

$$\hat{F}_{cal}(q,S) = \sigma\left(\frac{z}{T}\right) \tag{18}$$

where T is the temperature parameter learned on the validation set, and the goal is to minimize the expected calibration error. To control the drift of the surrogate model in the later stage of training, we adopt a periodic audit mechanism. We sample 100 instances in each training cycle, and call GPT-4 to score F(q,S), which is used as the true quality reference. When the mean absolute error between $\hat{F}$ and the audit results exceeded 0.15, a recalipay was performed using isotonic regression to correct for nonlinear bias and maintain consistency of the surrogate function with the true mass. The calibrated F^cal will directly serve the subsequent evidence search, stop decision and reverse pruning process, so as to significantly reduce the reasoning cost while ensuring the stability of the evaluation.

3.6 Minimum sufficient evidence set search algorithm

After obtaining the predicted causal contribution $\hat{\tau}(d \mid q, C)$ and the quality proxy function $\hat{F}(q, C)$ of the candidate evidence, the construction of the minimum sufficient evidence set is modeled as a set optimization problem under quality constraints. The search process adopts a two-stage strategy of "forward greedy selection + reverse redundant pruning". Let the current selected evidence set be C and the remaining candidate set be R|C, then the utility function is defined for candidate evidence d at each step as follows.

$$U(d \mid q, C) = \hat{\tau}(d \mid q, C) - \mu \cdot cost(d) + \kappa \cdot \hat{\sigma}(d \mid q, C) \tag{19}$$

where cost(d)= $\mid d \mid /1000$ represents the evidence length cost, μ is the length penalty coefficient, κ is the uncertainty exploration coefficient, and $\hat{\sigma}(d \mid q, C)$ represents the uncertainty of the candidate evidence contribution prediction. The design simultaneously considers three types of factors: the marginal gain of evidence on answer quality, the token cost of evidence occupation, and the degree of uncertainty of the model about the current judgment. In each iteration, the following is chosen from the unselected candidates:

$$d^* = \arg\max_{d \in R \setminus C} U(d \mid q, C) \tag{20}$$

And add it to the current evidence set $C \leftarrow C \cup \{d^*\}$, and then recalculate $\hat{F}(q, C)$. If the forward search satisfies $\hat{F}(q, C) \geq \theta$, the current set has reached the quality threshold. If the gain does not exceed δ for two consecutive rounds after crossing the threshold, the forward expansion is stopped to avoid continuing to introduce evidence of very low marginal contributions under noise perturbations.

To improve the robustness of the search process, MC Dropout is used in this paper to estimate the uncertainty of the candidate evidence contribution. Specifically, performing T=10 random inactivation forward propagation for the same input yields $\{\hat{\tau}_t(d \mid q, C)\}_{t=1}^{T}$, and the predicted fluctuation is characterized by the standard deviation:

$$\hat{\sigma}(d \mid q, C) = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\hat{\tau}_t(d \mid q, C) - \bar{\tau}(d \mid q, C))^2} \tag{21}$$

where $\bar{\tau}(d \mid q, C)$ is the mean of T predictions. When $\hat{\sigma}(d \mid q, C)$ is high, it means that the model is not stable to judge the causal effect of the evidence, and the uncertainty term in the utility function can avoid premature exclusion of potential key evidence. After the forward phase, reverse pruning is performed on the current set. For any evidence $d \in C$, if deletion still satisfies:

$$\hat{F}(q, C\backslash\{d\}) \geq \theta + \varepsilon \tag{22}$$

The evidence is removed; Here, ε is the safety margin, which is used to cancel the surrogate function estimation error. The pruning is performed iteratively in the reverse order of evidence addition, and repeated until there are no nodes that can be deleted. Finally, the minimum sufficient evidence subset S that can not be reduced is obtained.

Algorithm 1 Minimal Sufficient Evidence Set Search

Input: query q, candidate set R, quality threshold θ

Output: minimal sufficient evidence set S

Initialize $C \leftarrow \emptyset$

For each $d \in R \setminus C$, compute $\hat{\tau}(d \mid q, C), \hat{\sigma}(d \mid q, C)$, and $U(d \mid q, C)$

Select $d^* = \arg\max U(d \mid q, C)$, update $C \leftarrow C \cup \{d^*\}$

Recompute $\hat{F}(q, C)$; $\hat{F}(q, C) \geq \theta$ and marginal gain $< \delta$ for two consecutive rounds, stop forward search

Traverse C in reverse order; if $\hat{F}(q, C\backslash\{d\}) \geq \theta + \varepsilon$, remove d

Repeat Step 5 until no evidence can be removed

Return S=C

3.7 Model Training and Complexity Analysis

The model is trained by a multi-task joint optimization mechanism with the common goals of marginal causal contribution learning, ranking consistency constraint and graph structure preservation. Assuming that the predictive marginal contribution of candidate evidence is $\hat{\tau}(d_i \mid q, C_i)$ and the counterfactual supervision signal is $\Delta y(d_i)$, the regression loss is defined as follows.

