Explainable Multi-Hop Question Answering: A Rationale-Based Approach

Kyubeen Han †, Youngjin Jang †, And Harksoo Kim *

Posted Date: 24 September 2025

doi: 10.20944/preprints202509.1957.v1

Keywords: unsupervised learning; machine reading comprehension; multi-hop question answering; explainable AI; rationale inference

 doi:10.20944/preprints202509.1957.v1

Article

Explainable Multi-Hop Question Answering: A Rationale-Based Approach

Kyubeen Han [1,†], Youngjin Jang [1,†] and Harksoo Kim [2,*]

[1] Department of Artificial Intelligence, Konkuk University, 120 Neungdong-ro, Gwangjin-gu, Seoul 05029, Republic of Korea
[2] Division of Computer Science and Engineering & Department of Artificial Intelligence, Konkuk University, 120 Neungdong-ro, Gwangjin-gu, Seoul 05029, Republic of Korea
* Correspondence: nlpdrkim@konkuk.ac.kr
† These authors contributed equally to this work.

Abstract

Multi-hop question answering tasks involve identifying relevant supporting sentences from a given set of documents, which serve as the rationale for deriving answers. Most research in this area consists of two main components: a rationale identification module and a reader module. Since the rationale identification module often relies on retrieval models or supervised learning, annotated rationales are typically essential. However, approaches that rely on annotations face challenges when adapting to open-domain settings. Moreover, when models are trained on annotated rationales, explainable artificial intelligence (XAI) requires clear explanations of how these rationales are derived. Therefore, traditional multi-hop question answering approaches that depend on annotated rationales are unsuitable for XAI, which demands transparency in the model's reasoning process. To address this issue, we propose a rationale reasoning framework that can effectively infer rationales and demonstrate the model's reasoning process, even in open-domain environments without annotations. The proposed model is applicable to various tasks without structural constraints, and experimental results demonstrate its significantly improved rationale reasoning capabilities in multi-hop question answering, relation extraction, and sentence classification tasks.

Keywords: unsupervised learning; machine reading comprehension; multi-hop question answering; explainable AI; rationale inference

1. Introduction

Recent advancements in artificial intelligence (AI) have led to significant improvements across various domains. In particular, large language models (LLMs) [1–4] have transformed the field of natural language processing (NLP). By learning from vast amounts of data, LLMs understand linguistic context, meaning, and structure, enabling accurate generation and comprehension of natural language. Despite their remarkable capabilities, LLMs encounter critical challenges related to the opacity and reliability of their reasoning processes [5–7]. For instance, even when users prompt LLMs to provide a rationale for a response, they often remain unsure how that rationale is generated [8]. This lack of transparency raises concerns about why the model produces a response and what rationale underlies its decisions. As a result, users may struggle to trust the model's responses, leading to potential risks regarding the accuracy and reliability of the information provided. This lack of reliability can reduce users' trust in the model, potentially leading them to seek alternative sources of information. To address these concerns, research in explainable artificial intelligence (XAI) aims to enhance the transparency of AI systems by making their reasoning processes more understandable, thereby fostering greater trust in their outputs [9–11]. Therefore, XAI must provide clear explanations of how AI systems reach their responses, offering intuitive insights that facilitate user trust and understanding.

In this paper, we define the criteria for XAI models as follows: (1) The rationales provided by the model must be reliable and understandable to users. (2) Users should be able to comprehend

 doi:10.20944/preprints202509.1957.v1

the model's operation through its reasoning process. (3) The model should be capable of providing rationales without relying on annotations attached to the data. (4) The rationales provided by the model should align with its predictive outputs. To meet these criteria, we propose a framework that satisfies these requirements and apply it to the multi-hop question answering task, a representative task related to XAI.

Multi-hop question-answering tasks involve identifying multiple supporting sentences from a given set of documents, which serve as evidence for inferring an answer. These supporting sentences serve as rationale for deriving accurate responses. Since this task is a complex process that requires understanding and utilizing the relationships between sentences across multiple documents, it is crucial to identify reliable rationales. Most multi-hop question answering research [11,12] consists of two main components: a supporting sentence identification module and a reading comprehension module. The supporting sentence identification module identifies the sentences necessary for inferring the answer. It identifies supporting sentences within the documents based on retrieval models or supervised learning [13,14]. The reader then uses the sentences selected by the supporting sentence identification module to infer responses to the given questions. However, these approaches have several limitations. Firstly, the supporting sentence identification module requires explicit annotations for rationale. Even with these annotations, clearly explaining how the rationale is derived remains challenging. Secondly, the reader module often considers the entire input sequence composed of the retrieved segment as the rationale. This makes it difficult to clearly explain the reader model's reasoning process, which may not meet the requirements of XAI. Consequently, traditional multi-hop question-answering approaches are not well-suited to open-domain environments where annotated rationales are difficult to obtain or where model transparency is crucial.

To address these issues, this paper proposes a framework for explainable multi-hop question answering that makes the model's reasoning process explicit and enables rationale inference without annotations. The proposed model is designed based on a pointer network and is trained using a training strategy that guides it to infer rationale sentences. Additionally, the model ensures factuality by extracting rationales from documents and presenting them at the sentence level to enhance user understanding. Furthermore, the process of selecting rationale sentences using the pointer network provides insights into the model's reasoning process, and demonstrates strong performance without the need for explicit annotations. This approach meets the requirements of XAI and achieves strong rationale inference performance in experiments.

2. Related Works

XAI aims to explain a model's decision in ways that users can understand [15–17]. Most XAI research [18–21] focuses on interpreting the model's inference process or identifying key words, phrases, or sentences that influence predictions, rather than relying directly on annotated rationale labels. For example, [22] proposed aligning textual elements in natural language inference to analyze sentence correspondence, while using masking techniques to provide token- or phrase-level rationale. Similarly, [10] used trainable masks to block or activate specific nodes or edges in a graph, thereby removing irrelevant information and extracting key rationales. However, these studies have limited scope and applicability.

Multi-hop Question Answering requires combining multiple sentences from a given document set to answer complex questions. This task requires the clear identification of supporting sentences necessary for deriving the answer. [23] attempted to explain the rationale by classifying whether each sentence qualifies as supporting evidence. However, their approach does not sufficiently address interactions between sentences and has limitations in explaining complex, multi-document questions. [13] demonstrated the potential of utilizing annotation-based rationales through supervised learning. However, annotation-dependent methods face challenges in open-domain environments. To address this, [24] and [25] introduced complex datasets without annotations and emphasized the need for effective unsupervised learning methods. [26] and [23] proposed methods for finding rationale at the

 doi:10.20944/preprints202509.1957.v1

sentence level but mainly focused on dependencies between two pieces of information. In contrast, models like HUG [11] sought to address limitations by performing multi-hop inference through document and sentence predictions, though the explainability of these models' operations remains insufficiently explored. Existing XAI approaches still face challenges in explainability and transparency. For instance, RAG [14] treats evidence as latent variables but overlooks inter-document connections. [9] successfully generated pseudo-evidence annotations through semi-supervised learning for specific QA tasks, but this approach is limited to certain types of multi-hop questions. To overcome these limitations, ongoing research aims to develop explainable multi-hop QA methods [11,27,28]. However, a general approach [29,30] that clearly explains supporting sentences and is applicable across various domains is still needed. Therefore, this paper proposes a rationale reasoning framework that effectively infers the rationale without annotations in open-domain environments and provides explanations of the model's reasoning process. The proposed model is applicable to various tasks without structural constraints and has demonstrated superior performance on multi-hop question answering, relation extraction, and sentence classification tasks.

3. Methodology

Multi-hop question answering identifies the supporting sentences from a document set $D = \{d_0, d_1, ..., d_{|D|-1}\}$ and infers the answer to question Q. Here, $|D|$ refers to the size of the document set. Figure 1 illustrates the structure of the proposed model. The proposed model comprises a pre-trained language model [31] and a pointer network [32]. The pre-trained language model encodes the documents, while the pointer network selects the information necessary for answer inference. The final hidden states of the pointer network, which accumulate rationale information, are used for the final answer.

