LOG: A Local-to-Global Optimization Approach for Retrieval-based Explainable Multi-Hop Question Answering

Hao Xu[1*], Yunxiao Zhao[1*†], Jiayang Zhang[1], Zhiqiang Wang[1,2], Ru Li[1,2]
1. School of Computer and Information Technology, Shanxi University, Taiyuan, China
2. Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan, China
xuhao0050@163.com, yunxiaomr@163.com, {liru, wangzq}@sxu.edu.cn

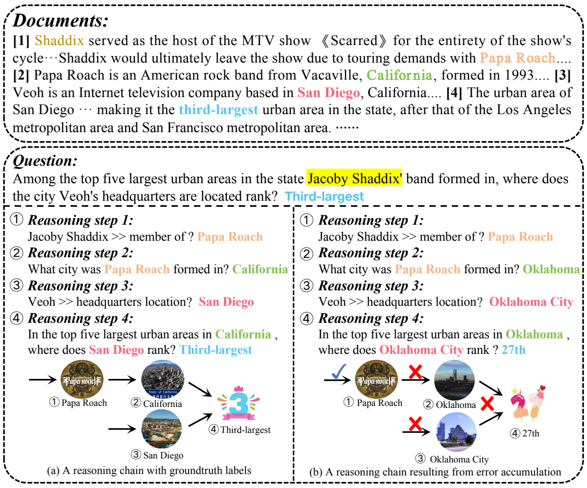
Abstract

Multi-hop question answering (MHQA) aims to utilize multi-source intensive documents retrieved to derive the answer. However, it is very challenging to model the importance of knowledge retrieved. Previous approaches primarily emphasize single-step and multi-step iterative decomposition or retrieval, which are susceptible to failure in long-chain reasoning due to the progressive accumulation of erroneous information. To address this problem, we propose a novel **L**ocal-t**O**-**G**lobal optimized retrieval method ($\mathcal{LOG}$) to discover more beneficial information, facilitating the MHQA. In particular, we design a pointwise conditional $\mathcal{V}$-information based local information modeling to cover usable documents with reasoning knowledge. We also improve tuplet objective loss, advancing multi-examples-aware global optimization to model the relationship between scattered documents. Extensive experimental results demonstrate our proposed method outperforms prior state-of-the-art models, and it can significantly improve multi-hop reasoning, notably for long-chain reasoning. Our code is available at https://github.com/yunxiaomr/LOG.

1 Introduction

Multi-Hop Question Answering (MHQA) requires multi-hop reasoning using intensive knowledge from multiple scattered documents to derive the answer (Xu et al., 2024; Shi et al., 2024; Yang et al., 2018). As shown in Figure 1(a), a multi-hop example with four hops is illustrated, where the MHQA model needs to reason the multi-hop question from different scattered documents. This multi-hop paradigm can facilitate the development of explainable systems (Thayaparan et al., 2022).

Traditional approaches have produced promising results on the MHQA task, notably using graph neural networks (Qiu et al., 2019; Fang et al., 2020; Huang and Yang, 2021; Ramesh et al., 2023) as an encoder to obtain the representation of contexts and predict the final multi-hop answer. However, these methods primarily focus on modeling implicit, black-box reasoning while neglecting the use of interpretable, step-by-step reasoning. Besides, they are also dependent on constructing graphs at the computational efficiency.



Documents:
[1] Shaddix served as the host of the MTV show 《Scarred》for the entirety of the show's cycle···Shaddix would ultimately leave the show due to touring demands with Papa Roach.... **[2]** Papa Roach is an American rock band from Vacaville, California, formed in 1993.... **[3]** Veoh is an Internet television company based in San Diego, California.... **[4]** The urban area of San Diego ··· making it the third-largest urban area in the state, after that of the Los Angeles metropolitan area and San Francisco metropolitan area. ······

Question:
Among the top five largest urban areas in the state Jacoby Shaddix' band formed in, where does the city Veoh's headquarters are located rank? Third-largest

① ***Reasoning step 1:***
Jacoby Shaddix >> member of ? Papa Roach
② ***Reasoning step 2:***
What city was Papa Roach formed in? California
③ ***Reasoning step 3:***
Veoh >> headquarters location? San Diego
④ ***Reasoning step 4:***
In the top five largest urban areas in California , where does San Diego rank? Third-largest

(a) A reasoning chain with groundtruth labels

① ***Reasoning step 1:***
Jacoby Shaddix >> member of ? Papa Roach
② ***Reasoning step 2:***
What city was Papa Roach formed in? Oklahoma
③ ***Reasoning step 3:***
Veoh >> headquarters location? Oklahoma City
④ ***Reasoning step 4:***
In the top five largest urban areas in Oklahoma , where does Oklahoma City rank ? 27th

(b) A reasoning chain resulting from error accumulation

Figure 1: A multi-hop example with 4 hops is illustrated, with (a) showing the reasoning chains with groundtruth labels and (b) showing those with accumulated errors.

On the other hand, some iterative-enhanced methods involving multi-step decomposition (Fu et al., 2021; Mao et al., 2022; Deng et al., 2022; Wu et al., 2024) and iterative retrieval (Qi et al., 2021; Trivedi et al., 2022; Zhao et al., 2023; Zhang et al., 2024) have been proposed to explicitly and controllably model the relationships between multiple-hop steps. For example, they leverage structured knowledge, such as Wikipedia (Asai et al., 2020; Qi et al., 2021) and abstract meaning representation (Deng et al., 2022), to bridge multi-hop nodes by methods like recomposition (Perez et al., 2020) and iterative enhancement (Tu et al., 2020). Although the MHQA has been improved, these methods remain

[*] Equal contribution. Author ordering determined by coin flip, following previous research (Kingma and Ba, 2015).
[†] Corresponding author.

Proceedings of the 31st International Conference on Computational Linguistics, pages 9085–9095
January 19–24, 2025. ©2025 Association for Computational Linguistics

vulnerable to *accumulation of erroneous information* through single-step or multi-step iterations, which can ultimately lead to failures of multi-hop inference. This is because of the incomplete modeling of required knowledge for reasoning, resulting in cascading biases. As Figure 1(b) shows, due to the failure to retrieve the correct documents for *reasoning step 2* and *reasoning step 3*, the relevant knowledge can not be modeled, which ultimately causes an incorrect answer in the *reasoning step 4*.