$$L_{reg} = \frac{1}{N}\sum_{i=1}^{N} \text{Huber}\left(\hat{\tau}(d_i \mid q, C_i), \Delta y(d_i)\right) \tag{23}$$

Its role is to directly fit the model output to the causal supervision signal obtained from the delete-replace intervention. To ensure the relative order of evidence ranking, the interval ranking loss is further introduced:

$$L_{margin} = \sum_{(d^+,d^-)} \max(0, \hat{\tau}(d^- \mid q, C) - \hat{\tau}(d^+ \mid q, C) + m) \tag{24}$$

where d+ denotes the positive samples in the minimum sufficient evidence set, d− denotes the hard negative samples or random negative samples, and the interval parameter is taken as m=0.5. The graph representation learning part uses the graph reconstruction loss:

$$L_{graph} = -\frac{1}{|V|}\sum_{v \in V} \log p(v \mid neighbors(v)) \tag{25}$$

It is used to constrain R-GCN to preserve the relational structure of the evidence graph during propagation. The overall training objective is written as follows:

$$L = L_{reg} + \lambda_1 L_{margin} + \lambda_2 L_{graph} \tag{26}$$

where λ1=0.3 and λ2=0.1. This joint goal enables the model to have the ability of numerical regression, ranking and distinguishing, and graph relationship modeling.

A three-stage strategy is used in the training process. In the first stage, standard correlation labels are used to perform warm-up pre-training on the cross-encoder for 2 epochs to stabilize the joint encoding representation of query-evidence. In the second stage, causal fine-tuning was performed for 5 epochs based on counterfactual supervision signals, and the optimization objective was$L_{reg} + \lambda_1 L_{margin}$. In the third stage, the cross-encoder body is frozen, only R-GCN and FiLM parameters are trained for 2 epochs, and then the full model is jointly fine-tuned for another epoch at a lower learning rate. AdamW was used as the optimizer, the initial learning rate was set to 2×10^{-5}, the joint fine-tuning phase was reduced to 5×10^{-6}, the batch size was 16, and early stopping was performed according to the change of the validation set MAE in five epochs. Key parameters included α=0.7, K0=100, K1=50, K2=25, θ=0.75, ε=0.05, μ=0.05, κ=0.2, δ=0.02 predominate, the delta = 0.02; The quality function weights are chosen as w1=0.4,w2=0.3,w3=0.2, and w4=0.1, and the F1 and token efficiency are jointly balanced on the validation set by Pareto grid search. The quality proxy function F^ performs calibration every 1000 steps: 100 validation instances are sampled, the true quality score is given by GPT-4, and then the prediction bias is corrected by temperature scaling with isotonic regression if necessary.

In terms of complexity, the system inference link consists of four stages: retrieval, reordering, graph propagation and set search. In the candidate recall stage, K0 paragraphs are retrieved from the whole database. If the corpus size is N, the complexity of the fusion of BM25 sparse recall and dense retrieval can be approximated as O(logN+K0). The lightweight reordering stage performs cross-encoder scoring on K0 candidates, and the complexity is about O(K0·L), where L is the average sequence encoding length. The construction of evidence graph and the three-layer R-GCN propagation were carried out on K1 nodes. If the number of edges was |E|, the propagation complexity was about O(3|E|d), where d was the hidden dimension. The minimum sufficient evidence set search performs forward greedy and reverse pruning on K2 candidates, and its worst-case complexity is about $O(K_2^2)$. Since the system adopts a stepwise compression mechanism of K0=100→K1=50→K2=25, the size of candidates that finally enter the set search stage has been significantly reduced, and the reasoning overhead is

mainly concentrated in the first two levels of retrieval and finalization. At the same time, the minimum sufficient set usually only retains 3-8 pieces of evidence, which can significantly reduce the context length and generation cost, and form a better accuracy-cost trade-off under a fixed token budget.

4 Experiment design and result analysis

4.1 Dataset and task setup

In order to systematically evaluate the applicability of the method in different knowledge-intensive scenarios, five representative datasets including HotpotQA, HOVER, FEVEROUS, QASPER and ALCE are selected to cover multi-hop question answering, fact checking, scientific literature question answering and long answer citation generation. HotpotQA is a typical multi-hop reasoning question answering task, which requires the model to synthesize two or more Wikipedia articles to generate the answer. The experiment uses the distractor setting, each query contains 10 gold standard evidence and 8 distractor evidence, and the development set and test set size are both 7,405. HOVER is oriented to the claim verification task. It requires the system to retrieve supporting evidence from Wikipedia and decide whether the claim is SUPPORTED, REFUTED or NOT_ENOUGH_INFO. Experiments are conducted on the complete test set, which contains 7,171 samples, and each claim corresponds to a maximum of 100 candidate paragraphs. FEVEROUS further introduces tabular and textual mixed evidence on the basis of traditional fact checking, and the task requires not only to complete the authenticity judgment, but also to identify effective evidence sets. The experiment uses a blind test set with 5,000 instances.