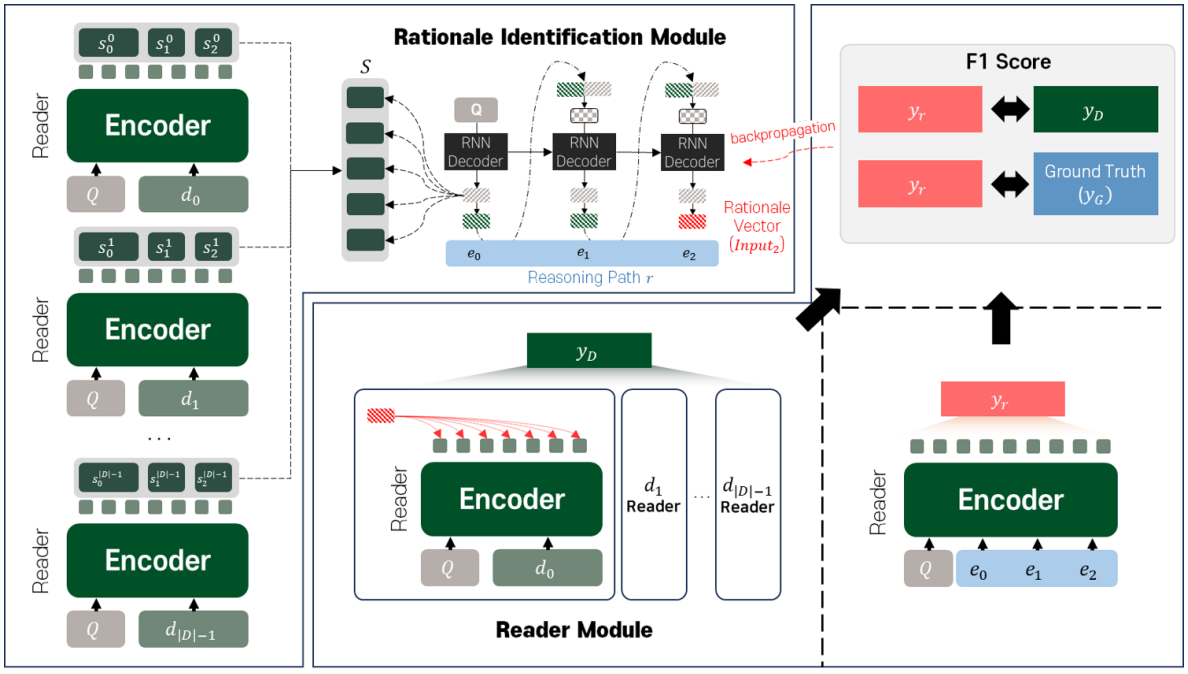

Figure 1. An overview of our proposed framework. The model architecture consists of two components: a rationale identification module based on a pointer network and a reader module based on a pre-trained language model.

Input Encoder Specifically, the input sequence is defined as follows:

$$T = \{\{t_0^0, ..., t_{L-1}^0\}, ..., \{t_0^{|D|-1}, ..., t_{L-1}^{|D|-1}\}\} \in \mathbb{R}^{|D| \times L} \quad (1)$$

where T consists of a [CLS] token, a tokenized question, an [SEP] token, a tokenized document, and a final [SEP] token. Here, L refers to the maximum number of tokens in the input sequence. The sequence

 doi:10.20944/preprints202509.1957.v1

T is then processed by an encoder to obtain a token vector sequence V. Subsequently, sentence vectors $S = \{s_0^0, s_1^0, ..., s_{|d|-1}^{|D|-1}\} \in \mathbb{R}^{|S| \times h}$ are generated using the encoded token vector sequence V using mean pooling. Here, $|d|$ refers to the maximum number of sentences constituting each document, $|S|$ refers to the total number of sentences in the document set D, and h denotes the output vector size of the pre-trained model.

Pointer Network The pointer network selects sentences through attention mechanisms with S and is guided to choose supporting sentences using the proposed training strategy. The initial state of the pointer network is set to *None*, and the initial input, $input_0$, is as shown in Equation 2 below:

$$input_0 = \frac{1}{|D|} \sum_{d=0}^{|D|-1} s_0^d \tag{2}$$

The output of the pointer network, h_n, is as shown in Equation 3.

$$h_n = GRU(input_n, h_{n-1}) \tag{3}$$

Here, n refers to the number of decoding steps, and $input_n$ and h_{n-1} represent the input to the RNN [33] and the previous RNN output, respectively. The context vector C_n is generated by incorporating important sentence information through an attention operation between the output h_n produced by the RNN and the sentence representations S, as shown in Equation 4 below:

$$\begin{aligned} score(S, h_n) &= h_n W_a S^T \\ \alpha_n &= softmax(score(S, h_n)) \\ C_n &= \sum_k^{|D|} \alpha_{n,k} S_k \end{aligned} \tag{4}$$

In this equation, W_a represents a trainable weight matrix, and $score(S, h_n)$ denotes the importance of each sentence, calculated as the inner product between the sentence representations S and the RNN output h_n at the current decoding step. α_n refers to the $score(S, h_n)$ normalized via the softmax function, and C_n represents the weighted sum vector of S with respect to α_n. The context vector C_n is concatenated with the RNN output h_n to form the next input vector $input_{n+1}$, as described in Equation 5 below:

$$input_{n+1} = W_{in}(C_n \oplus h_n) \tag{5}$$

In the equation above, W_{in} represents a trainable weight matrix. As shown in Figure 1, we consider the final input, $input_N$, accumulated through the pointer network, as the rationale vector. This is used to generate the probability distribution ($\hat{y}_{start}$, $\hat{y}_{end}$) for the answer positions, and the corresponding equation is given by Equation 6.

$$\begin{aligned} \hat{y}_{start} &= VW_s(input_N)^T \in \mathbb{R}^{|D| \times L} \\ \hat{y}_{end} &= VW_e(input_N)^T \in \mathbb{R}^{|D| \times L} \end{aligned} \tag{6}$$

In the above equation, the probability distributions $\hat{y}_{start}$ and $\hat{y}_{end}$ are obtained by multiplying the token representations V with the trainable weight matrices W_s and W_e, respectively, followed by multiplication with the rationale vector $input_N$.

 doi:10.20944/preprints202509.1957.v1

Training In this paper, we train the proposed model using two loss functions: one for answer inference and another for guiding the pointer network in selecting the supporting sentences. First, the model is trained by minimizing the cross-entropy between the predicted probability distributions, $\hat{y}_{start}$ and $\hat{y}_{end}$, and the ground-truth answer positions y_{start} and y_{end}, defined as follows:

$$\begin{aligned} L_s &= -\sum y_{start} \log(\hat{y}_{start}) \\ L_e &= -\sum y_{end} \log(\hat{y}_{end}) \\ L_{span} &= (L_s + L_e)/2 \end{aligned} \tag{7}$$

The proposed framework is trained to minimize the negative log-likelihood (NLL) described above, ensuring effective task performance. Specifically, during the training, the model adjusts its predicted label distribution to closely align with the correct label distribution by incorporating the information accumulated through the sentence selections of the pointer network. This mechanism enhances the model's ability to generate reliable and well-informed predictions. Furthermore, to guide the pointer network in accurately selecting rationale sentences, an auxiliary loss function is introduced. This auxiliary loss function plays a crucial role in refining the sentence selection process, encouraging the model to make more precise decisions when identifying important rationale sentences. As a result, the selected sentences contribute meaningfully to the model's final predictions, ultimately improving both interpretability and reliability.

The corresponding equation is presented in Equation 8 below:

$$\begin{aligned} L_D(r) &= \sum_{n=0}^{N-1} [\alpha_d \sum_{s \in r} \log(P(s|e_n)) + (1-\alpha_d) \sum_{s \in S-r} \log(P(s|e_n))] \\ L_G(r) &= \sum_{n=0}^{N-1} [\alpha_g \sum_{s \in r} \log(P(s|e_n)) + (1-\alpha_g) \sum_{s \in S-r} \log(P(s|e_n))] \end{aligned} \tag{8}$$

$$\begin{aligned} \alpha_d &= \begin{cases} 1 & \text{if } F_1(y_r, y_D) > Threshold \\ 0 & \text{otherwise} \end{cases} \\ \alpha_g &= \begin{cases} 1 & \text{if } F_1(y_r, y_G) > Threshold \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{9}$$

In this equation, $P(s|e_n)$ represents the probability that the pointer network selects sentence s at the n-th decoding step. y_D is the predicted answer based on the entire input documents D, while y_r is the predicted answer from the reasoning path r selected by the pointer network. In other words, if an inference similar to the full context can be achieved with only the sentences selected by the pointer network, the probability $P(s|e_n)$ of the selected sentences is increased; otherwise, probabilities of unselected sentences increase. Similarly, L_G is calculated by comparing the annotated ground truth y_G with y_r. The final loss function L_{total} is defined as shown in Equation 10.

$$L_{total} = L_{span} + \lambda \cdot (L_D(r) + L_G(r)) \tag{10}$$

In the above formula, λ represents a weighting coefficient.

