International Journal of Machine Learning and Cybernetics
https://doi.org/10.1007/s13042-025-02712-y

ORIGINAL ARTICLE

Improving multi-hop question answering with prompting explicit and implicit knowledge aligned human reading comprehension

Guangming Huang[1] · Yunfei Long[1] · Cunjin Luo[1]

Received: 9 January 2025 / Accepted: 6 June 2025

Abstract

Language models (LMs) utilize chain-of-thought (CoT) to imitate human reasoning and inference processes, achieving notable success in multi-hop question answering (QA). Despite this, a disparity remains between the reasoning capabilities of LMs and humans when addressing complex challenges. Psychological research highlights the crucial interplay between explicit content in texts and prior human knowledge during reading. However, current studies have inadequately addressed the relationship between input texts and the pre-training-derived knowledge of LMs from the standpoint of human cognition. In this paper, we propose a **P**rompting **E**xplicit and **I**mplicit knowledge (PEI) framework, which employs CoT prompt-based learning to bridge explicit and implicit knowledge, aligning with human reading comprehension for multi-hop QA. PEI leverages CoT prompts to elicit implicit knowledge from LMs within the input context, while integrating question type information to boost model performance. Moreover, we propose two training paradigms for PEI, and extend our framework on biomedical domain QA to further explore the fusion and relation of explicit and implicit biomedical knowledge via employing biomedical LMs in the Knowledge Prompter to invoke biomedical implicit knowledge and analyze the consistency of the domain knowledge fusion. The experimental results indicate that our proposed PEI performs comparably to the state-of-the-art on HotpotQA, and surpasses baselines on 2WikiMultihopQA and MuSiQue. Additionally, our method achieves a significant improvement compared to baselines on MEDHOP. Ablation studies further validate the efficacy of PEI framework in bridging and integrating explicit and implicit knowledge.

Keywords Multi-hop question answering · Prompt tuning · Chain-of-thought · Implicit knowledge

1 Introduction

Multi-hop question answering (QA) poses a significant challenge, requiring sophisticated reasoning and inference across multiple sources to derive a coherent and accurate answer [63]. Chain-of-thought (CoT) mimics human reasoning by generating a series of intermediate natural language steps that guide the model toward the final answer for complex reasoning tasks. Recent studies utilizing CoT prompt-based learning on language models (LMs) have shown considerable effectiveness in tackling multi-hop QA [9, 51, 66].

Despite the advancements in LMs, there remains a significant gap between their reasoning abilities and human cognitive processes in addressing intricate problems. Current research has yet to adequately investigate the interplay between input texts and the pre-training-derived knowledge of LMs, particularly through the lens of cognitive science.

In studies of human reading comprehension, Smith [48] suggests that information is often reiterated during reading, resulting in redundancies at various linguistic levels, including letter-to-letter, word-to-word, sentence-to-sentence, and text-to-text. As a result, readers are able to reduce their dependence on explicit information details within the text by integrating external sources of information, such as world knowledge [17]. According to the findings of Clarke and Silberstein [8], readers engage in reading comprehension and question-answering process while reading, drawing upon both the explicit information conveyed in the text and their

✉ Yunfei Long
yl20051@essex.ac.uk

Guangming Huang
gh22231@essex.ac.uk

Cunjin Luo
cunjin.luo@essex.ac.uk

[1] School of Computer Science and Electronic Engineering, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK

Published online: 16 June 2025

pre-existing language knowledge, background knowledge, and world knowledge derived from that explicit information. Certain studies have pointed out that a critical factor in reading ability is what the reader brings to the text, or what is generally referred to as prior knowledge [1, 2, 64]. Related experimental findings further reveal a significant positive correlation between human reading comprehension and prior knowledge [1].

For instance, as depicted in Fig. 1, consider the question "*Was Morris Lee born in the capital of the Democratic Republic of the Congo?*". A human reader would retrieve relevant information from the provided passages and, based on the auxiliary verb "*was*" in the yes-no question, infer the answer "*yes*" or "*no*" drawing upon linguistic knowledge (as part of implicit knowledge), even in the absence of information regarding the capital of Congo.