The QASPER test set contains 1,726 questions, and the average document length is 5,427 tokens, which puts higher requirements on evidence screening and long context modeling. ALCE belongs to the long answer question answering task with automatic citation generation, which requires the model to select corresponding supporting evidence for each statement when generating multi-sentence responses. The experiments are evaluated on ASQA and ELI5 subsets with 948 and 500 items respectively. HOVER and FEVEROUS emphasize evidence verification and conclusion judgment. QASPER emphasizes scientific semantic understanding of long documents. ALCE further requires answer generation and evidence reference to be completed simultaneously. Through unified evaluation on these five types of tasks, the performance of the proposed method in terms of evidence selection minimization, support adequacy and cross-scenario generalization ability can be comprehensively tested.

4.2 Comparison of methods and experimental environments

The comparison methods are divided into three categories. Traditional relevance retrieval methods include BM25 + Top-K, DPR + Top-K and MMR. BM25 is used for sparse recall, DPR is used for dense semantic recall, and MMR is used to suppress redundancy on the basis of relevance. ColBERT + Cross-Encoder pipeline was used in the neural reranking method to improve the candidate ranking accuracy with fine-grained interaction. The verifiable RAG method selects Self-RAG and LLM-as-Judge, and controls the retrieval generation process through self-reflection and the large model prompting evidence screening to complete the comparison respectively. The above Settings cover the main technical routes from relevance retrieval, neural refinement to verifiable generation.

The experimental environment uses a unified retrieval-reranking -search framework. In the candidate recall phase, BM25 and dense retrieval were used to generate the initial candidate pool, and K0=100 was set. Compressed to K1=50 after lightweight reordering; The graph

enhanced causal reordering retains K2=25. The encoder backbone used BERT-large or DeBERTa-v3, and the graph model used 3-layer R-GCN with 256 hidden dimensions. The quality assessment and scoring module uses a two-layer MLP. In the training phase, AdamW optimizer was used, learning rate was 2×10^{-5}, batch size was 16, and early stopping was performed according to the validation set MAE. The graph module freezes training before jointly fine-tuning with a learning rate of 5×10^{-6}. Key super parameters including α=0.77, θ=0.75, ε=0.05, μ=0.05, κ=0.2 and δ=0.02.

4.3 Evaluation index

In order to comprehensively evaluate the performance of the model in the construction task of "sufficient evidence set of verifiable causality", this paper divides the evaluation indicators into four groups. The first group is the answer quality indicators, which are used to measure the correctness of the final generated results or judgment results, mainly including Accuracy, Exact Match, F1-score, etc. The second group is the evidence minimality index, which is used to measure whether the evidence set is compact enough under the premise of satisfying the answer quality constraint, mainly using the average number of evidence, the average context length and Minimality@k. The third group is Deletion Sensitivity index, which is used to test the necessity of the selected evidence, that is, the decline of the answer quality after deleting a certain evidence, which can be described by the average deletion loss and the normalized deletion sensitivity. The fourth group is Citation Coverage and verifiability indicators, which is used to measure whether the key entities, values and conclusions in the answers are fully supported by evidence, mainly including citation coverage, NLI-based Entailment and Contradiction Rate. Through the joint evaluation of the above four groups of indicators, the experiment can not only investigate whether the model "answers correctly", but also further characterize "whether the evidence is few enough, whether it is irreplaceable, and whether it really supports the answer", so as to more accurately reflect the comprehensive performance of the method in the selection of sufficient evidence sets for verifiable causality.

4.4 Overall experimental results

The overall experimental results show that the proposed method achieves the improvement of answer quality and the compression of evidence cost on multiple datasets at the same time, which verifies the effectiveness of causal sufficient evidence selection for retrieval enhancement generation.

Table 1: Answer Quality and Evidence Selection Performance

Method	HotpotQA EM	HotpotQA F1	#Pass	HOVER EM	HOVER F1	#Pass	QASPER F1	#Pass	ALCE-ASQA Cit-P	ALCE-ASQA Cit-R	#Pass
BM25 + Top-5	42.1	55.3	5.0	68.4	73.2	5.0	28.7	5.0	62.3	51.8	5.0
DPR + Top-5	48.7	61.2	5.0	72.6	77.8	5.0	31.4	5.0	67.1	54.3	5.0
ColBERT + Rerank	51.3	63.8	5.0	74.2	79.1	5.0	33.2	5.0	68.9	56.7	5.0
MMR	49.8	62.1	4.8	73.1	78.3	4.7	32.1	4.9	66.5	58.2	4.8
Self-RAG	52.7	64.9	4.3	75.8	80.4	4.1	34.6	4.5	71.2	59.8	4.2
LLM-as-Judge	53.1	65.3	3.9	76.2	80.9	3.7	35.1	4.2	72.8	61.4	3.8
CER (Ours)	54.8	67.1	3.2	77.9	82.3	2.9	36.7	3.4	75.6	64.9	3.1
Improvement	+1.7	+1.8	-18%	+1.7	+1.4	-22%	+1.6	-19%	+2.8	+3.5	-18%