Beam Search Decoding This section explains the process of extending the pointer network, as shown in Figure 2. Through beam search decoding [34], the model generates K reasoning paths $R = \{r^1, r^2, ..., r^K\}$, evaluates each path, selects the optimal reasoning path r^{k*} for answer inference, and uses it during training. Each reasoning path r^k is evaluated by comparing the predicted answer

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 September 2025 doi:10.20944/preprints202509.1957.v1

y_D based on the entire document, the predicted answer y_{r^k} based on r^k, and the ground-truth answer y_G, using the sum of F_1 scores. This process is illustrated in Figure 3.

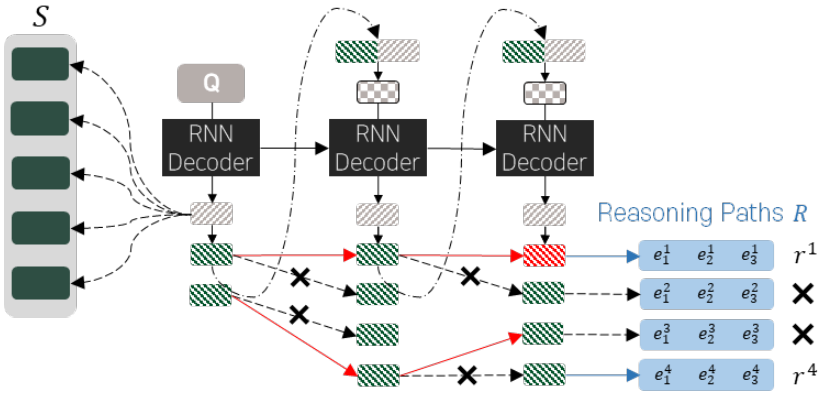

Figure 2. Expansion of reasoning paths R via beam search decoding.

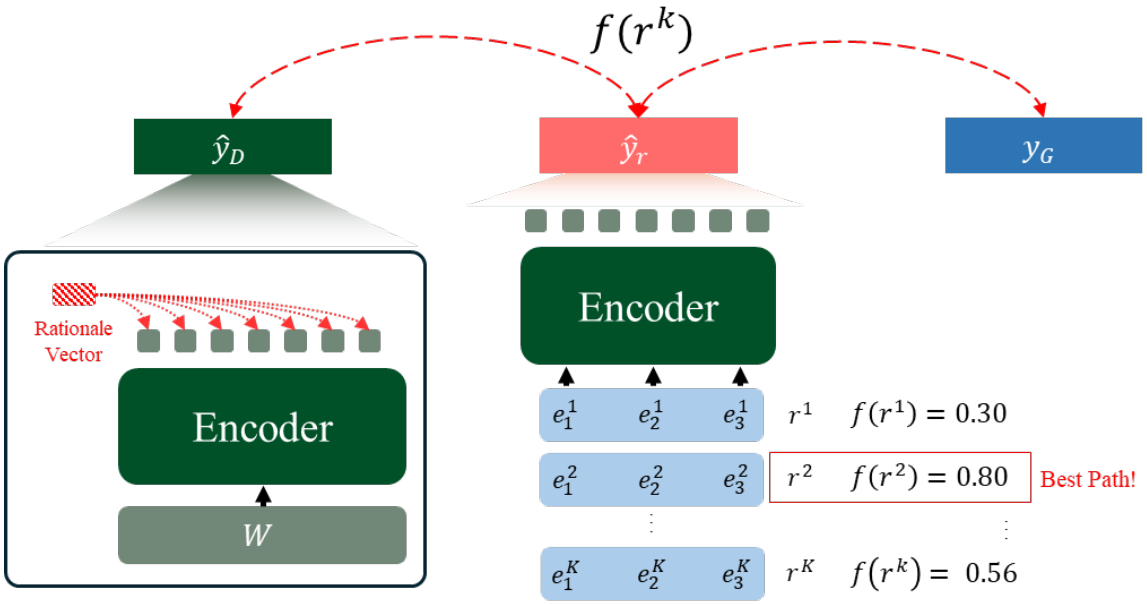

Figure 3. Selection of the best reasoning path r^{k*} based on scoring function f.

We select the best path, r^{k*}, which maximizes the F_1 score among the reasoning paths and use it to train the model using the following equation.

$$\begin{aligned} f(r^k) &= F_1(y_{r^k}, y_D) + F_1(y_{r^k}, y_G) \\ k^* &= \arg\max_K f(r^k) \\ L &= L_{span} + \lambda \cdot (L_D(r^{k^*}) + L_G(r^{k^*}) \end{aligned} \tag{11}$$

4. Experiments

4.1. Dataset and Experimental Setup

The primary task addressed in this study is multi-hop question answering. To assess the effectiveness of the proposed approach, we conducted experiments using two multi-hop question answering datasets [12,35]. Additionally, to emphasize that the proposed model is task-agnostic and capable of inferring rationales for its predictions across various tasks, we conducted additional experiments on a

 doi:10.20944/preprints202509.1957.v1

document-level relation extraction dataset [36] and a binary sentiment analysis dataset [37]. Through these experiments, we aimed to demonstrate the model's consistent performance across diverse tasks and confirm that its rationale extraction capability is not limited to a specific task.

The proposed model utilizes the pre-trained ELECTRA-base [31] as its encoder. The weighting coefficient λ in Equations 10, and 11 was set to 0.1. Furthermore, the pointer network was constrained to extracting two or three sentences, depending on the dataset used in the experiments.

HotpotQA This dataset [12] is a large-scale multi-hop question answering dataset designed to integrate information across documents to answer questions. It contains approximately 112,000 questions and corresponding answers covering a wide range of topics. For each question, 10 related documents are provided, containing supporting sentences that the model must reference to infer the correct answer. In our experiments, we used 90,564 samples for training and 7,405 samples for development under the distractor setting.

MuSiQue This dataset [35] is designed to train and evaluate question answering tasks requiring multi-hop reasoning over text, similar to HotpotQA. It comprises approximately 25,000 questions with corresponding answers, where each question consists of several sub-questions. This structure enables the evaluation of the model's ability to integrate information from various sources to derive answers. The proposed model is trained to infer answers to the initially provided questions without directly addressing the sub-questions. The dataset consists of 19,144 samples and 2,417 development samples.

DocRed This dataset [36] is a large-scale resource for document-level relation extraction, focusing on identifying relationships between entities across entire documents rather than within individual sentences. It contains 132,375 entities and 56,354 relationships extracted from 5,053 Wikipedia documents, encompassing 96 distinct relation types. Notably, each relation instance is accompanied by supporting sentences for relation extraction. The objective task includes identifying pairs of entities that have existing relations among all possible combinations of entities. However, in our experiments, relation extraction was performed only on entity pairs with existing relations, and documents with fewer than three sentences were excluded. The final dataset used for training comprises 56,195 samples, while the evaluation dataset includes 17,803 samples.

IMDB This dataset [37] is designed for sentiment analysis, aiming to predict the sentiment (positive/negative) by analyzing movie review texts. It includes a total of 50,000 movie reviews, with each review labeled as either positive or negative. The dataset is split into 25,000 training samples and 25,000 testing samples, with each set containing an equal number of positive and negative reviews.

Detailed statistics of the datasets used in the experiments are presented in Table 1.

Table 1. Statistics of the experimental datasets.

Dataset	Train set	Test set	Sampled Test set
HotpotQA	90,564	7,405	1000
MuSiQue	25,494	3,911	1000
DocRED	56,195	17,803	1000
IMDB	25,000	-	1000

As shown in Table 1, we used four publicly available datasets for our experiments, utilizing the train and test sets for quantitative evaluation of both answer prediction and rationale inference. However, since the IMDB dataset does not include rationale annotations, it cannot be directly used for quantitative evaluation of rationale inference. Therefore, we sampled a subset of the test set (Sampled Test set) and evaluated rationale inference using a GPT-based model.