To alleviate this problem, we focus on retrieval-enhanced MHQA, which aims to utilize useful scattered documents via retrieve-then-read paradigm to reason the answer. A key problem is to highlight valuable information. To this end, as shown in Figure 2, we propose a novel **L**ocal-t**O**-**G**lobal optimized retrieval method ($\mathcal{LOG}$), which includes *local information modeling* and *global objective optimization* to discover beneficial information for multi-hop reasoning. Our contributions are as follows: 1) We propose a novel $\mathcal{LOG}$ method that models usable knowledge from a local-to-global perspective, facilitating the MHQA; 2) We introduce pointwise conditional $\mathcal{V}$-information to quantify the contribution of local documents to the target prediction; we design multi-examples-aware objective optimization to model the relationship between documents globally; 3) Extensive experimental results show $\mathcal{LOG}$ outperforms prior state-of-the-art models, and it can significantly improve multi-hop reasoning, especially for long-chain reasoning.

2 Methodology

Notations. Following (Zhao et al., 2023; Mavi et al., 2024), let $\mathbb{C}$ denote the set of all documents, $\mathbb{S}$ denote the set of all questions and $\mathbb{A}$ denote the set of all possible answers. The MHQA task aims to approximate a function $f : \mathbb{S} \times \mathbb{C}^n \mapsto \mathbb{A} \cup \{\varnothing\}$, which needs to satisfy:

$$f(C,q) = \begin{cases} a \in \mathbb{A} & \exists P_q = \{p_1, \cdots, p_k\} \subseteq C, \\ & k > 1, \\ & P_q \models (q \to a) \\ \varnothing & otherwise \end{cases} \tag{1}$$

Typically, our retrieval-enhanced MHQA can be represented as $f(C,q) = f_h(q, f_g(C,q))$ where a retriever $f_g : \mathbb{S} \times \mathbb{C}^n \mapsto \mathbb{C}^k$ and a reader $f_h : \mathbb{S} \times \mathbb{C}^k \mapsto \mathbb{A} \cup \{\varnothing\}$ are employed. In this process, a reasoning chain for a question $R'_q = \{r'_{q,i}\}_{i=1}^k$ is defined as an ordered sequence of the set P_q defined above, such that: $\forall\, j,\, 1 \leq j < k,\, r'_{q,j} \to r'_{q,j+1}$ represents the j^{th} reasoning step and $r'_{q,k} \to a$ is the k-th reasoning step.

2.1 Pointwise conditional $\mathcal{V}$-information based Local Information Modeling

Local information modeling aims to seek usable documents' information for target prediction under a constrained question in the MHQA. To model this information, inspired by (Chen et al., 2023; Jiang et al., 2024), the mutual information between the documents P_q and their labels Y, $I(Y; P_q)$, can be helpful for quantization. However, we are more interested in modeling the informativeness obtained by the model under certain constraints. Therefore, we first quantify the contribution of documents under the condition with a constrained question using conditional $\mathcal{V}$-information (`CVI`) (Hewitt et al., 2021). The computation process can be represented by

$$\begin{aligned} & I_{\mathcal{V}}(P_q \to Y|q) \\ & = H_{\mathcal{V}}(Y|q) - H_{\mathcal{V}}(Y|q, P_q) \end{aligned} \tag{2}$$

where $H_{\mathcal{V}}(\cdot|\cdot)$ is conditional $\mathcal{V}$ entropy[1]. Because the `CVI` directly measures the contribution of multiple documents to question q at a macroscopic level, the imbalance between irrelevant and relevant document samples can lead to a reduction in overall usable information, as the abundance of irrelevant documents dilutes the informativeness. Thus we propose document-aware pointwise conditional $\mathcal{V}$-information (`PCVI`) for each document to model locally valuable information that contributes to the muti-hop question q. Specifically, the contribution of i-th document p_i to target prediction can be formalized as the following formula,

$$\begin{aligned} & \texttt{PCVI}(p_i \to y \mid q) \\ & = -\log_2 f[q](y) + \log_2 f[p_i, q](y) \end{aligned} \tag{3}$$

where $f \in \mathcal{V}$, s.t.$\mathbb{E}[-\log f[q](y)] = H_{\mathcal{V}}(y\,|q)$ and $\mathbb{E}[-\log f[p_i, q](y)] = H_{\mathcal{V}}(y|p_i, q)$. Furthermore, we optimize the final local loss by minimizing the negative mutual information of `PCVI`.

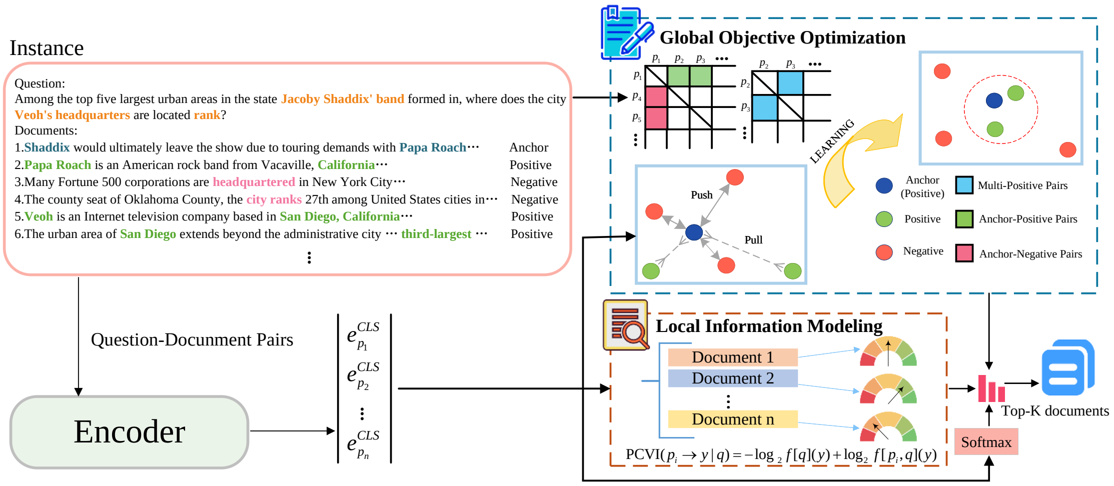
$$\mathcal{L}_{\mathcal{V}} = -\frac{1}{n}\sum_{i=1}^{n} \texttt{PCVI}(i), \tag{4}$$

where n indicates the total number of documents.

2.2 Multi-Positive-Negative Examples-aware Global Objective Optimization

Global objective optimization aims to model the relationship between documents globally, optimizing the distance between different documents.