Therefore, an inherent and inseparable connection prevails between the explicit information within text context and pre-existing prior knowledge of human being. The prior knowledge lessens the dependence on explicit details, thus reducing the necessities for redundant information during inference and reasoning. In addition, the harmonious fusion of explicit information and prior knowledge improve the effectiveness of the reading process, contributing to enhanced comprehension and deeper engagement.

Building on insights from the theories of human cognition mentioned above, we introduce a novel framework, referred to as **P**rompting **E**xplicit and **I**mplicit knowledge (PEI), to address the challenges multi-hop QA. In this framework, readers are analogized to LMs, where their prior knowledge represents implicit knowledge gained through pre-training, and the explicit information within passages serves as the input context conveying explicit knowledge. While acknowledging the inherent differences between LMs and human beings, and recognizing the limitations in directly considering readers as LMs, Jin and Rinard [24] argue that LMs surpass beyond mere "stochastic parrots" [4], as they possess the capacity to acquire meaningful semantic information during pre-training. Complex question answering encompasses high-complexity, non-factoid inquiries that require multi-step decomposition and integration of multiple information sources. The decomposition process is fundamental to this problem-solving paradigm, as it transforms initially intractable complex questions into manageable subproblems. Furthermore, as evidenced in recent chain-of-thought approaches and prompt-based methodologies for LLMs, these explicit reasoning steps not only enhance the models' problem-solving capabilities but also render the solutions auditable, verifiable, and interpretable when errors occur [5, 42].

To make use of these knowledge sources, we utilize CoT prompting to capture explicit knowledge and activate implicit knowledge. Intuitively, this approach effectively bridges these knowledge types, thereby enhancing the reasoning performance of PEI framework for multi-hop QA. Additionally, PEI reduces dependency on the explicit information details contained within input passages by enabling the selective removal of irrelevant or "redundant" information unrelated to the questions, aligning with Smith [48]'s theory. To further demonstrate the significant role of implicit knowledge in boosting the proposed framework's performance, we conduct ablation studies, which corroborates our hypothesis (refer to Sect. 4.7).

As illustrated in Fig. 2, our proposed PEI framework consists of three main components: (i) **The Type Prompter** is designed to identify and learn the weights of reasoning types for given questions; (ii) **The Knowledge Prompter** acquires implicit knowledge by leveraging explicit knowledge, which employs an encoder-decoder foundation language module, mirroring human reading comprehension based on psychological cognition theories; (iii) **The Unified Prompter** fuses explicit, implicit knowledge and question types for multi-hop QA, which uses the same foundation model as Type Prompter. Moreover, our proposed framework offers flexibly

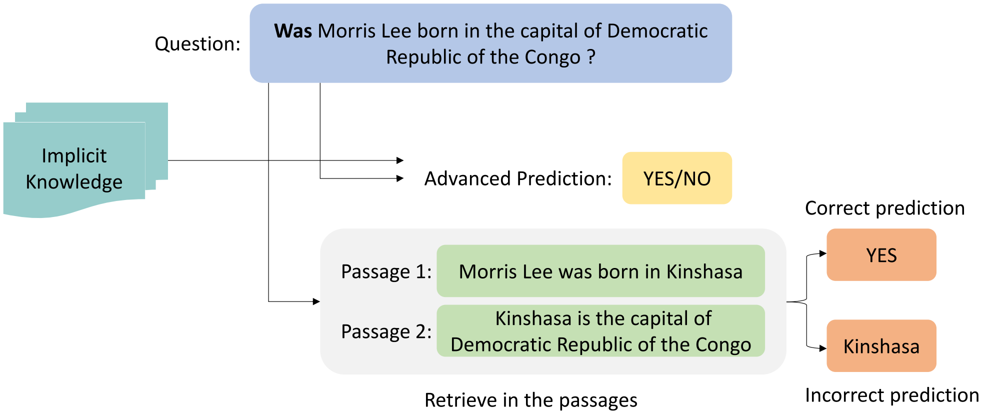
Question:
Was Morris Lee born in the capital of Democratic Republic of the Congo ?
Implicit Knowledge
Advanced Prediction:
YES/NO
Correct prediction
YES
Passage 1:
Morris Lee was born in Kinshasa
Passage 2:
Kinshasa is the capital of Democratic Republic of the Congo
Kinshasa
Retrieve in the passages
Incorrect prediction