4.2. Baseline

The proposed framework is designed to be adaptable across various tasks. To assess its generalizability, we adjusted the output layer according to each task. In this process, unlike in the multi-hop question answering task, the rationale vector $input_N$ is concatenated with the $[CLS]$ vector to generate

 doi:10.20944/preprints202509.1957.v1

the final probability distribution $\hat{y}$. The detailed structure of this process is illustrated in Figure 4 below.

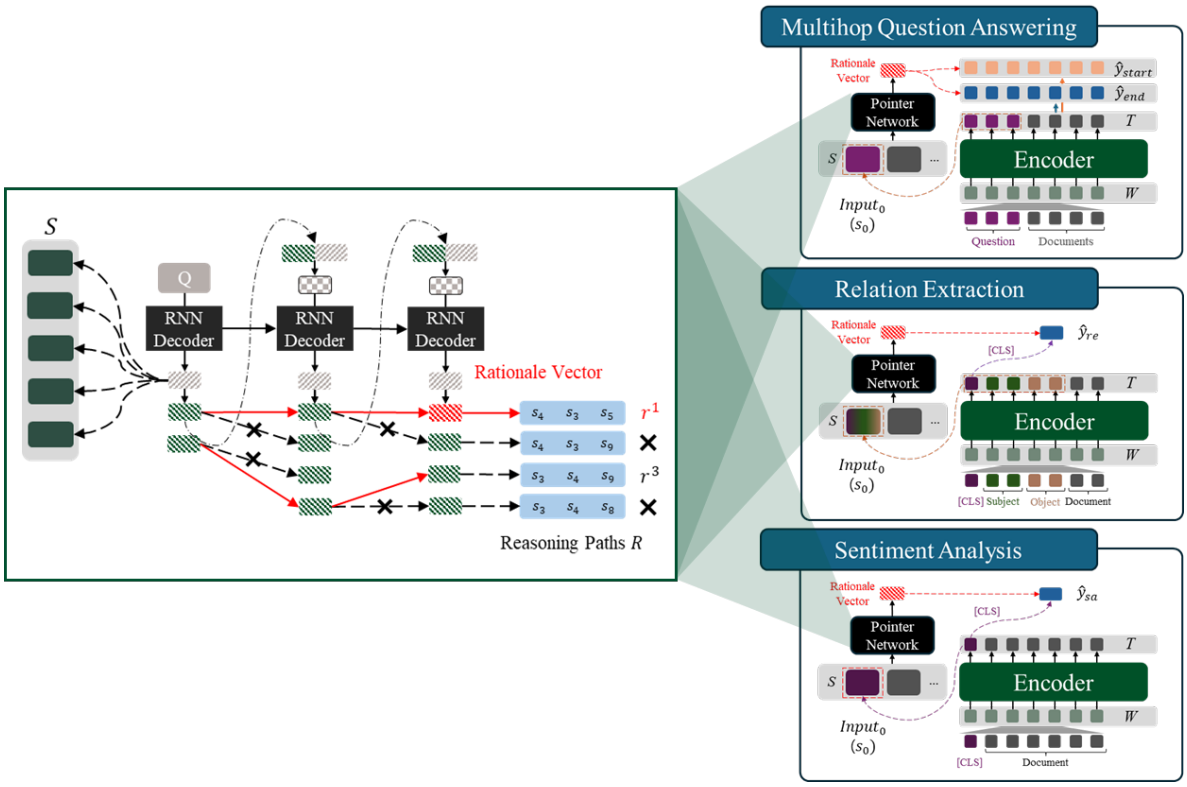

Figure 4. Task Specific Prediction.

At this stage, for the classification task, the best reasoning path r^{k*} is determined using KL Divergence Loss [38] instead of the F_1 score. Specifically, the selected sentence subset r^{k*} is the one that yields a prediction most similar to the prediction obtained by using the entire input sequence. Detailed implementation for each task is provided in Appendix A.

4.3. Metrics

As a performance metric, we set the F_1 score for the model's predictions. For HotpotQA, MuSiQue and DocRED, which contain annotated rationale, we compute not only the F_1 score of the answers but also the precision, recall, and F_1 score of the rationale. The F_1 score of the rationale is calculated at the sentence level. Furthermore, we define a model as having high rationale inference capability if it provides appropriate supporting rationale, even when its predictions are incorrect. Therefore, we determined that comparing the model's rationale inference capabilities based only on annotated rationales is insufficient, so we utilize GPT-4o mini for this evaluation. Specifically, we instruct GPT-4o mini to evaluate the supporting rationale for predictions by providing the task description, the model's predicted answers, and the inferred rationale sentences. We conduct the evaluation on a scale from 0 to 2, depending on how well the rationale sentences support the predictions. The evaluation criteria are set as follows:

- 0 points: The provided rationale does not substantiate or is irrelevant to the model's prediction at all.
- 1 point: The provided rationale somewhat supports the model's prediction but is not decisive; it is partially inferred but unclear.
- 2 points: The provided rationale fully substantiates the model's prediction, and the same answer can be inferred solely based on the given rationale.

 doi:10.20944/preprints202509.1957.v1

The GPT-4o mini evaluation was performed on a dataset (Sampled Dataset) consisting of 1,000 randomly selected samples from each evaluation dataset. The evaluation scores were normalized to a range of 0 to 1. The prompts used in the experiments are detailed in Appendix B.

4.3.1. Prompt Details for GPT-Based Evaluation

In this section, we provide a detailed explanation of the prompts used to guide GPT-4o mini in evaluating the model's rationale inference capabilities. The prompts were structured to present GPT-4o mini with three key pieces of information: the task description, the model's predicted answer, and the rationale sentences inferred by the model. The goal of the evaluation was to measure how effectively the inferred rationale sentences supported the model's predictions. Each sample provided GPT-4o mini with the model's predicted answer and the corresponding rationale, and the model was instructed to score the relevance of the rationale on a scale from 0 to 2, according to predefined criteria. The following elements were included in each prompt:

- Task Description: A clear explanation of the evaluation task.
- Predicted Answer: The model's predicted answer for the given question.
- Rationale Sentences: The evidence sentences inferred by the model, to be evaluated by GPT-4o mini.

These elements formed the core of each evaluation prompt.

4.4. Comparison Models

To evaluate the rationale inference performance of our proposed model against existing approaches, we selected the following comparison models. Models trained on annotated evidence from the dataset were excluded from this evaluation.

HotpotQA and MuSiQue: BM25 [39] and RAG [14] serve as baselines retrieval models that were not trained on the experimental dataset. RAG was modified to search at the sentence level to match the characteristics of the dataset, providing three fixed sentences for performance evaluation. The semi-supervised approach [9] involves constructing silver rationale annotations in advance and learning through an RNN decoder to infer the rationale. HUG [11] is a state-of-the-art unsupervised model for multi-hop question answering. It identifies the optimal rationale by combining sentences from all documents and predicts the answer based on this rationale. [40] applies a multi-stage retrieval method at both the document and sentence levels, based on BM25. Their approach demonstrated superior performance compared to existing retrieval methods and showed strong performance in answering complex questions such as those in multi-hop question answering. The upper bound represents the performance of models trained on annotated rationales, similar to the structure of the proposed model. It provides an upper bound on the rationale inference performance that the proposed model can potentially achieve.

DocRED and IMDB: GPT_{pred} illustrates the performance of GPT-4o mini [41] in a zero-shot setting. The prompts used in the experiments are detailed in Table A5 and Table A6. For DocRED, GPT_{pred} refers to the performance GPT-4o mini when instructed to perform both relation extraction and evidence inference simultaneously. For the IMDB dataset, GPT_{pred} is prompted to perform sentiment analysis while simultaneously identifying key sentences that influence the classification.