[1] The details on the calculation of $\mathcal{V}$ information and conditional $\mathcal{V}$ entropy can be found in Appendix A.1.



Instance

Question:
Among the top five largest urban areas in the state **Jacoby Shaddix' band** formed in, where does the city **Veoh's headquarters** are located **rank**?

Documents:

1.**Shaddix** would ultimately leave the show due to touring demands with **Papa Roach**··· Anchor

2.**Papa Roach** is an American rock band from Vacaville, **California**··· Positive

3.Many Fortune 500 corporations are **headquartered** in New York City··· Negative

4.The county seat of Oklahoma County, the **city ranks** 27th among United States cities in··· Negative

5.**Veoh** is an Internet television company based in **San Diego, California**··· Positive

6.The urban area of **San Diego** extends beyond the administrative city ··· **third-largest** ··· Positive

⋮

Figure 2: The overview of our proposed method $\mathcal{LOG}$.

Negative Examples-aware Global Objective. Inspired by (Schroff et al., 2015; Sohn, 2016), we first adapt the (N+1)-tuplet loss[2]. Given a set of documents $\{p_1 \cdots, p_n\}$ and the corresponding labels, we randomly select a document p with y = 1 as anchor and another with y = 1 as positive example p^+. Those documents with y = 0 serve as negative examples $\{p_i\}_{i=1}^{N-1}$. The (N+1)-tuplet can be modeled by

$$\mathcal{L}(\{p, p^+, \{p_i\}_{i=1}^{N-1}\}; f(\cdot;\theta)) \\ = \log(1 + \sum_{i=1}^{N-1} \exp(f^\top f_i - f^\top f^+)). \tag{5}$$

We encode all examples using a pre-trained model to obtain `[CLS]` token representations and then compute pairwise distances. The distance measurement function is as follows:

$$d(x, y) = \frac{x^T y}{\|x\| \|y\|} \tag{6}$$

Multi-Positive Examples-aware Global Objective. Though Eq.5 can distinguish irrelevant documents by pushing away negative examples, documents in the MHQA related to reasoning are often not just one but rather scattered. To aggregate these useful documents (i.e., positive examples), we model the multi-positive examples based on Eq.5, which can be represented:

$$\mathcal{L}_{Tri}(\{p, p^+, \{p_i\}_{i=1}^{M-1}, \{p_j\}_{j=1}^{N-M}\}; f(\cdot;\theta)) \\ = \log(1 + \sum_{j=1}^{N-M} \exp(f(p)^\top f(p_j) \\ - \frac{1}{M}(\sum_{i=1}^{M-1} f(p)^\top f(p_i) + f(p)^\top f(p^+))), \tag{7}$$

where M and N denote the number of positive examples and all examples excluding anchor.

2.3 Training and Prediction

For training, we perform joint modeling using the local loss $\mathcal{L}_\mathcal{V}$ and the global loss $\mathcal{L}_{Tri}$. We also include the cross-entropy loss $\mathcal{L}_{CE}$ to supervise two-classfication labels. The final loss function is $\mathcal{L} = \lambda_1\mathcal{L}_\mathcal{V} + \lambda_2\mathcal{L}_{Tri} + \lambda_3\mathcal{L}_{CE}$, where $\lambda_1, \lambda_2, \lambda_3 \in (0, 1)$. For prediction, the encoder first obtains the embedding $e_{p_i}^{CLS}$ of (q, p_i) pair, then the probability distribution p_i of (q, p_i) pair is computed by performing a binary classification with $e_{p_i}^{CLS}$ as input by

$$p_i = \mathrm{softmax}(W_i e_{p_i}^{CLS} + b_i), \tag{8}$$

where W_i and b_i are learnable parameters.

3 Experiments

3.1 Datasets, Baselines and Evaluation Metrics

Datasets and Baselines. We first compare our proposed $\mathcal{LOG}$ on MusiQue-Ans (Trivedi et al., 2022), which has 19,938 train, 2,417 development, and 2,459 test examples, respectively. The dataset includes 2-4 hop questions, answers, and a collection of 20 documents as context per question. Specifically, we compare three end-to-end models: **FiD**, **FiD$_{\mathbf{+PT}}$** (Izacard and Grave, 2021) and **EE** (Trivedi et al., 2022); three decomposition-based models: **EX(EE)** (Trivedi et al., 2022), **EX(SA)** (Trivedi et al., 2022) and **HPE** (Liu et al., 2023b); two retrieval-based models: **SA** (Trivedi et al., 2022) and **HUG** (Zhao et al., 2023); and an early method: **RNN-based baseline** (Yang et al., 2018). In addition, we also conduct experiments on HotpotQA in

[2] The details of the (N+1)-tuplet loss are in Appendix A.2.

Methods	RE Performance		QA Performance	
	Support_EM	Support_F1	Answer_EM	Answer_F1
RNN (Yang et al., 2018)	-	41.9	-	13.6
EE (Trivedi et al., 2022)	21.5	67.6	34.6	42.3
SA* (Trivedi et al., 2022)	30.4	72.3	39.3	47.3
EX(EE) (Trivedi et al., 2022)	48.8	77.8	38.4	45.6
EX(SA) (Trivedi et al., 2022)	53.5	79.2	41.5	49.7
HUG (Zhao et al., 2023)	-	44.4	-	39.1
$\mathcal{LOG}$	33.3	76.7	**42.7**	**50.7**
$\mathcal{LOG}$ (-w/o $\mathcal{L}$)	33.1	76.6	42.1	50.4
$\mathcal{LOG}$ (-w/o $\mathcal{G}$)	32.1	75.7	40.8	49.2

Table 1: RE (Retrieval) and QA (Question Answering) performance on the development set of MusiQue-Ans in comparison with previous work. * indicates the backbone model of our proposed $\mathcal{LOG}$.

Methods	Overall Performance	
	Support_F1	Answer_F1
FiD (Izacard and Grave, 2021)	-	45.3
FiD$_{+PT}$ (Izacard and Grave, 2021)	-	48.8
EE (Trivedi et al., 2022)	69.4	40.7
SA* (Trivedi et al., 2022)	74.7	49.7
EX(EE) (Trivedi et al., 2022)	78.1	46.4
EX(SA) (Trivedi et al., 2022)	80.6	49.0
HPE (Liu et al., 2023b)	-	50.1
$\mathcal{LOG}$ (ours)	78.6	**53.3**

Table 2: Overall performance on the test set of MusiQue-Ans in comparison with previous work.

the distractor setting (Yang et al., 2018). Appendix B provides further details.