Fig. 1 An example of the significance of implicit knowledge in reading comprehension

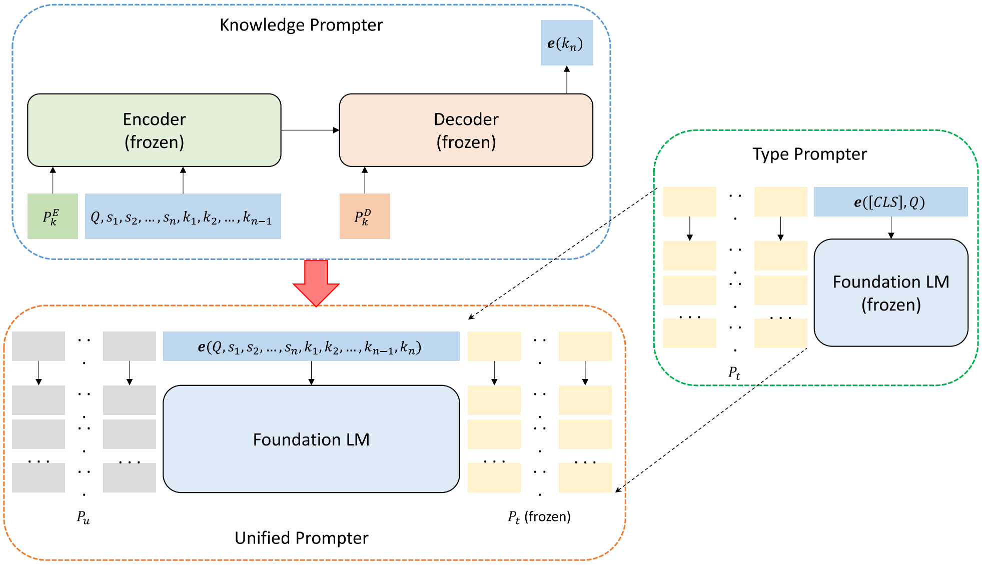

Fig. 2 The overview of our proposed PEI framework for multi-hop QA. The right green dashed block is the Type Prompter; the top blue dashed block refers to the Knowledge Prompter; and the bottom orange dashed block is the Unified Prompter

replace the foundation models to adapt to different requirements, such as those specific to the biomedical domain or constraints related to computational cost. Hence, We propose two training paradigms to PEI framework, which employs various sizes of foundation LMs (i.e., Llama 3.1-8B and ELECTRA).

Tackling questions in the biomedical field frequently requires multi-step reasoning. For example, a clinician might inquire, “Which tests are required for patients exhibiting [specific symptoms]?” To answer, models must: (i) deduce the possible diseases involved and (ii) determine the appropriate tests for differential diagnosis. To address these challenges, we extend our proposed PEI framework to biomedical domain for further exploring the fusion and connection of explicit and implicit biomedical knowledge in addition to general domain. Biomedical QA significantly deviates from general QA in content, scope, and methodology because of the complex nature of biomedical information [20, 25]. Zweigenbaum [69] was the first to highlight the distinctive characteristics of Biomedical QA compared to general domain QA. Biomedical QA deals with specific technical terms (e.g., “pharmacokinetics,” “tyrosine kinase inhibitors” and “monoclonal antibodies”) and entails the comprehension of detailed, evidence-based information, such as the interpretation of clinical trials and understanding action mechanisms. Additionally, biomedical information is frequently incomplete, ambiguous, or subject to constant change. Therefore, it is essential to comprehensively evaluate our proposed framework within the biomedical domain to ensure its effectiveness and reliability. We utilize a biomedical encoder-decoder foundation model within the Knowledge Prompter to harness specialized implicit knowledge, which is then integrated into the Unified Prompter with explicit knowledge derived from the provided texts. Moreover, we examine the consistency of the domain knowledge integration.