This evaluation assesses the model's ability to infer rationales in an unsupervised manner without access to annotated evidence, demonstrating its effectiveness in extracting supporting rationales directly from the text. In this experiment, GPT is given ground-truth labels and supporting sentences from the dataset, independent of GPT_{pred}'s inference results, to assess how effectively the provided rationale supports the correct answers.

 doi:10.20944/preprints202509.1957.v1

4.5. Experimental Results

4.5.1. Quantitative Evaluation

Table 2 presents the results for HotpotQA. The proposed model demonstrates the highest performance in rationale inference compared to unsupervised learning-based models. Specifically, compared to the semi-supervised model, the proposed model achieves superior performance, providing a significant advantage over methods that rely on explicit silver-label learning. Compared to the approach by [40], which searches for five fixed sentences, the proposed model exhibits higher recall performance in rationale sentence inference. Since the proposed model extracts only three sentences, the performance gap is expected to be even larger when extracting the same number of sentences. The proposed model and HUG achieve comparable performance. However, HUG, which explores the optimal reasoning path by considering all combinations of the given documents and sentence pairs, has high time complexity. In contrast, the proposed model, which adds only a short decoding process to the reading model, offers similar performance to HUG while providing a speed advantage. Furthermore, considering that recall is a more important performance metric than precision when providing rationales to users, the proposed model demonstrates approximately 95% of the recall performance of the upper bound, even without annotated evidence.

Table 2. Performance comparison on HotpotQA.

Models	Answer	Rationale		
	F_1	Precision	Recall	F_1
BM25	-	-	-	40.5
RAG-small	62.8	-	-	49.0
Semi-supervised	66.0	-	-	64.5
You (2023)	50.9	-	74.2	-
HUG	66.8	-	-	67.1
Proposed Model	60.2	60.1	76.5	67.2
Upperbound	61.1	82.8	80.5	80.7

Table 3 presents the results for MuSiQue. Similar to HotpotQA, the proposed model demonstrates the highest performance compared to unsupervised rationale inference models.

Table 3. Performance comparison on MuSiQue.

Models	Answer F_1	Rationale F_1
BM25	-	12.9
RAG-small	24.2	32.0
HUG	25.1	34.2
Proposed Model	25.0	35.4

Table 4 presents the experimental results for DocRED. The proposed model achieves an Answer F_1 score of 81.8, significantly outperforming GPT_{pred}, which score of 41.8. This demonstrates the superior performance of the proposed model in relation extraction compared to GPT-based models. Additionally, the rationale F_1 scores for GPT_{pred} is 21.8, which is considerably lower than the 54.9 achieved by our model. This result indicates that the proposed model excels in rationale sentence extraction relative to GPT models. However, since these experiments are based on rationale sentences attached to the data, they provide limited insight into the comparative rationale inference capabilities of each model. To address this, the next section will further analyze how well the inferred answers and rationale sentences correspond to each other.

 doi:10.20944/preprints202509.1957.v1

Table 4. Performance comparison on DocRED.

Models	Answer F_1	Rationale F_1
GPT_{pred}	41.8	21.8
Proposed Model	81.8	54.9

4.5.2. GPT Score Evaluation

To evaluate how effectively the rationale sentences inferred by the model support the final predictions, we conducted an experiment using the GPT-4o mini model. This experiment measures the degree to which the inferred rationale sentences convincingly justify the predicted answers based on both model-generated rationales. Additionally, we included GPT_{gold}, which utilizes ground truth answers and rationale sentences from the dataset, and GPT_{pred}, which performs the dataset task while simultaneously inferring rationale sentences, as comparison models. Table 5 presents the evaluation results, using the GPT-4o mini model, showing how well the rationale sentences inferred by the model support the predicted answers across four datasets (HotpotQA, MuSiQue, DocRED, and IMDB). We considered GPT_{pred}, which performs the dataset task while simultaneously inferring rationale sentences, and GPT_{gold}, which uses the ground truth answers and rationale sentences attached to the dataset, as comparison models.

Table 5. Performance comparison of rational inference using GPT scores.

Models	Dataset				Overall
	HotpotQA	MuSiQue	DocRED	IMDB	
GPT_{pred}	92.4	81.0	51.0	95.4	79.9
GPT_{gold}	95.5	82.6	65.0	-	-
Proposed Model	87.5	55.9	87.2	88.7	79.8

First, regarding GPT_{gold}, we observe that its score does not reach 100%, even when it has access to the correct answers and rationales. This suggests a misalignment between the evaluation criteria of the datasets and those used by the GPT-4o mini model. Next, GPT_{pred} achieves over 90% of GPT_{gold}'s performance across most datasets. This demonstrates that GPT-based models can achieve relatively high rationale inference without strictly adhering to the rationale sentences provided in the ground truth data. In particular, for the HotpotQA and MuSiQue datasets, the performance gap between GPT_{pred} and GPT_{gold} is minimal, indicating that GPT models can infer evidence sentences that closely approximate the quality of the actual data. However, for the DocRED dataset, there is a significant performance gap between GPT_{pred} and GPT_{gold} (51.0 vs. 65.0). Considering the inherent limitations of GPT in relation extraction tasks, this result reflects the model's constraints in inferring accurate rationale sentences. Although GPT_{gold} results are unavailable for the IMDB dataset, GPT_{pred} achieves a high score of 95.4. This suggests that GPT models are capable of inferring highly relevant rationales in tasks involving single-document analysis, performing better in simpler structures compared to those requiring complex document linking. Overall, while GPT-based models exhibit strong performance across various datasets, there is a notable decrease in rational inference accuracy for tasks with lower performance, and a greater divergence from the annotated rationales is observed in these cases. The proposed model also demonstrates generally strong rationale inference performance, though a significant decrease is observed in the MuSiQue experiments. However, compared to the answer inference performance in earlier experiments, the rationale inference performance remains higher. This suggests that the proposed model can infer suitable rationale even when the predicted answer is incorrect. Finally, the overall scores for GPT-4o mini and the proposed model are 79.9 and 79.8, respectively, indicating very similar performance. Nevertheless, despite having only approximately 1/80 the parameters of GPT-4o mini, the proposed model demonstrates comparable rationale inference capabilities. Additionally, the proposed model can be considered an effective rationale inference framework suitable for XAI, as it demonstrates the model's reasoning process through the pointer

 doi:10.20944/preprints202509.1957.v1

network indirectly guided to select rationale sentences. The detailed prompts for the experiment are provided in Appendix C.

4.5.3. Ablation Study

To analyze the impact of individual components on the performance of the proposed model, we conducted an ablation study by selectively removing or modifying key elements. This analysis allows us to assess the significance of each component in the learning and inference processes. Table 6 presents the results of the ablation experiments conducted on the proposed model. It is evident that removing the loss function significantly reduces the effectiveness of learning in the rationale sentence extraction process. This suggests that without an appropriate loss function tailored for the pointer network, the model's ability to accurately infer rationale sentences is compromised. Additionally, the results show that beam search decoding allows for more accurate evidence inference compared to greedy decoding, enabling more refined selections. Overall, the proposed model demonstrates high performance in rationale sentence extraction by effectively utilizing both the loss function and beam search decoding, with particularly strong results in terms of recall. This leads to the conclusion that applying an appropriate learning loss and employing beam search for decoding play crucial roles in rationale sentence extraction using a pointer network.

Table 6. Ablation study of the proposed model on experimental datasets.

Models	Answer	Evidence		
	F_1	Precision	Recall	F_1
Proposed Model	60.2	60.1	76.5	67.2
- Loss	60.5	47.7	61.1	52.9
- Beam	60.8	55.3	68.2	61.1
- Loss & Beam	60.4	47.3	60.9	52.7

Given the importance of the loss function in guiding the pointer network, we further investigate how the threshold values used in the loss function affect rationale inference performance. Specifically, we examine the impact of the threshold in Equation 9 on the model's ability to extract rationale sentences accurately. Figure 5 illustrates the changes in rationale inference performance based on the threshold values used to determine α_d and α_g in the loss function (Equation 8). According to the graph, the model achieves high performance when the threshold value is between 0.4 and 0.6, with the highest rationale inference performance occurring at 0.5. Moreover, the performance sharply declines as the threshold approaches either 0 or 1. This indicates that overly strict or lenient criteria for evaluating rationale sentences can hinder the learning of the pointer network. Thus, selecting an appropriate threshold value is crucial for improving the performance of the pointer network.