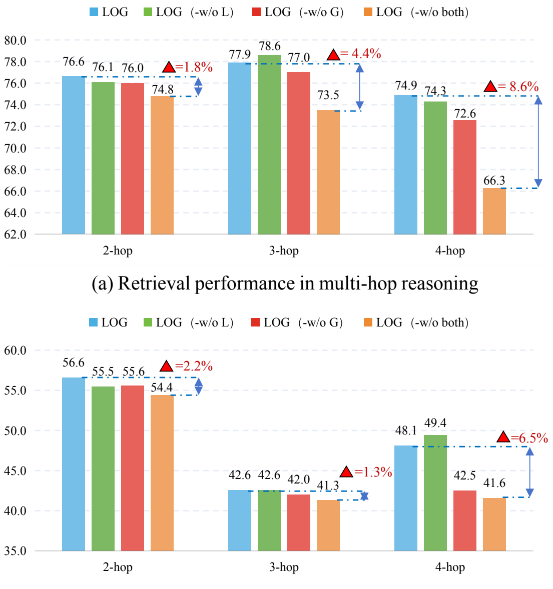
Evaluation Metrics. We employ Exact Match (EM) and F1 score to evaluate retrieval performance and also use EM and F1 for answer and support identification as our evaluation metrics.

3.2 Performance Comparison

Retrieval Results. We first compare the retrieval performance of $\mathcal{LOG}$. As shown in Table 1, we observe that $\mathcal{LOG}$'s performance improved by 2.9% in EM and 4.4% in F1 compared to the backbone model SA, highlighting the advantages of our proposed $\mathcal{LOG}$ in retrieval.

Multi-hop QA Results. We then compare the QA performance of $\mathcal{LOG}$. We can see that compared to recent decomposition-based and retrieval-based methods, $\mathcal{LOG}$ achieves the best performance on MusiQue-Ans (Table 1), and it shows a significant improvement (3.4%) over the backbone model. Besides, the results on HotpotQA demonstrate that $\mathcal{LOG}$ also maintains competitive performance, which focuses on shorter 2-hop questions[3].

Online Results. Furthermore, we also submit our test results online as shown in Table 2, which shows that our model once again obtains the best performance under the paradigm of iterative retrieval consistently.



(a) Retrieval performance in multi-hop reasoning

(b) QA performance in multi-hop reasoning

Figure 3: The results of retrieval and QA performance on the development set, including 2-4 hops.

Methods	MuSiQue-Ans	
	Support_F1	Answer_F1
$\mathcal{LOG}$	78.6	53.3
-w/o $\mathcal{L}$	77.3	52.4
-w/o $\mathcal{G}$	77.6	52.5
-w/o *both*	74.7	49.7

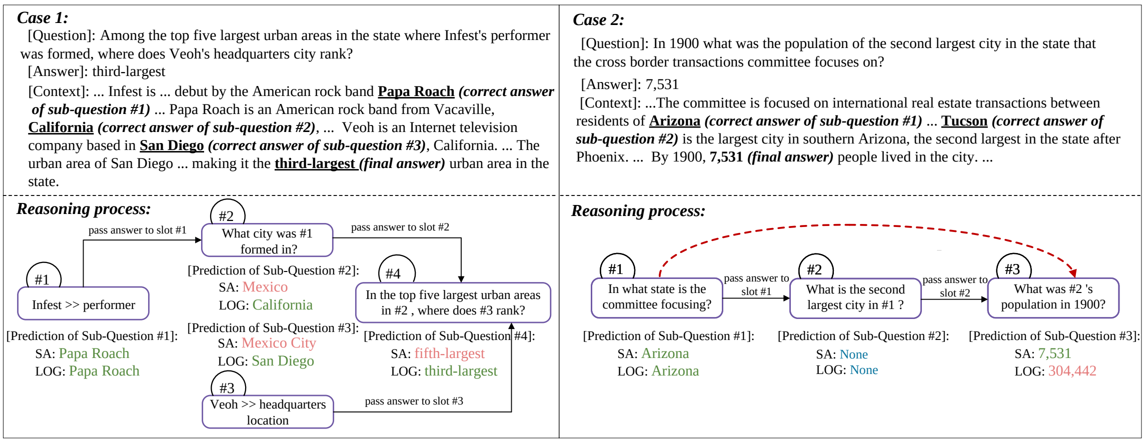
Table 3: Ablation study results on the test set.

3.3 Emperimental Analysis

Ablation Study. To further verify the effectiveness of $\mathcal{LOG}$, we also conduct the ablation study. As depicted in Table 3, the deceasing results for removing the Local module(-w/o $\mathcal{L}$), Global module(-w/o $\mathcal{G}$), or both modules underscore the critical roles played by our proposed two key components.

Different-hop Results. As shown in Figure 3, we also analyze $\mathcal{LOG}$'s performance on 2-4 hop questions. We can observe that our model $\mathcal{LOG}$ shows

[3] Please see more details in Appendix C.



Case 1:

[Question]: Among the top five largest urban areas in the state where Infest's performer was formed, where does Veoh's headquarters city rank?

[Answer]: third-largest

[Context]: ... Infest is ... debut by the American rock band **Papa Roach** ***(correct answer of sub-question #1)*** ... Papa Roach is an American rock band from Vacaville, **California** ***(correct answer of sub-question #2)***, ... Veoh is an Internet television company based in **San Diego** ***(correct answer of sub-question #3)***, California. ... The urban area of San Diego ... making it the **third-largest** ***(final answer)*** urban area in the state.

Reasoning process:

#1
Infest >> performer
pass answer to slot #1
#2
What city was #1 formed in?
pass answer to slot #2
#3
Veoh >> headquarters location
pass answer to slot #3
#4
In the top five largest urban areas in #2 , where does #3 rank?
[Prediction of Sub-Question #1]:
SA: Papa Roach
LOG: Papa Roach
[Prediction of Sub-Question #2]:
SA: Mexico
LOG: California
[Prediction of Sub-Question #3]:
SA: Mexico City
LOG: San Diego
[Prediction of Sub-Question #4]:
SA: fifth-largest
LOG: third-largest

Case 2:

[Question]: In 1900 what was the population of the second largest city in the state that the cross border transactions committee focuses on?