Our contributions are summarized as follows:

- We introduce the PEI framework that offers a proficient method for multi-hop QA, based on the human reading process, by modeling the input passages or context as explicit knowledge and invoking the pre-trained knowledge of LMs as implicit knowledge that mirroring with human prior knowledge.
- We propose two training paradigms to PEI framework, which employs various sizes of foundation LMs (i.e., Llama 3.1-8B and ELECTRA). The experiment results show that the performance of Llama-based PEI with prompt tuning is slightly lower that of the standard PEI, but significantly reduces the number of trainable parameters while maintaining comparable reasoning performance.
- Our PEI framework demonstrates performance on par with state-of-the-art baseline evaluating on the benchmark HotpotQA dataset. Furthermore, PEI shows consistent effectiveness and robustness on single-hop sub-questions and additional multi-hop datasets (2WikiMultiHopQA and MuSiQue).
- We further evaluate PEI framework on biomedical domain to comprehensively explore the explicit and implicit knowledge fusion in specify domain. The experimental results shows that the proposed PEI framework

significantly outperforms the baselines on MEDHOP dataset.

- The ablation studies corroborate that implicit knowledge improves the reasoning abilities of PEI framework, thereby supporting our hypothesis regarding PEI, which grounded in the human cognition theories for reading comprehension.

2 Related work

2.1 Chain-of-thought prompting

Prompt tuning[1] has been acknowledged as a potent method to tune LMs to harness pertinent knowledge for targeted downstream tasks [62]. CoT prompting, which is a prompt-driven strategy, has surfaced as a technique to extract implicit knowledge from large language models (LLMs) for intricate reasoning tasks. It mirrors the sequential and coherent thought processes of humans by creating intermediate reasoning steps in natural language that culminate in the final result [56, 67]. Manual-CoT [56] aimed to extract CoT reasoning capability via manual demonstrations. Subsequently, Kojima et al [28] showed that LLMs can effectively act as zero-shot reasoners, producing rationales that inherently contain CoT reasoning by using the phrase "*Let's think step by step*" to encourage a detailed thought process before deriving answers. AutoCoT, an automatic CoT prompting approach [67], employed various question sampling and reasoning chain construction to form demonstrations, thereby reducing the need for human input. Recent research have delved into CoT prompt learning for multi-hop QA [51, 53]. Building on aforementioned studies, our study investigates the application of CoT prompting to extract implicit knowledge from LMs. Unlike CoT, which produces intermediate steps in natural language, our approach produces continuous embeddings to represent implicit knowledge.

2.2 Prompt-based learning for multi-hop QA

Significant advancements in recent research have been made by incorporating prompts for multi-hop QA [21, 33, 68]. For instance, PromptRank [27] developed an instruction-based prompt, integrating a candidate document pathway to calculate the relevance between a given question and the documented path. This relevance is evaluated via the conditional likelihood of the question in relation to the path prompt, as assessed by a language model. In contrast, IRCoT [51] implemented a system that alternated between CoT generation and knowledge retrieval steps, leveraging CoT prompting to direct the retrieval process. Wang et al [53] proposed an iterative CoT prompting method that progressively extracts knowledge from LMs using a sequence-to-sequence BART-large model, thereby recalling natural language sequences for multi-hop QA. Each triplet in the evidence path is transformed into a natural language statement through a straightforward template, cumulatively generating the final statement. Building on this concept, our method employs a similar encoder-decoder foundation LM (i.e., BART-large) for recalling implicit knowledge by an approach of iterative prompting.

Comparing to the aforementioned studies, our proposed method distinguishes in the following three aspects: (i) PEI eliminates the need to transform triple evidence paths into natural language statements; (ii) we utilizes input passages to explicitly draw upon implicit knowledge from LMs, which previous methods have yet explored; (iii) PEI represents the recalled implicit knowledge as continuous embeddings, as opposed to using natural language statements or lexical knowledge [19]. Consequently, our framework is not dependent on natural language statements originating from evidence paths.