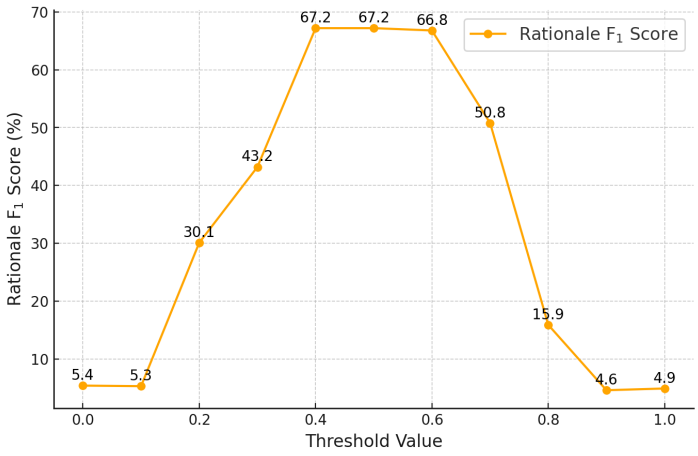

Figure 5. Rationale inference performance on the HotpotQA dataset with varying threshold values in the loss function.

 doi:10.20944/preprints202509.1957.v1

4.5.4. Further Analysis

To evaluate whether the model maintains consistent predictions when provided with only the selected rationale sentences instead of the entire document, we conducted an additional experiment. This analysis investigates the degree to which the selected rationale sentences contribute to the model's reasoning process. Table 7 presents the results verifying whether the model can infer identical predictions using only the selected rationale sentences, compared to when the entire document is provided, as mentioned in the introduction. The experiment was conducted on the HotpotQA evaluation dataset. In the table, "Proposed Model (All Document $\Leftrightarrow$ Only Rationale Sentences Input)" shows the F_1 score comparing the predictions inferred by the proposed model using the entire document versus those inferred using only the selected rationale sentences. Additionally, for comparison, we measure the F_1 scores between the inferred predictions and the dataset's ground truth in both cases: "Ground Truth $\Leftrightarrow$ All Document Input" and "Ground Truth $\Leftrightarrow$ Only Rationale Sentences Input". Here, "Test set F_1" represents the evaluation results on the entire HotpotQA test set, while "Sampled Test set F_1" measures the performance only on the subset of the test set where the predictions inferred from the entire document exactly match the ground truth.

Table 7. Consistency Evaluation on HotpotQA.

Models	Test set F_1	Sampled Test set F_1
Proposed Model (All Document Input $\Leftrightarrow$ Only Rationale Sentences Input)	84.9	93.3
Proposed Model (Ground Truth $\Leftrightarrow$ Only Rationale Sentences Input)	60.2	90.3
Proposed Model (All Document Input $\Leftrightarrow$ Ground Truth)	60.2	100

The evaluation results in Table 7 demonstrate the consistency of predictions made by the proposed model when using the entire document compared to when using only the selected rationale sentences. The key findings are as follows: In the results for "Proposed Model (All Document Input $\Leftrightarrow$ Only Rationale Sentences Input)", the Test set F_1 score is 84.9, showing a high level of consistency between the answers predicted based on the entire document and those predicted using only the rationale sentences. Furthermore, the Sampled Test set F_1 score increases to 93.3, indicating that the selected rationale sentences effectively preserve essential information for generating accurate predictions. This score is particularly noteworthy in cases where the predictions based on the entire document exactly match the Ground Truth. As a result, the proposed model demonstrates strong consistency in predictions even when only the selected rationale sentences are provided, achieving results comparable to when the entire document is used. Additionally, the increase in the Sampled Test set F_1 score highlights the robustness of rationale-based predictions in scenarios where predictions align with the Ground Truth.

5. Conclusion

We propose a framework for effective rationale inference in explainable multi-hop question answering. Our model utilizes a pointer network to accumulate important information and infer answers, effectively selecting evidence sentences even in the absence of annotations. Furthermore, the model's reasoning process is interpretable through the pointer network's sentence selection. Compared to prior multi-hop question-answer models, our approach achieves state-of-the-art performance across multiple metrics and provides supporting evidence even when predictions are incorrect. Moreover, compared to the recently released GPT-4o mini, our model-despite having only approximately 1/80 of the parameters-exhibited comparable rationale inference capabilities.

6. Patents

This section is not mandatory, but may be added if there are patents resulting from the work reported in this manuscript.

 doi:10.20944/preprints202509.1957.v1

Author Contributions: Conceptualization, K.H., Y.J. and H.K.; methodology, K.H. and Y.J.; software, K.H. and Y.J.; validation, K.H., Y.J. and H.K.; formal analysis, K.H. and Y.J.; investigation, K.H. and Y.J.; resources, H.K.; data curation, K.H. and Y.J.; writing—original draft preparation, K.H. and Y.J.; writing—review and editing, K.H., Y.J. and H.K.; visualization, K.H. and Y.J.; supervision, H.K.; project administration, H.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Institute for Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2022-II220369, (Part 4) Development of AI Technology to support Expert Decision-making that can Explain the Reasons/Grounds for Judgment Results based on Expert Knowledge).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original data presented in this study are openly available. The HotpotQA dataset can be accessed at https://hotpotqa.github.io. The MuSiQue dataset is available at https://github.com/stonybrooknlp/musique. The DocRED dataset can be accessed at https://github.com/thunlp/DocRED. The IMDB dataset is available via the Large Movie Review Dataset at https://ai.stanford.edu/~amaas/data/sentiment.

Acknowledgments: This work was supported by Institute for Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. RS-2022-II220369, (Part 4) Development of AI Technology to support Expert Decision-making that can Explain the Reasons/Grounds for Judgment Results based on Expert Knowledge)

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Implementation Details

In this appendix, we provide detailed implementation specifications for each task. These details are intended to facilitate reproducibility and further understanding of our approach. The proposed model employs a pre-trained ELECTRA-base [31] to process input sequences across different tasks. The dataset structure and mini-batch formation differ for each task, as follows:

- Multi-Hop Question Answering:
  - The input sequence T follows the format: '$[CLS]$ Question $[SEP]$ Document $[SEP]$'. The encoded token representations$T$ are further processed to extract key evidence for reasoning.
  - Since each question is associated with multiple documents, each mini-batch contains multiple document-question pairs for the same question.
- Document-Level Relation Extraction:
  - The input T consists of one or more sentences containing entity mentions, where the relationships between entities must be inferred. ELECTRA encodes these token embeddings V, capturing contextual information to analyze entity relationships.
  - In this task, each data sample is processed as an individual document, and the mini-batch contains different documents.
- Binary Sentiment Classification:
  - The input T consists of a review text or a short comment, where the final token representations V are used for sentiment classification.
  - Similar to relation extraction, each sample corresponds to a single document, and the mini-batch contains different texts.

Figure A1 illustrates the process of generating sentence vectors using mean-pooling.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 September 2025 doi:10.20944/preprints202509.1957.v1

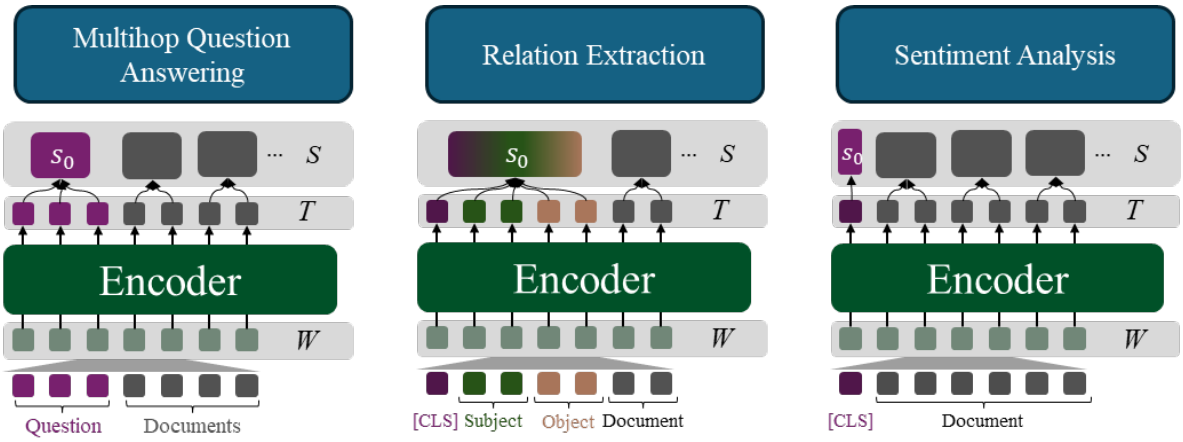

Figure A1. An overview of model input structures and sentence representation generation via mean-pooling for different tasks.