[Answer]: 7,531

[Context]: ...The committee is focused on international real estate transactions between residents of **Arizona** ***(correct answer of sub-question #1)*** ... **Tucson** ***(correct answer of sub-question #2)*** is the largest city in southern Arizona, the second largest in the state after Phoenix. ... By 1900, **7,531** ***(final answer)*** people lived in the city. ...

Reasoning process:

#1
In what state is the committee focusing?
pass answer to slot #1
#2
What is the second largest city in #1 ?
pass answer to slot #2
#3
What was #2 's population in 1900?
[Prediction of Sub-Question #1]:
SA: Arizona
LOG: Arizona
[Prediction of Sub-Question #2]:
SA: None
LOG: None
[Prediction of Sub-Question #3]:
SA: 7,531
LOG: 304,442

Figure 4: Case Study. The green font represents the correct predicted answer, the red font represents the incorrect predicted answer, and the blue font indicates that the model skipped the reasoning step without producing a predicted answer. The red dashed line represents the actual reasoning path of the SA and $\mathcal{LOG}$ models.

significant improvements in both retrieval and QA tasks involving 2-4 hop queries. Notably, $\mathcal{LOG}$ exhibits the most substantial improvement (8.6% and 6.5%) in long-step reasoning with 4 hops.

Case Study. Figure 4 shows two cases of reasoning process. Case 1 is an example of error propagation. For the second and third sub-questions, the answers predicted by the SA model are wrong, affecting the subsequent reasoning process, thus outputting the wrong final answer. In contrast, our proposed method effectively capture more usable reasoning knowledge, enabling it to arrive at the correct final answer. Case 2 demonstrates the challenge of reasoning shortcuts in MHQA, where the model is prone to skipping intermediate reasoning steps and taking shortcuts to obtain answers. $\mathcal{LOG}$ does not specifically address this issue; therefore, it obtains an incorrect final answer. For more cases and analysis, please refer to Appendix D.

4 Related Work

Multi-Hop Question Answering. Traditional approaches to solving the multi-hop QA problem can be mainly categorized as multi-step decomposition (Fu et al., 2021; Mao et al., 2022; Deng et al., 2022; Wu et al., 2024; Yan et al., 2024), graph-based method (Qiu et al., 2019; Fang et al., 2020; Huang and Yang, 2021; Ramesh et al., 2023), and iterative retrieval-based method (Qi et al., 2021; Trivedi et al., 2022; Zhao et al., 2023; Zhang et al., 2024). These methods have produced promising results. EX(EE) and EX(SA) (Trivedi et al., 2022) perform explicit multi-step reasoning by first decomposing the question into a DAG having single-hop sub-questions, and then calling single-hop model repeatedly. Beam Retrieval (Zhang et al., 2024) maintains multiple partial hypotheses of relevant documents at each step via beam search. Meanwhile, Various methods also demonstrate good interpretability. BreakRC (Wolfson et al., 2020), QDAMR (Deng et al., 2022), and HPE (Liu et al., 2023b) decompose multi-hop questions through structured semantic parsing methods, followed by iterative execution to obtain the final answer. SNMN (Jiang and Bansal, 2019) and Mao et al. (Mao et al., 2022) propose novel neural symbolic reasoning methods based on Neural Module Networks to enhances interpretability. SG Prompt (Li and Du, 2023) improves reasoning ability and interpretability by incorporating the sequentialized semantic graph into the prompt. HUG (Zhao et al., 2023) probabilistically models the dependency between documents and between sentences within a document, without requiring rationale supervision.

5 Conclusion

We propose a novel local-to-global optimized retrieval method to cover beneficial information, facilitating the explainable MHQA. We introduce pointwise conditional $\mathcal{V}$-information to quantify the contribution of individual documents to the target prediction, and improve tuple loss to model the relationship between different documents. The experimental results show that our method outperforms previous models and significantly improves MHQA, especially for long chains.

Limitation

Though multi-hop question answering can improve interpretability by providing a decomposition-based reasoning process (Thayaparan et al., 2022), the decomposition-based explanations are not always faithful to the model's predictions. Therefore, exploring the model's self-explanation mechanisms represents a promising direction for future research (Liu et al., 2024; Zhao et al., 2024; Liu et al., 2023a; Storek et al., 2023; Lei et al., 2016). In addition, it may be a good idea to improve structural-level semantic correlation between question and candidates by a hierarchy in the question-document pairs. Finally, utilizing our method to advance the reasoning and explanations of large language models remains the next direction to explore. These are the focus of our future research.

Ethics Statement

This paper does not involve the presentation of a new dataset and the utilization of demographic or identity characteristics information.

Acknowledgements

We would like to thank all the anonymous reviewers for their valuable feedback. This work was supported by the National Natural Science Foundation of China (Nos.62376144, 61906111, 62272285), the Science and Technology Cooperation and Exchange Special Project of Shanxi Province (No.202204041101016) and the Key Research and Development Project of Shanxi Province under Grant (No.202102020101008).

References

Akari Asai, Kazuma Hashimoto, Hannaneh Hajishirzi, Richard Socher, and Caiming Xiong. 2020. Learning to retrieve reasoning paths over wikipedia graph for question answering. In *International Conference on Learning Representations*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Hanjie Chen, Faeze Brahman, Xiang Ren, Yangfeng Ji, Yejin Choi, and Swabha Swayamdipta. 2023. REV: Information-theoretic evaluation of free-text rationales. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 2007–2030.

Zhenyun Deng, Yonghua Zhu, Yang Chen, Michael Witbrock, and Patricia Riddle. 2022. Interpretable amr-based question decomposition for multi-hop question answering. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4093–4099.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding dataset difficulty with $\mathcal{V}$-usable information. In *Proceedings of the 39th International Conference on Machine Learning*, pages 5988–6008.

Yuwei Fang, Siqi Sun, Zhe Gan, Rohit Pillai, Shuohang Wang, and Jingjing Liu. 2020. Hierarchical graph network for multi-hop question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8823–8838.

Ruiliu Fu, Han Wang, Xuejun Zhang, Jun Zhou, and Yonghong Yan. 2021. Decomposing complex questions makes multi-hop QA easier and more interpretable. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 169–180.

Asish Ghoshal, Srinivasan Iyer, Bhargavi Paranjape, Kushal Lakhotia, Scott Wen-tau Yih, and Yashar Mehdad. 2022. Quaser: Question answering with scalable extractive rationalization. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 1208–1218.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

John Hewitt, Kawin Ethayarajh, Percy Liang, and Christopher Manning. 2021. Conditional probing: measuring usable information beyond a baseline. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1626–1639.