Table 1 shows that CER achieves the best results on HotpotQA, HOVER, QASPER and ALCE-ASQA: The EM and F1 of HotpotQA reached 54.8 and 67.1, respectively, which were 1.7 and 1.8 percentage points higher than the strongest baseline LLM-as-Judge. HOVER achieves EM and F1 of 77.9 and 82.3, respectively, which are 1.7 and 1.4 percentage points higher than the baseline. QASPER's F1 improved to 36.7. citation precision and citation recall of ALCE-ASQA reach 75.6% and 64.9%, respectively, which are 2.8 and 3.5 percentage points higher than the baseline. At the same time, the average number of evidence in the four types of tasks of CER is reduced to 3.2, 2.9, 3.4 and 3.1 respectively, which is 18%-22% smaller than that of LLM-as-Judge of 3.9, 3.7, 4.2 and 3.8. If compared with the fixed Top-5 method, the compression is more significant. This result shows that the evidence selection based on causal contribution modeling does not lose the answer quality due to compressing the context, but improves the overall generation accuracy and citation quality by eliminating redundant and weak supporting evidence.

From the perspective of evidence minimization and causal necessity, CER also shows obvious advantages.

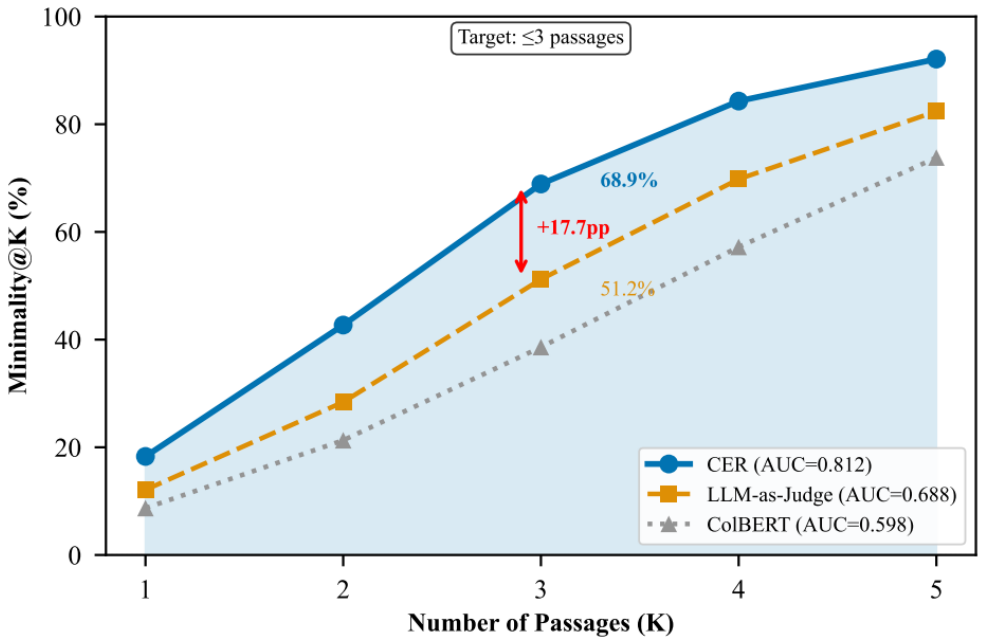

Figure 5: Minimality comparison under fixed quality thresholds

Figure 5 shows that on HotpotQA, when the target quality threshold is set to F1≥60, 68.9% of the samples can reach the target with no more than three pieces of evidence, while the LLM-as-Judge and ColBERT+Rerank are 51.2% and 38.6%, respectively. CER leads by 17.7 percentage points at K=3 and achieves an AUC of 0.812.

Table 2: Deletion Sensitivity Analysis

Method	HotpotQA ΔF1	HOVER ΔAcc	QASPER ΔF1	Avg. Impact
ColBERT + Rerank	8.3 ± 3.7	7.1 ± 4.2	6.9 ± 3.9	7.4
MMR	9.1 ± 3.4	7.8 ± 3.8	7.5 ± 3.6	8.1
Self-RAG	11.2 ± 3.1	9.4 ± 3.5	9.1 ± 3.3	9.9
LLM-as-Judge	12.6 ± 2.8	10.7 ± 3.2	10.3 ± 3.0	11.2
CER (Ours)	15.8 ± 2.1	13.9 ± 2.4	13.6 ± 2.3	14.4

Table 2 further shows that the average deletion sensitivity of CER reaches 14.4, which is significantly higher than that of LLM-as-Judge (11.2) and ColBERT+Rerank (7.4), indicating that the selected evidence will cause a greater decline in answer quality after deletion, so it has stronger causal necessity and lower redundancy.