2.3 Biomedical multi-hop QA

Biomedical multi-hop QA is a specialized domain of multi-hop QA that involves reasoning over multiple interconnected biomedical facts to answer complex queries [25]. Recent studies utilize a combination of LMs and structured knowledge representations such as biomedical knowledge graphs [11, 14], or external medical knowledge [15]. Du et al [11] proposed Adversarial Entity Graph Convolutional Networks (AEGCN), constructing an enriched entity graph with innovative edge relationships derived from supporting text while leveraging adversarial entities during training to enhance the model's resistance to interference. MedKGQA [15] addressed drug-drug interaction (DDI) prediction and medical reasoning by combining external medical knowledge bases with "drug-protein" triplets and graph neural networks (GNNs) to navigate and extract answers from biomedical pathways effectively.

Comparing with the previous methods introduced in related works, our proposed PEI framework employs CoT prompting to elicit implicit biomedical knowledge from LMs, reducing the cost of constructing of knowledge bases. Meanwhile, PEI framework is flexible to utilize various foundation LMs to adapt specialty knowledge requirements.

[1] The term "prompt tuning" refers to a broad array of methods rather than a specific approach.

3 Methodology

3.1 Problem statement

Multi-hop QA represents a challenging natural language processing task wherein a system must generate responses by retrieving and reasoning across multiple evidence fragments from disparate textual sources. In contrast to single-hop QA, which extracts answers from individual passages, multi-hop QA necessitates complex inferential reasoning to synthesize information across disconnected documents. Formally, given a complex query Q and a collection of supporting sentences $S_n = [s_1, s_2, ..., s_i, ..., s_n]$, the objective is to derive the correct answer A through sequential inferential processes that traverse the relevant textual evidence, and the corresponding supporting sentences S_k, where $S_k \subseteq S_n$.

3.2 Framework overview

As illustrated in Fig. 2, our proposed PEI framework comprises three primary components: (i) Type Prompter identifies reasoning types for given questions and learns their respective weights; (ii) Knowledge Prompter acquires implicit knowledge by leveraging explicit knowledge through an encoder-decoder foundation language model, reflecting the human reading process as informed by psychological cognition theories; and (iii) Unified Prompter, which integrates explicit knowledge, implicit knowledge, and question types for multi-hop QA using the identical foundation backbone model as Type Prompter. PEI framework facilitates flexible substitution of foundation models to accommodate diverse requirements, such as domain-specific adaptation for biomedical applications or optimization of computational resource allocation.

Pre-training on single-hop QA. To analyze the capabilities of QA models throughout each step of reasoning processes for multi-hop QA, our study utilizes a foundational LM, namely ELECTRA[2] [7], trained on the single-hop QA dataset SQuAD [45]. Following this, we employ the pre-trained ELECTRA model as the foundation LM for our Type Prompter component. By deploying the ELECTRA model trained on single-hop tasks, we endeavor to investigate the interplay between the model's behavior and reasoning processes across multi-hop reasoning stages.

Note that the LLMs (i.e. Llama 3.1 [12]) also be employed as the foundation model, however, we do not fine-tune Llama on SQuAD since the computational cost of full parameter fine-tuning for LLMs is costly. Additionally, SQuAD dataset is a sub-task of the LLM benchmark.

3.3 Type prompter

The Type Prompter is designed to enhance the training of acquired weights for soft prompts, allowing them to adeptly learn the unique features of different question types. As illustrated in Fig. 2, the yellow blocks P_t denote trainable prompt embeddings, whereas the blue blocks represent the input embeddings and frozen foundation LM.

The P-tuning v2 approach [34] is utilized to trainable soft prompts P_t, learning the weights and capturing specific-type information of the given queries. Initially, the foundation LM remains frozen while the trainable soft prompt P_t is optimized. Upon training, the updated P_t is linked to the Unified Prompter module, while preserving its fixed nature throughout subsequent operations. The input sequence for the model includes both the trainable prompt embeddings and the token embeddings of the given question Q:

$$\mathbf{H}_{in} = [\mathbf{e}(P_t);\mathbf{e}([CLS]);\mathbf{e}(Q)] \tag{1}$$

where $\mathbf{e}(P_t)$ is the trainable prefix embeddings (prompt tokens added before the input text), $\mathbf{e}(Q)$ denotes the token embeddings of the input question Q, and the specific token $[CLS]$ is used as classification.