Yongjie Huang and Meng Yang. 2021. Breadth first reasoning graph for multi-hop question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5810–5821.

Gautier Izacard and Édouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 874–880.

Yichen Jiang and Mohit Bansal. 2019. Self-assembling modular networks for interpretable multi-hop reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4474–4484.

Zhengping Jiang, Yining Lu, Hanjie Chen, Daniel Khashabi, Benjamin Van Durme, and Anqi Liu. 2024. RORA: Robust free-text rationale evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 1070–1087.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, pages 18661–18673.

Tushar Khot, Daniel Khashabi, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2021. Text modular networks: Learning to decompose tasks in the language of existing models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1264–1279.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas. Association for Computational Linguistics.

Ruosen Li and Xinya Du. 2023. Leveraging structured information for explainable multi-hop question answering and reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6779–6789.

Yen-Ting Lin, Alexandros Papangelis, Seokhwan Kim, Sungjin Lee, Devamanyu Hazarika, Mahdi Namazifar, Di Jin, Yang Liu, and Dilek Hakkani-Tur. 2023. Selective in-context data augmentation for intent detection using pointwise V-information. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1463–1476.

Wei Liu, Zhiying Deng, Zhongyu Niu, Jun Wang, Haozhao Wang, YuanKai Zhang, and Ruixuan Li. 2024. Is the mmi criterion necessary for interpretability? degenerating non-causal features to plain noise for self-rationalization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Wei Liu, Haozhao Wang, Jun Wang, Ruixuan Li, Xinyang Li, YuanKai Zhang, and Yang Qiu. 2023a. MGR: Multi-generator based rationalization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 12771–12787, Toronto, Canada. Association for Computational Linguistics.

Ye Liu, Semih Yavuz, Rui Meng, Dragomir Radev, Caiming Xiong, Shafiq Joty, and Yingbo Zhou. 2023b. Hpe: Answering complex questions over text by hybrid question parsing and execution. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4437–4451.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Jianguo Mao, Wenbin Jiang, Xiangdong Wang, Hong Liu, Yu Xia, Yajuan Lyu, and QiaoQiao She. 2022. Explainable question answering based on semantic graph by global differentiable learning and dynamic adaptive reasoning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5318–5325.

Vaibhav Mavi, Anubhav Jangra, Jatowt Adam, et al. 2024. Multi-hop question answering. *Foundations and Trends® in Information Retrieval*, (5):457–586.

Sewon Min, Victor Zhong, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2019. Multi-hop reading comprehension through question decomposition and rescoring. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6097–6109.

Kamal Nigam, John Lafferty, and Andrew McCallum. 1999. Using maximum entropy for text classification. In *IJCAI-99 workshop on machine learning for information filtering*, volume 1, pages 61–67. Stockholom, Sweden.

Ethan Perez, Patrick Lewis, Wen-tau Yih, Kyunghyun Cho, and Douwe Kiela. 2020. Unsupervised question decomposition for question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8864–8880.

Peng Qi, Haejun Lee, Tg Sido, and Christopher Manning. 2021. Answering open-domain questions of varying reasoning steps from text. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3599–3614.

Lin Qiu, Yunxuan Xiao, Yanru Qu, Hao Zhou, Lei Li, Weinan Zhang, and Yong Yu. 2019. Dynamically fused graph network for multi-hop reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6140–6150.

Gowtham Ramesh, Makesh Narsimhan Sreedhar, and Junjie Hu. 2023. Single sequence prediction over reasoning graphs for multi-hop qa. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 11466–11481.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

Zhengliang Shi, Shuo Zhang, Weiwei Sun, Shen Gao, Pengjie Ren, Zhumin Chen, and Zhaochun Ren. 2024. Generate-then-ground in retrieval-augmented generation for multi-hop question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 7339–7353.

Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, page 1857–1865.

Adam Storek, Melanie Subbiah, and Kathleen McKeown. 2023. Unsupervised selective rationalization with noise injection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12647–12659, Toronto, Canada. Association for Computational Linguistics.

Mokanarangan Thayaparan, Marco Valentino, Deborah Ferreira, Julia Rozanova, and André Freitas. 2022. Diff-explainer: Differentiable convex optimization for explainable multi-hop inference. *Transactions of the Association for Computational Linguistics*, 10:1103–1119.

Jiachen Tian, Shizhan Chen, Xiaowang Zhang, Zhiyong Feng, Deyi Xiong, Shaojuan Wu, and Chunliu Dou. 2021. Re-embedding difficult samples via mutual information constrained semantically oversampling for imbalanced text classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3148–3161.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, pages 539–554.

Ming Tu, Kevin Huang, Guangtao Wang, Jing Huang, Xiaodong He, and Bowen Zhou. 2020. Select, answer and explain: Interpretable multi-hop reading comprehension over multiple documents. In *AAAI 2020*.

Tomer Wolfson, Mor Geva, Ankit Gupta, Matt Gardner, Yoav Goldberg, Daniel Deutch, and Jonathan Berant. 2020. Break it down: A question understanding benchmark. *Transactions of the Association for Computational Linguistics*, 8:183–198.

Jian Wu, Linyi Yang, Yuliang Ji, Wenhao Huang, Börje F Karlsson, and Manabu Okumura. 2024. Gendec: A robust generative question-decomposition method for multi-hop reasoning. *arXiv preprint arXiv:2402.11166*.

Shicheng Xu, Liang Pang, Huawei Shen, Xueqi Cheng, and Tat-Seng Chua. 2024. Search-in-the-chain: Towards accurate, credible and traceable large language models for knowledge-intensive tasks. In *International World Wide Web Conference*.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. 2020. A theory of usable information under computational constraints. In *International Conference on Learning Representations*.

Zhichao Yan, Jiapu Wang, Jiaoyan Chen, Xiaoli Li, Ru Li, and Jeff Z Pan. 2024. Atomic fact decomposition helps attributed question answering. *arXiv preprint arXiv:2410.16708*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.

Jiahao Zhang, Haiyang Zhang, Dongmei Zhang, Liu Yong, and Shen Huang. 2024. End-to-end beam retrieval for multi-hop question answering. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1718–1731.

Wenting Zhao, Justin Chiu, Claire Cardie, and Alexander M Rush. 2023. Hop, union, generate: Explainable multi-hop reasoning without rationale supervision. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16119–16130.