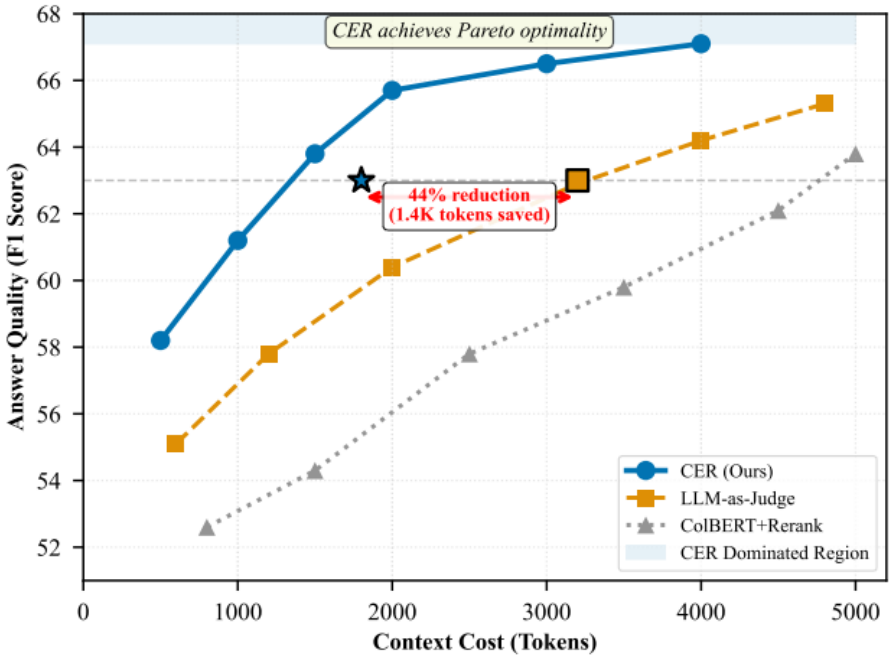

Figure 6: Accuracy-Cost Pareto frontier under fixed token budgets

The Accuracy-Cost Pareto curve given in Figure 6 further illustrates the advantage of the proposed method under a fixed token budget: On HotpotQA, when F1 remains at 63.0, CER only needs 1.8K tokens, while LLM-as-Judge needs 3.2K tokens, decreasing token overhead by 44%, saving about 1.4K tokens. Better quality-cost tradeoff is reflected.

Table 3: Robustness Under Adversarial Conditions (HotpotQA F1)

Method	Clean	Adv. Distract.	Near-Dup.	Outdated	Avg. Drop
ColBERT + Rerank	63.8	54.2 (-9.6)	58.7 (-5.1)	57.3 (-6.5)	-7.1
Self-RAG	64.9	57.8 (-7.1)	60.3 (-4.6)	59.1 (-5.8)	-5.8
LLM-as-Judge	65.3	59.7 (-5.6)	61.8 (-3.5)	60.9 (-4.4)	-4.5
CER (Ours)	67.1	63.4 (-3.7)	64.8 (-2.3)	64.2 (-2.9)	-3.0

The robustness experimental results are shown in Table 3. Under the three scenarios of adversarial interference, near-duplicate noise and outdated information, the average performance degradation of CER is only 3.0, while LLM-as-Judge is 4.5, ColBERT+Rerank is 7.1. It shows that the method still maintains high stability under the condition of conflict evidence and noise evidence injection. In general, CER achieves synergistic improvement in the four dimensions of answer quality, evidence quantity, token cost and anti-interference ability, indicating that the modeling of verifying causal sufficient evidence set can effectively improve the overall performance of retrieval enhancement generation system.

4.5 Ablation experiment

To examine the actual contribution of each computational module to the overall performance,

this paper conducts ablation analysis around counterfactual intervention, evidence graph modeling, set search and quality proxy calibration, and the results are shown in Table 4.