Next, we describe the output layer, which computes the final probability distribution for each task using the rationale vector generated by the Pointer Network. We implement and validate the proposed framework across three tasks: (1) multi-hop question answering, (2) relation extraction, and (3) sentiment analysis. The output layer architecture for each task is illustrated in Figure A2 below.

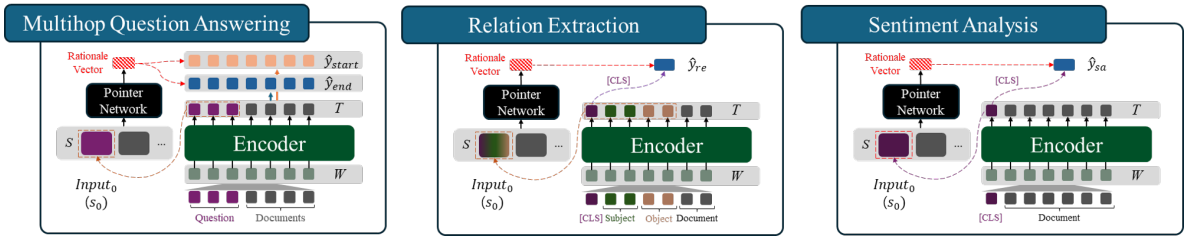

Figure A2. Task-specific output layer architecture utilizing the rationale vector to compute the final probability distribution.

- For detailed implementation equations of multi-hop question answering, refer to Equation 6.
- Relation extraction involves classifying the relationship between a subject and an object within a given sentence or document. To achieve this, the input sequence follows the structure shown in Figure A2. A probability distribution over relation labels C_{re} is then computed based on $t_{[CLS]}$, which contains the contextual information of the input sequence. This process is represented by Equation A1 below.

$$\hat{y}_{re} = W_{re}(t_{[CLS]} \oplus input_N)^T \in \mathbb{R}^{|C_{re}|} \tag{A1}$$

- Sentiment analysis, as a form of sentence classification, identifies the sentiment within the input text and categorizes it as positive, negative, or neutral. The input and output structure for this task is shown in Figure A2. Similar to relation extraction, a probability distribution over sentiment labels C_{sa} is generated according to Equation A2.

$$\hat{y}_{sa} = W_{sa}(t_{[CLS]} \oplus input_N)^T \in \mathbb{R}^{|C_{sa}|} \tag{A2}$$

Appendix B. GPT-Based Evaluation Instructions and Inputs for All Tasks

This appendix provides detailed instructions and input formats for all tasks used in the GPT-based evaluation. Each task includes a description, the specific instruction given to the model, and an

 doi:10.20944/preprints202509.1957.v1

example input. The following Table A1, Table A2 and Table A3 present examples used in the evaluation of the multi-hop question answering, relation extraction and sentiment analysis tasks, respectively.

Table A1. Instructions for GPT-Based Evaluation in Multi-hop QA.

Instruction
Determine how validly the provided <Sentences>support the given <Answer>to the <Question>. Your task is to assess the supporting sentences based on their relevance and strength in relation to the answer, regardless of whether the answer is correct. The focus is on evaluating the validity of the evidence itself. Even if the answer is incorrect, a supporting sentence can still be rated highly if it is relevant and strong. <Score criteria> - 0 : The sentences do not support the answer. They are irrelevant, neutral, or contradict the answer. - 1 : The sentences provide partial or unclear support for the answer. The connection is weak, lacking context, or not directly related to the answer. - 2 : The sentences strongly support the answer, making it clear and directly inferable from them. <Output format> <Score>: <0, 1, or 2>
Input
<Question> When did the park at which Tivolis Koncertsal is located open? </Question> <Answer> 15 August 1843 </Answer> <Sentences> Tivolis Koncertsal is a 1,660-capacity concert hall located at Tivoli Gardens in Copenhagen, Denmark. The building, which was designed by Frits Schlegel and Hans Hansen, was built between 1954 and 1956. The park opened on 15 August 1843 and is the second-oldest operating amusement park in the world, after ... </Sentences> <Score>:

Table A2. Instructions for GPT-based evaluation in Relation Extraction.

Instruction
Determine whether the relationship between the given <Subject> and <Object> can be inferred solely from the provided <Sentences>. The <Relationship>may not be explicitly stated, and it might even be incorrect. However, your task is to evaluate whether the sentences themselves suggest the given relationship, regardless of its accuracy. <Score criteria> - 0: The sentences do not suggest the relationship at all. The sentences are neutral, irrelevant, or contradict the relationship. - 1: The sentences somewhat suggest the relationship but are not conclusive. The relationship is partially inferred but not clearly established. - 2: The sentences fully suggest the relationship. The relationship can be clearly and directly inferred from the sentences alone. <Output format> Score: <0, 1, or 2>
Input
<Sentences> The discovery of the signal in the chloroplast genome was announced in 2008 by researchers from the University of Washington. Somehow, chloroplasts from V. orcuttiana, swamp verbena (V. hastata) or a close relative of these had admixed into the G. bipinnatifida genome. </Sentences> <Subject> mock vervains </Subject> <Object> Verbenaceae </Object> <Relationship> parent taxon : closest of the taxon in question </Relationship> <Score>:

 doi:10.20944/preprints202509.1957.v1

Table A3. Instructions for GPT-based evaluation in the Sentiment Analysis.

Instruction
Determine whether the given <Sentiment> can be derived solely from the <Supporting Sentences> for the given <Review>. The given <Sentiment>may not be the correct answer, but evaluate whether the <Supporting Sentences> alone can support it. <Score criteria> - 0: The supporting sentences do not support the sentiment at all. The facts are neutral, irrelevant to the sentiment, or contradict the sentiment. - 1: The supporting sentences somewhat support the sentiment but are not conclusive. The sentiment is partially inferred but not clearly. The facts suggest the sentiment but do not decisively establish it. - 2: The supporting sentences fully support the sentiment. The sentiment can be clearly and directly inferred from the facts alone. <Output format> Score: <0, 1, or 2>
Input
<Sentiment> positive </Sentiment> <Supporting Sentences> A trite fish-out-of-water story about two friends from the midwest who move to the big city to seek their fortune. They become Playboy bunnies, and nothing particularly surprising happens after that. </Supporting Sentences> <Score>:

Appendix C. GPT-Based Answer and Supporting Sentence Extraction for All Tasks

In this section, we provide the prompts for GPT*pred*, which simultaneously infers both the answer and the corresponding rationale sentences. The prompts for different tasks are shown in Tables A4, Table A5, and Table A6.

Table A4. Instructions for GPT-based Answer and Supporting Sentence Extraction in Multi-hop QA.

Instruction
Answer the given <Question> using only the provided <Reference documents>. Some documents may be irrelevant. Keep the answer concise, extracting only key terms or phrases from the <Reference documents> rather than full sentences. Extract exactly 3 supporting sentences—no more, no less. For each supporting sentence, provide its sentence number as it appears in the reference documents. <Output format> <Answer>: <Generated Answer> <Supporting Sentences>: <Sentence Number 1>, <Sentence Number 2>, <Sentence Number 3>
Input
<Question> When did the park at which Tivolis Koncertsal is located open? </Question> <Reference documents> Document 1 : Tivolis Koncertsal [1] Tivolis Koncertsal is a 1,660-capacity concert hall located at Tivoli Gardens in Copenhagen, Denmark. [2] The building, which was designed by Frits Schlegel and Hans Hansen, was built between 1954 and 1956. Document 2 : Tivoli Gardens [3] Tivoli Gardens (or simply Tivoli) is a famous amusement park and pleasure garden in Copenhagen, Denmark. [4] The park opened on 15 August 1843 and is the second-oldest operating amusement park in the world, after ... Document 3 : Takino Suzuran Hillside National Government Park [5] Takino Suzuran Hillside National Government Park is a Japanese national government park located in Sapporo, Hokkaido. [6] It is the only national government park in the northern island of Hokkaido. [7] The park area spreads over 395.7 hectares of hilly country and ranges in altitude between 160 and 320 m above sea level. [8] Currently, 192.3 is accessible to the public. ... </Reference documents> <Answer>: <Supporting Sentences>:

 doi:10.20944/preprints202509.1957.v1

Table A5. Instructions for GPT-based Answer and Supporting Sentence Extraction in Relation Extraction.