Yunxiao Zhao, Zhiqiang Wang, Xiaoli Li, Jiye Liang, and Ru Li. 2024. AGR: Reinforced causal agent-guided self-explaining rationalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 510–518, Bangkok, Thailand. Association for Computational Linguistics.

Appendix

A Definitions

A.1 $\mathcal{V}$-information based Theory

A classification task can be viewed as minimizing the information entropy between prediction and truth (Ghoshal et al., 2022). Information entropy has been extensively applied to text-related tasks (Tian et al., 2021; Nigam et al., 1999). In this work, we introduce $\mathcal{V}$-information and conditional entropy to model informativeness under certain constraints. Specifically, the $\mathcal{V}$-information theory aims to model usable information under computational constraints (Xu et al., 2020; Ethayarajh et al., 2022; Lin et al., 2023; Jiang et al., 2024). It examines how much usable information about a random variable Y can be derived from another random variable X by applying functions belonging to a specified set.

Definition 1 ($\mathcal{V}$-information) *Let X, Y denote random variables with sample spaces $\mathcal{X}$, $\mathcal{Y}$, respectively, $\varnothing$ denotes a null input that provides no information about Y, Ω denotes a specified functions set, the $\mathcal{V}$-information from X to Y is*

$$I_{\mathcal{V}}(X \to Y) = H_{\mathcal{V}}(Y) - H_{\mathcal{V}}(Y|X),$$

where $H_{\mathcal{V}}(Y|X)$ is conditional $\mathcal{V}$-entropy.

Definition 2 (Conditional $\mathcal{V}$-entropy) *The conditional $\mathcal{V}$-entropy can be defined as follows:*

$$H_{\mathcal{V}}(Y|X) = \inf_{f \in \mathcal{V}} \mathbb{E}[-\log_2 f[X](Y)].$$

Specifically, $H_{\mathcal{V}}(Y)$ is defined for $X = \varnothing$.

A.2 Tuplet Objective Optimization

The triplet loss aims to learn an embedding representation of the data that preserves the distance between similar data points close and dissimilar data points far on the embedding spaces (Schroff et al., 2015; Sohn, 2016; Khosla et al., 2020). (Sohn, 2016) proposes an (N+1)-tuplet loss, which extends triplet loss by allowing joint comparison among more than one negative example.

Definition 3 ((N+1)-Tuplet Loss) *Given a training example $\{x, x^+, x_1, \cdots, x_{N-1}\}$, where x is an anchor example, x^+ is a positive example to x and $\{x_i\}_{i=1}^{N-1}$ are negative. The (N+1)-tuplet loss is defined as follows:*

$$\mathcal{L}(\{x, x^+, \{x_i\}_{i=1}^{N-1}\}; f(\cdot;\theta))$$

where $f(\cdot;\theta)$ is an embedding kernel defined by deep neural network .

B Experimental Details

B.1 Baselines

We compare our approach with nine baselines:

RNN (2018), a strong RNN-based baseline, which is a non-transformer model.

FiD (2021), an end-to-end model, which takes question and context as input, and generate the answer as a sequence of tokens.

FiD$_{\text{+PT}}$ (2021) is FiD pre-trained on the reader network using a subset of probably asked questions.

EE (2022), an end-to-end model, which takes question and context as input, and runs it through a transformer.

SA (2022), a retrieval-based model that includes a selector and a reader.

EX(EE) (2022), a multistep reasoning model that decomposes question into single-hop questions, using an End2End model for iterative execution.

EX(SA) (2022), a multistep reasoning model that decomposes question into single-hop questions, using a selector and a reader for iterative execution.

HPE (2023b), a multistep reasoning model that parses question into H-expressions, followed by hybrid execution to get the final answer.

HUG (2023), a retrieval-based model that explicitly considers all possible document sets and sentence subsets to generate an answer.

B.2 Implementation

Following previous work, we re-implement the backbone model SA and further develop our $\mathcal{LOG}$ model based on it. Specifically, we implement a retriever using RoBERTa-large (Liu, 2019) and a reader using Longformer-Large (Beltagy et al., 2020) to fairly compare. We use AdamW (Loshchilov and Hutter, 2019) as the optimizer with a learning rate of 2e-5 and batch size of 16. We set the maximum length of the input sequence to 300. The retriever chooses the top-K documents, where K is 7. Our all experiments are run on NVIDIA Tesla V100 GPUs with 32GB.

C More Experiments

C.1 Performance on Dataset HotpotQA

We compare our proposed $\mathcal{LOG}$ on the distractor setting of HotpotQA, which has 90,564 train, 7,405 development, and 7,405 test examples, respectively. It includes 2-hop questions, answers, and a collection of 10 documents as context per question. We compare two end-to-end models:

SNMN (Jiang and Bansal, 2019) and EE (Trivedi et al., 2022); four decomposition-based models: DecompRC (Min et al., 2019), BreakRC (Wolfson et al., 2020), ModularQA (Khot et al., 2021) and Mao et al. (Mao et al., 2022); two retrieval-based models: SA (Trivedi et al., 2022) and HUG (Zhao et al., 2023); and an early method: RNN-based baseline (Yang et al., 2018).

As shown in Table 4, this experimental setup focuses on the dataset with only 2-hop multi-hop reasoning. We find that our proposed $\mathcal{LOG}$ outperforms the all baselines, demonstrating that our method also maintains competitive performance on questions with shorter hops.

Methods	HotpotQA-Distractor	
	Answer_EM	Answer_F1
RNN (Yang et al., 2018)	-	51.0
DecompRC (Min et al., 2019)	-	61.7
SNMN (Jiang and Bansal, 2019)	-	63.1
BreakRC (Wolfson et al., 2020)	39.2	51.4
ModularQA (Khot et al., 2021)	-	61.8
Mao et al. (Mao et al., 2022)	55.4	69.1
EE (Trivedi et al., 2022)	-	72.9
SA* (Trivedi et al., 2022)	63.6	76.9
HUG (Zhao et al., 2023)	-	73.5
$\mathcal{LOG}$	**63.8**	**77.5**

Table 4: QA performance on the development set of HotpotQA-Distractor in comparison with previous work. * indicates the backbone model of our proposed $\mathcal{LOG}$.