Table 4: Ablation Study on HotpotQA. Parentheses show change from full CER

Configuration	F1	Avg. #Pass	Del. Sens.
Full CER	67.1	3.2	15.8
- Replacement intervention	65.4 (-1.7)	3.5 (+0.3)	14.1 (-1.7)
- Evidence graph	64.8 (-2.3)	3.8 (+0.6)	13.6 (-2.2)
- Backward pruning	66.3 (-0.8)	4.1 (+0.9)	15.2 (-0.6)
- Uncertainty weighting	65.9 (-1.2)	3.4 (+0.2)	14.9 (-0.9)
- Calibration audits	64.2 (-2.9)	3.3 (+0.1)	13.8 (-2.0)
Deletion only ((\alpha=1.0))	65.6 (-1.5)	3.4 (+0.2)	14.3 (-1.5)
Relevance baseline (no counterfactual)	61.2 (-5.9)	4.7 (+1.5)	9.4 (-6.4)

The complete model achieves F1 of 67.1, average number of evidence 3.2, and deletion sensitivity of 15.8 on HotpotQA, which are the best results among all configurations. After removing the substitution intervention, F1 decreased to 65.4, the average number of evidence increased to 3.5, and the deletion sensitivity decreased to 14.1, indicating that it was difficult to fully suppress retrieval bias by only relying on the deletion counterfactual signal, and the substitution intervention had an obvious effect on identifying the "semantic close but factual conflict" pseudo-supporting evidence. Table 4 also shows the Deletion only (α=1.0) configuration, whose F1 is 65.6, the average number of evidence is 3.4, and the deletion sensitivity is 14.3, which are also lower than the full model, indicating that there is a complementary relationship between deletion intervention and replacement intervention, and the combination of the two can provide more stable causal supervision signals.

The contribution of graph structure modeling is equally obvious as the search pruning module. After removing the evidence graph, F1 decreases by 2.3 percentage points to 64.8, the average number of evidence increases to 3.8, and the deletion sensitivity decreases to 13.6, indicating that graph neural network plays a key role in complementary evidence recognition, redundant relationship compression, and conflict evidence suppression. After removing the reverse pruning, the F1 only decreases to 66.3, but the average number of evidence increases to 4.1, which is the highest among all configurations, indicating that the pruning mechanism in the minimum set search is mainly responsible for controlling the evidence size and maintaining the compact of the set. After removing the uncertainty weighting, F1 drops to 65.9, indicating that the uncertainty information provided by MC Dropout helps to avoid premature exclusion of potentially key evidence. After removing the quality proxy calibration, F1 is reduced to 64.2, and the deletion sensitivity is reduced to 13.8, which shows that temperature scaling, isometric regression and periodic auditing play a decisive role in maintaining the stability of quality estimation. On the whole, each module is not a simple superposition relationship, but jointly supports the formation of the minimum sufficient evidence set at the four levels of "causal supervision, structural modeling, set optimization, and quality calibration".

4.6 Robustness and parameter sensitivity analysis

In order to evaluate the stability of the model under different search conditions and noise interference environments, this paper conducts robustness and parameter sensitivity analysis from four dimensions: the quality threshold θ, the candidate scale K2, the proportion of conflict evidence and the proportion of near-duplicate interference. The effect of parameter variation on model performance versus evidence scale is shown in Figure 7. As can be seen from Figure

7(a), as the quality threshold θ increases from 0.65 to 0.85, the average number of evidence items of the model continues to increase, and the F1 index on HotpotQA shows a trend of first increasing and then decreasing, and achieves an optimal balance at θ=0.75. As can be seen from Figure 7(b), when the candidate size K2 increases from 10 to 25, the performance of the model is significantly improved. When K2 continues to increase, the performance gain tends to be saturated, but the inference overhead continues to rise, indicating that too large candidate pool will weaken the search efficiency.

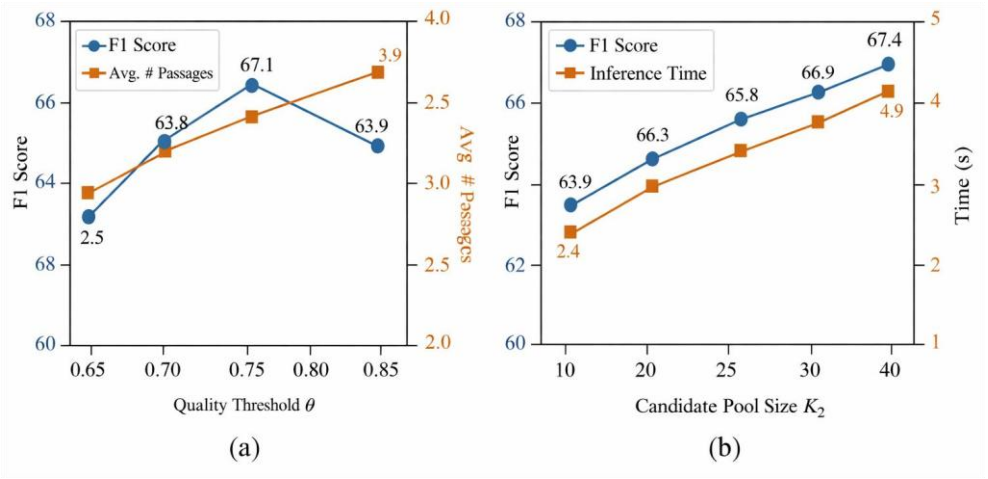

Figure 7: Parameter sensitivity analysis under different quality thresholds and candidate sizes.