The extended input $\mathbf{H}_{in}$ is passed through the LM to compute hidden representations:

$$\mathbf{H}_{out} = Model(\mathbf{H}_{in}) \tag{2}$$

here, $\mathbf{H}_{out} \in \mathbb{R}^{(d+l)\times m}$, where We denote d as the embedding dimension of the foundation LM, l denotes the length of trainable prompt P_t and m is the hidden dimension of the LM.

In this module, the total number of trainable parameters can be calculated as $\Theta(d \cdot h \cdot l)$, where h as the number of layers within the LM.

Comparing to full-parameter fine-tuning, Type Prompter module that employing p-tuning v2 decreases the number of training parameters while effectively capturing type-specific information. Additionally, it allows for the transfer updated weights of P_t, encompassing type-specific information, to the Unified Prompter module. Furthermore, by utilizing p-tuning v2, a broader feature spectrum can be efficiently captured and learned compared to the prompt-tuning [29].

3.4 Knowledge prompter

The Knowledge Prompter leverages textual input to activate and integrate LMs' innate prior knowledge, thereby enhancing the fusion of explicit and implicit information for effective reading comprehension.

[2] The foundational LMs could be displace with more advanced models. Consistent with previous studies [9] on prompt-based learning, we selected ELECTRA.

Figure 2 illustrates how the Knowledge Prompter employs an iterative encoder-decoder LM to retrieve implicit knowledge through prefix tuning[3] [32]. Trainable prompt embeddings, labeled as P_k^E for the encoder and P_k^D for the decoder, are integrated within each LM layer. This method facilitates the efficient retrieval and application of explicit knowledge during the iterative phases of encoding and decoding.

Given a multi-hop query Q and a serious of supporting sentences $S_n = [s_1, s_2, ..., s_i, ..., s_n]$, we aim to retrieve and extract a sequence of knowledge $K_n = [k_1, k_2, .., k_i, ..., k_n]$ that provides sufficient information for determining the response to both Q and S_n, where n represents the number of supporting sentences. Our focus lies in the development of prompt-based learning method, where we intend to construct trainble prompts P_k^E and P_k^D to lead the encoder-decoder LM in recalling the desired knowledge K_n. Notably, we maintain fixed parameters for the encoder-decoder LM, thereby allowing us to direct its retrieval process via trainable prompts.

Motivated by the sequential nature observed in multi-step reasoning tasks [53], we adopt an iterative approach as below:

$$P(k_j|Q, S_j, K_{j-1}) = \prod_{j=1}^{n} P(k_j|Q, s_1, ...,s_j, k_1, ..., k_{j-1}) \tag{3}$$

$$decoder(k_j) = encoder(Q, S_j, K_{j-1}) \tag{4}$$

where at each step j, LM recalls the next piece of knowledge k_j conditioned on the query Q and supporting sentences $s_1, ..., s_j$ and gathered knowledge $k_1, ..., k_{j-1}$.

More specially, when $j = 1$, it is written as following based on Equation (3) and (4):

$$decoder(k_1) = encoder(Q, s_1) \tag{5}$$

3.5 Unified prompter

As illustrated in Fig. 2, we suture the Unified Prompter module with P_t, integrating weights tailored for specific reasoning types. Additionally, the implicit knowledge K_n derived from the Knowledge Prompter module serves as supplementary input. This fusion of information intuitively boost reasoning capabilities of PEI framework based on human reading process.

In Unified Prompter module, the trainable prompt embeddings are denoted as P_u, where the updated prompt embeddings P_t are frozen. To preserve the learned weights of P_t derived from the Type Prompter, we adopt the identical architecture of the foundation LM, allowing for seamless concatenation of P_t with the Unified Prompter.

The input sequence for the Unified Prompter includes the trainable prompt embeddings P_u and the frozen prompt embeddings P_t:

$$\mathbf{H} = [\mathbf{e}(P_u);\mathbf{e}(Q, S_n, K_n);\mathbf{e}(P_t)] \tag{6}$$

Subsequently, we perform two training settings depending on various foundation LM. For ELECTRA, we employ joint p-tuning and full parameter fine-tuning to this module. For Llama, we use p-tuning to optimize the trainable P_u and freeze the foundation LM.