C.2 Performance on Different-architecture Reader

As presented in Table 5, we integrate $\mathcal{LOG}$ with three distinct reader architectures (RoBERTa (Liu, 2019), DeBERTa (He et al., 2021), and Longformer (Beltagy et al., 2020)), resulting in notable answer_F1 score improvements: 3.4 points (from 47.9 to 51.3) with RoBERTa, 4.0 points (from 51.2 to 55.2) with DeBERTa, and 3.4 points (from 47.3 to 50.7) with Longformer on the MuSiQue-Ans dataset. The results verify the effectiveness and adaptability of $\mathcal{LOG}$.

Methods	MuSiQue-Ans	
	Answer_EM	Answer_F1
SA_RoBERTa	40.2	47.9
SA_DeBERTa	41.2	51.2
SA_Longformer	39.3	47.3
$\mathcal{LOG}$_RoBERTa	43.4	51.3
$\mathcal{LOG}$_DeBERTa	46.6	55.2
$\mathcal{LOG}$_Longformer	42.7	50.7

Table 5: Different-architecture reader results on the development set of MuSiQue-Ans.

D Case Studies

As depicted in Figure 5, we also provide more detailed case studies, involving 2-hop, 3-hop, and 4-hop scenarios, to illustrate the reasoning processes of both our baseline model (SA) and $\mathcal{LOG}$. For instance, in the fifth example (3 hops), the SA model correctly answers the first query with *Sazerac Company*, but makes errors on the subsequent queries, providing *St. Louis* for the second and *Billiken* for the third. In contrast, our proposed model $\mathcal{LOG}$ consistently delivers accurate answers: *Sazerac Company* for the first query, *New Orleans* for the second, and *Fleur-de-lis* for the third.

Similarly, in the seventh example (4 hops), the SA model starts correctly by answering the first query as *Guadalajara*, but errors accumulate in the subsequent steps, leading to incorrect answers: *The Atlantic Ocean*, *King João I*, and *Prince Henry the Navigator*. In contrast, our proposed model $\mathcal{LOG}$ maintains accuracy across all four queries, answering with *Guadalajara*, *North America*, *John Cabot*, and *Sebastian Cabot*. These results further highlight that SA is susceptible to an error accumulation effect, whereas $\mathcal{LOG}$ excels in capturing more detailed and accurate information.

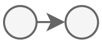
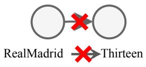
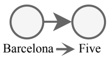
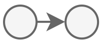
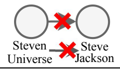
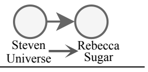
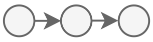
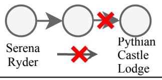
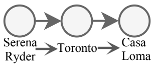
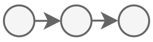
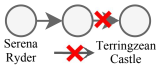
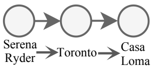
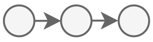
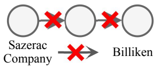
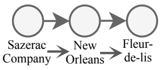
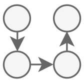
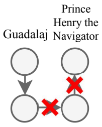
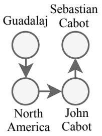
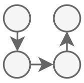
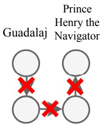
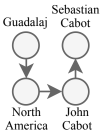
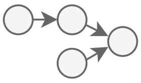
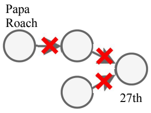
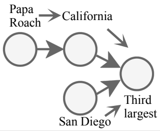
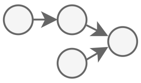
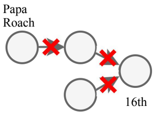
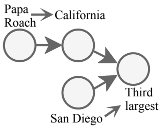
Graph	Question	Decompositon	SA Model	LOG Model
	How many UEFA Super Cup awards have been received by the team that has won the treble competitions twice? Five	**1.** What team has won the treble competitions twice ? Barcelona **2.** How many UEFA Super Cup awards does Barcelona have? Five		
	Who is the creator of the main character in the Steven the Sword Fighter series? Rebecca Sugar	**1.** Steven the Sword Fighter >> part of the series? Steven Universe **2.** Steven Universe >> creator ? Rebecca Sugar		
	What is the name of the castle in the city where the performer of Fall was born? Casa Loma	**1.** Fall >> performer ? Serena Ryder **2.** Serena Ryder >> place of birth? Toronto **3.** What is the name of the castle in Toronto? Casa Loma		
	What is the name of the castle found in the birthplace of the performer of Fallin g Out? Casa Loma	**1.** Falling Out >> performer ? Serena Ryder **2.** Serena Ryder >> place of birth ? Toronto **3.** what is the name of the castle in Toronto ? Casa Loma		
	What is the symbol of the Saints from the city where the headquarters of the manufacturer of McAfee's Benchmark called ? Fleur-de-lis	**1.** McAfee's Benchmark >> manufacturer ? Sazerac Company **2.** Sazerac Company >> headquarters location ? New Orleans **3.** what is the New Orleans saints symbol called ? Fleur-de-lis		
	Who was the child of the Italian navigator who sailed for England, and explored the eastern coast of the continent where the birthplace of Eduardo Cuervo is found? Sebastian Cabot	**1.** Eduardo Cuervo >> place of birth ? Guadalajara **2.** Guadalajara >> continent ? North America **3.**Who was the italian navigator sailing for englan d that explored the eastern coast of North America ? John Cabot **4.** John Cabot >> child? Sebastian Cabot		
	Who is the son of the Italian navigator who explored the eastern coast of the continent Ulises Solís' birthplace is located in for England? Sebastian Cabot	**1.** Ulises Solís >> place of birth ? Guadalajara **2.** Guadalajara >> continent ? North America **3.**Who was the italian navigator sailing for englan d that explored the eastern coast of North America ? John Cabot **4.** John Cabot >> child ? Sebastian Cabot		
	Among the top five largest urban areas in the state Jacoby Shaddix' band formed in, where does the city Veoh's headquarters are located rank? Third-largest	**1.** Jacoby Shaddix >> member of ? Papa Roach **2.** What city was Papa Roach formed in? California **3.** Veoh >> headquarters location? San Diego **4.**In the top five largest urban areas in California , where does San Diego rank? Third-largest		
	Where does the city where Veoh's headquarters is located rank in the top five largest urban areas of the state where The Paramour Sessions' performers were formed? Third-largest	**1.** The Paramour Sessions >> performer? Papa Roach **2.** What city was Papa Roach formed in? California **3.** Veoh >> headquarters location ? San Diego **4.** In the top five largest urban areas in California , where does San Diego rank? Third-largest		

Figure 5: Some visual case studies. We cover 2-hop, 3-hop, and 4-hop reasoning scenarios.