The performance changes under noise injection conditions are shown in Figure: 8. As can be seen from Figure 8(a), with the proportion of conflicting evidence increasing from 0 to 40%, although the F1 of CER decreases, the overall decrease is significantly lower than that of correlation retrieval and common neural re-ranking methods, indicating that the proposed method has a stronger ability to suppress the sources of factual conflicts. Figure 8(b) shows that CER still maintains a relatively stable performance when the proportion of near-duplicate interference gradually increases, indicating that counterfactual supervision and evidence graph modeling can effectively weaken the misleading effect of semantic duplicate evidence on evidence selection results. In conclusion, CER shows good robustness under the condition of parameter perturbation and noise interference.

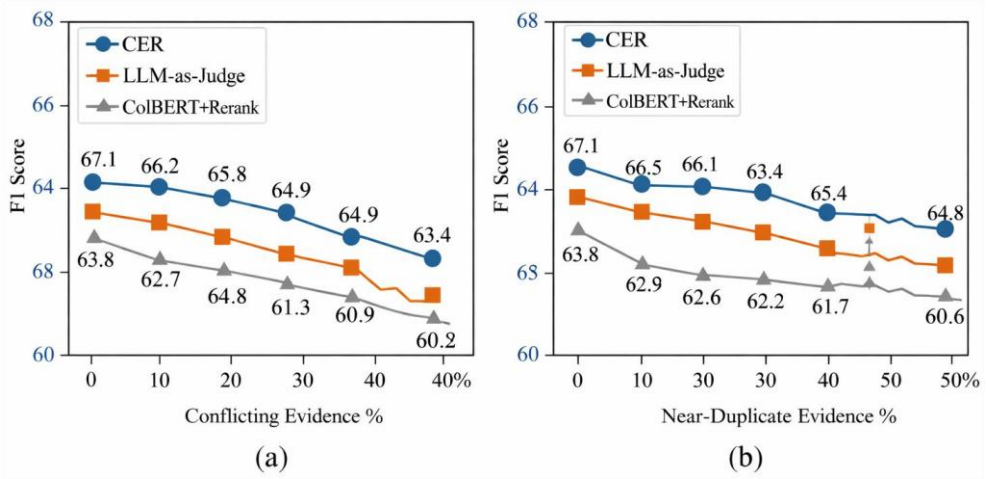

Figure 8: Robustness analysis under conflicting and near-duplicate evidence perturbations

4.7 Discussion of Results

To sum up, it can be seen that the high correlation does not mean that the evidence is causally necessary for the answer to hold. Relevance retrieval pays more attention to the semantic proximity between the query and the paragraph, so it is easy to retain texts with similar topics, repeated information or only background explanation. Such evidence can improve the surface matching score, but may not lead to a significant decline in answer quality after removal. The deletion sensitivity results show that the evidence retained by CER will cause more obvious performance attenuation once removed, indicating that its screening results are closer to the true support chain. The role of graph neural network is to transform the evidence from independent scoring objects into linkage nodes in the relationship graph, and the context information is propagated through entity overlap, time consistency, source consistency and contradiction, so that the model can identify the complementary evidence combinations of "single insufficient and joint valid", and suppress the local high correlation but global redundant candidates. The minimum sufficient search further pushes the evidence selection from the ranking problem to the constrained optimization problem, and eliminates the paragraphs with low marginal contribution under the premise of ensuring the answer quality reaches the standard, so it can effectively reduce the average number of evidence and token overhead, and shorten the manual verification link. For auditors, shorter and irreducible evidence sets mean clearer supporting boundaries, lower verification costs and stronger traceability, which is the fundamental reason why CER outperforms the comparison methods on the accuracy-cost Pareto frontier.

5 Conclusion

Focusing on the key problem of "relevant evidence is not equal to necessary evidence" in retrieval enhancement generation, this paper constructs a task of selecting verifiable causal sufficient evidence set, which transforms the traditional relevance ranking into the minimum sufficient evidence subset search under quality constraints. In terms of method, a unified algorithm framework is formed, which consists of counterfactual supervision signal construction, pre-trained language model causal rerankinging, evidence graph complementary modeling, quality surrogate function calibration, and minimum sufficient set search. The experimental results show that the proposed method achieves both answer quality improvement and evidence cost compression on HotpotQA, HOVER, QASPER and ALCE tasks, and has a better Accuracy-cost Pareto frontier under the same token budget. The average evidence size is reduced by about 18%-22%, the context overhead is reduced by up to 60%, and it shows stronger stability in the scene of conflict evidence and near-repeated interference. The above results show that modeling evidence from the perspective of causal contribution rather than semantic correlation can more effectively improve the verifiability, compactness and robustness of retrieval enhancement generation system. At the same time, there are still some aspects to be deepened in this paper. First, although the quality proxy function reduces the generation overhead in the search phase, there may still be an estimation bias between the quality and the true answer quality. Secondly, the delete-replace counterfactual supervision needs to construct additional intervention samples, which leads to high training and calibration costs. Third, the current framework mainly focuses on textual evidence, and has not integrated tabular evidence, numerical evidence and multimodal evidence. Further research can be carried out in two directions: one is to build an end-to-end joint optimization mechanism of selector-generator to reduce the accumulation of proxy errors; The second is to extend to tables, images and multi-source heterogeneous knowledge scenarios, and develop a verifiable RAG method for multi-