Instruction
Determine the relationship between the given <Subject>and <Object>. The relationship must be selected from the following list: 'head of government', 'country', 'place of birth', 'place of death', 'father', 'mother', 'spouse', ... After selecting the appropriate relationship, provide two key sentence numbers that best support this relationship. <Output format> <Relationship>: <Extracted Relationship> <Supporting Sentences>: <Sentence Number 1>, <Sentence Number 2>
Input
<Document> [1] Since the new chloroplast genes replaced the old ones, it may be that the possibly ... [2] Glandularia, common name mock vervain or mock verbena, is a genus of annual and perennial herbaceous flowering ... [3] They are native to the Americas. [4] Glandularia species are closely related to the true vervains and sometimes still ... ... </Document> <Subject> mock vervains </Subject> <Object> Verbenaceae </Object> <Relationship>: <Supporting Sentences>:

Table A6. Instructions for GPT-based Answer and Supporting Sentence Extraction in Sentiment Analysis.

Instruction
Classify the sentiment of the given <Sentence> as either 'positive' or 'negative'. After selecting the appropriate sentiment, extract **only two** key sentences that best support this sentiment. <Output format> <Sentiment>: <Extracted Sentiment> <Supporting Sentences>: <Sentence Number 1>, <Sentence Number 2>
Input
<Document> [1] This movie was awful. [2] The ending was absolutely horrible. [3] There was no plot to the movie whatsoever. [4] The only thing that was decent about the movie was the acting done by Robert ... ... </Document> <Sentiment>: <Supporting Sentences>:

References

1. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* **2023**.
2. Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* **2023**.
3. Zhao, W.X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. A survey of large language models. *arXiv preprint arXiv:2303.18223* **2023**.
4. Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* **2024**.
5. Bender, E.M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, 2021, pp. 610–623.

 doi:10.20944/preprints202509.1957.v1

6. Bommasani, R.; Hudson, D.A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M.S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* **2021**.
7. Rudin, C.; Chen, C.; Chen, Z.; Huang, H.; Semenova, L.; Zhong, C. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surveys* **2022**, *16*, 1–85.
8. Shwartz, V.; Choi, Y. Do neural language models overcome reporting bias? In Proceedings of the Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 6863–6870.
9. Chen, J.; Lin, S.t.; Durrett, G. Multi-hop question answering via reasoning chains. *arXiv preprint arXiv:1910.02610* **2019**.
10. Wu, H.; Chen, W.; Xu, S.; Xu, B. Counterfactual supporting facts extraction for explainable medical record based diagnosis with graph network. In Proceedings of the Proceedings of the 2021 conference of the north American chapter of the association for computational linguistics: human language technologies, 2021, pp. 1942–1955.
11. Zhao, W.; Chiu, J.; Cardie, C.; Rush, A.M. Hop, Union, Generate: Explainable Multi-hop Reasoning without Rationale Supervision. In Proceedings of the Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 16119–16130.
12. Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W.; Salakhutdinov, R.; Manning, C.D. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In Proceedings of the Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 2369–2380.
13. Qi, P.; Lin, X.; Mehr, L.; Wang, Z.; Manning, C.D. Answering Complex Open-domain Questions Through Iterative Query Generation. In Proceedings of the Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 2590–2602.
14. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.t.; Rocktäschel, T.; et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems* **2020**, *33*, 9459–9474.
15. Ribeiro, M.T.; Singh, S.; Guestrin, C. " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.
16. DW, G.D.A. DARPA's explainable artificial intelligence program. *AI Mag* **2019**, *40*, 44.
17. Jiang, Z.; Xu, F.F.; Araki, J.; Neubig, G. How can we know what language models know? *Transactions of the Association for Computational Linguistics* **2020**, *8*, 423–438.
18. Arras, L.; Montavon, G.; Müller, K.R.; Samek, W. Explaining Recurrent Neural Network Predictions in Sentiment Analysis. In Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis WASSA 2017: Proceedings of the Workshop. The Association for Computational Linguistics, 2017, pp. 159–168.
19. Scott, M.; Su-In, L.; et al. A unified approach to interpreting model predictions. *Advances in neural information processing systems* **2017**, *30*, 4765–4774.
20. Alvarez-Melis, D.; Jaakkola, T.S. On the Robustness of Interpretability Methods. *arXiv e-prints* **2018**, pp. arXiv–1806.
21. Gilpin, L.H.; Bau, D.; Yuan, B.Z.; Bajwa, A.; Specter, M.; Kagal, L. Explaining explanations: An overview of interpretability of machine learning. In Proceedings of the 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA). IEEE, 2018, pp. 80–89.
22. Jiang, Z.; Zhang, Y.; Yang, Z.; Zhao, J.; Liu, K. Alignment rationale for natural language inference. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 5372–5387.
23. Atanasova, P.; Simonsen, J.G.; Lioma, C.; Augenstein, I. Diagnostics-guided explanation generation. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2022, Vol. 36, pp. 10445–10453.
24. Welbl, J.; Stenetorp, P.; Riedel, S. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics* **2018**, *6*, 287–302.
25. Yu, X.; Min, S.; Zettlemoyer, L.; Hajishirzi, H. CREPE: Open-Domain Question Answering with False Presuppositions. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 10457–10480.

 doi:10.20944/preprints202509.1957.v1

26. Glockner, M.; Habernal, I.; Gurevych, I. Why do you think that? exploring faithful sentence-level rationales without supervision. *arXiv preprint arXiv:2010.03384* **2020**.
27. Min, S.; Zhong, V.; Zettlemoyer, L.; Hajishirzi, H. Multi-hop Reading Comprehension through Question Decomposition and Rescoring. In Proceedings of the Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 6097–6109.
28. Mao, J.; Jiang, W.; Wang, X.; Liu, H.; Xia, Y.; Lyu, Y.; She, Q. Explainable question answering based on semantic graph by global differentiable learning and dynamic adaptive reasoning. In Proceedings of the Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022, pp. 5318–5325.
29. Groeneveld, D.; Khot, T.; Mausam.; Sabharwal, A. A Simple Yet Strong Pipeline for HotpotQA. In Proceedings of the Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); Webber, B.; Cohn, T.; He, Y.; Liu, Y., Eds., Online, nov 2020; pp. 8839–8845. https://doi.org/10.18653/v1/2020.emnlp-main.711.
30. Yin, Z.; Wang, Y.; Hu, X.; Wu, Y.; Yan, H.; Zhang, X.; Cao, Z.; Huang, X.; Qiu, X. Rethinking label smoothing on multi-hop question answering. In Proceedings of the China National Conference on Chinese Computational Linguistics. Springer, 2023, pp. 72–87.
31. Clark, K. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555* **2020**.
32. Vinyals, O.; Fortunato, M.; Jaitly, N. Pointer networks. *Advances in neural information processing systems* **2015**, *28*.
33. Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* **2014**.
34. Brown, P.F.; Della Pietra, S.A.; Della Pietra, V.J.; Mercer, R.L. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics* **1993**, *19*, 263–311.
35. Trivedi, H.; Balasubramanian, N.; Khot, T.; Sabharwal, A. MuSiQue: Multihop Questions via Single-hop Question Composition. *Transactions of the Association for Computational Linguistics* **2022**, *10*, 539–554.
36. Yao, Y.; Ye, D.; Li, P.; Han, X.; Lin, Y.; Liu, Z.; Liu, Z.; Huang, L.; Zhou, J.; Sun, M. DocRED: A Large-Scale Document-Level Relation Extraction Dataset. In Proceedings of the Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 764–777.
37. Maas, A.L.; Daly, R.E.; Pham, P.T.; Huang, D.; Ng, A.Y.; Potts, C. Learning Word Vectors for Sentiment Analysis. In Proceedings of the Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011, pp. 142–150.
38. Kullback, S.; Leibler, R.A. On information and sufficiency. *The annals of mathematical statistics* **1951**, *22*, 79–86.
39. Robertson, S.E.; Walker, S.; Jones, S.; Hancock-Beaulieu, M.M.; Gatford, M.; et al. Okapi at TREC-3. *Nist Special Publication Sp* **1995**, *109*, 109.
40. You, H. Multi-grained unsupervised evidence retrieval for question answering. *Neural Computing and Applications* **2023**, *35*, 21247–21257.
41. OpenAI. Gpt-4o mini: advancing cost-efficient intelligence, 2024. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/.