modal evidence chains.

References

[1] Ram O, Levine Y, Dalmedigos I, et al. In-context retrieval-augmented language models[J]. Transactions of the Association for Computational Linguistics, 2023, 11: 1316-1331.

[2] Siriwardhana S, Weerasekera R, Wen E, et al. Improving the domain adaptation of retrieval augmented generation (RAG) models for open domain question answering[J]. Transactions of the Association for Computational Linguistics, 2023, 11: 1-17.

[3] Russo D, Tekiroğlu S S, Guerini M. Benchmarking the generation of fact checking explanations[J]. Transactions of the Association for Computational Linguistics, 2023, 11: 1250-1264.

[4] Park S I, Lee J Y. Toward robust ralms: Revealing the impact of imperfect retrieval on retrieval-augmented language models[J]. Transactions of the Association for Computational Linguistics, 2024, 12: 1686-1702.

[5] Rubin O, Berant J. Retrieval-pretrained transformer: Long-range language modeling with self-retrieval[J]. Transactions of the Association for Computational Linguistics, 2024, 12: 1197-1213.

[6] Adlakha V, BehnamGhader P, Lu X H, et al. Evaluating correctness and faithfulness of instruction-following models for question answering[J]. Transactions of the Association for Computational Linguistics, 2024, 12: 681-699.

[7] Glockner M, Staliūnaitė I, Thorne J, et al. Ambifc: Fact-checking ambiguous claims with evidence[J]. Transactions of the Association for Computational Linguistics, 2024, 12: 1-18.

[8] Zeng F, Gao W. Justilm: Few-shot justification generation for explainable fact-checking of real-world claims[J]. Transactions of the Association for Computational Linguistics, 2024, 12: 334-354.

[9] Katsis Y, Rosenthal S, Fadnis K, et al. MTRAG: A multi-turn conversational benchmark for evaluating retrieval-augmented generation systems[J]. Transactions of the Association for Computational Linguistics, 2025, 13: 784-808.

[10] Rosenthal S, Sil A, Florian R, et al. CLAPnq: C ohesive L ong-form A nswers from P assages in Natural Questions for RAG systems[J]. Transactions of the Association for Computational Linguistics, 2025, 13: 53-72.

[11] Amar S, Shapira O, Slobodkin A, et al. A Unifying Scheme for Extractive Content Selection Tasks[J]. Transactions of the Association for Computational Linguistics, 2025, 13: 1645-1671.

[12] Li X, Jin J, Zhou Y, et al. From matching to generation: A survey on generative information retrieval[J]. ACM Transactions on Information Systems, 2025, 43(3): 1-62.

[13] Huang L, Yu W, Ma W, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions[J]. ACM Transactions on Information Systems, 2025, 43(2): 1-55.

[14] Ji Z, Lee N, Frieske R, et al. Survey of hallucination in natural language generation[J]. ACM computing surveys, 2023, 55(12): 1-38.

[15] Lyu Y, Li Z, Niu S, et al. Crud-rag: A comprehensive chinese benchmark for retrieval-augmented generation of large language models[J]. ACM Transactions on Information Systems, 2025, 43(2): 1-32.

[16] Hosking T, Tang H, Lapata M. Hierarchical indexing for retrieval-augmented opinion summarization[J]. Transactions of the Association for Computational Linguistics, 2024, 12: 1533-1555.

[17] Lyu Q, Apidianaki M, Callison-Burch C. Towards faithful model explanation in nlp: A survey[J]. Computational Linguistics, 2024, 50(2): 657-723.

[18] Chaturvedi A, Bhar S, Saha S, et al. Analyzing semantic faithfulness of language models via input intervention on question answering[J]. Computational Linguistics, 2024, 50(1): 119-155.

[19] Wang D, Huang Q, Jackson M, et al. Retrieve what you need: A mutual learning framework for open-domain question answering[J]. Transactions of the Association for Computational Linguistics, 2024, 12: 247-263.

[20] Liu N F, Lin K, Hewitt J, et al. Lost in the middle: How language models use long contexts[J]. Transactions of the association for computational linguistics, 2024, 12: 157-173.