Prediction Module.[4] After encoding in Unified Prompter, we design a prediction module to jointly perform answer and supporting evidence prediction, followed by [9, 13]. To determine the answer span, two linear layers are utilized on the context representation to ascertain the start and end positions of the response. Meanwhile, a binary linear layer is deployed for predicting supporting evidence by assigning a binary relevance label at the beginning of each supporting sentence [SE].

4 Experiments

4.1 Dataset and metrics

HotpotQA [63] comprises a dataset of 113,000 question-answer pairs sourced from Wikipedia. Furthermore, HotpotQA includes sentence-level supporting facts critical for reasoning, thereby enabling QA systems to carry out inference with strong supervision and articulate their predictions.

2WikiMultiHopQA [18], comprises more than 192,000 entries, distributed across 167,000 for training, 12,500 for evaluation, and 12,500 for testing. While its structure is largely aligned with HotpotQA [63], this dataset introduces improvements by offering a wider spectrum of reasoning categories for questions and detailed annotations of the evidence trajectories linked to each question.

MuSiQue [50] consists of 25,000 examples featuring questions that require 2 to 4 reasoning steps. This collection is curated through a structured method of selecting compatible single-hop question pairs that manifest logical links, thereby crafting a comprehensive set of multi-hop inquiries.

Sub-question QA dataset [49] was developed to support the examination of multi-hop QA models' reasoning abilities at each phase of the reasoning process. To assess the models' effectiveness, the authors assembled a dedicated dataset composed of single-hop sub-questions. This collection

[3] We employ prefix tuning method for the Knowledge Prompter inspired by the context-aware prompter design [53].

[4] https://github.com/Tswings/PCL.

encompasses 1k samples, manually validated from the HotpotQA development set, thereby providing a high-quality evaluation resource for the research.

To achieve uniformity and comparability among the datasets employed in our experimental analyses, we classify the question types from the three datasets being examined into the overarching categories of comparison and bridge. This classification aids in establishing a standardized methodology for managing various question structures across the datasets subjected to our evaluation.

MEDHOP[58] is a benchmark resource designed for multi-hop reasoning in the biomedical domain. It belongs to the QAngaroo collection,[5] which emphasizes the integration of data from multiple documents or textual units to resolve intricate queries. The dataset comprises 1,620 training instances, 342 validation instances, and 546 test instances, culminating in a total of 2,508 instances. However, to the best of our knowledge, only MEDHOP evaluates multi-hop reasoning abilities, while almost all other Biomedical QA datasets focus on single-hop reasoning[25].

Metrics. Consistent with prior research [63], We report Exact Match (*EM*) and Partial Match ($F1$) to evaluate the efficacy and performance of our proposed framework concerning both answer and supporting facts prediction. Furthermore, the joint *EM* and $F1$ are used to assess the overall performance. Specially, we use accuracy to evaluate the performance of our PEI and baselines on MEDHOP following by Welbl et al [58].

4.2 Selected baselines

To comprehensively assess the performance of PEI, we compare with a series of selected and state-of-art baselines, including 1) general domain methods on HotpotQA, 2WikiMultiHopQA, MuSiQue and Sub-question QA, and 2) biomedical domain methods that are conducted on MEDHOP dataset.

These baselines for general multi-hop QA as follows:

- Baseline Model [63] serves as the initial baseline for HotpotQA.
- DecompRC [37] transforms complex queries into easier sub-questions, enabling resolution via existing single-hop reading comprehension frameworks.
- OUNS [41] represents an algorithm designed for One-to-N Unsupervised Sequence transduction. It converts intricate, multi-step queries into a range of straightforward, single-step questions to enhance the QA process by decomposing complexities.
- QFE [39] model incrementally identifies evidence sentences through an RNN with attention focused on the query, drawing inspiration from models used in extractive summarization.
- Longformer [3] employs an attention mechanism that increases linearly with the sequence length, facilitating the examination of extended texts in a multi-hop QA context.
- Beam Retrieval [66] integrates the retrieval operation in an end-to-end manner by synchronously optimizing an encoder and two classification outputs throughout all steps.

The following baselines apply GNN-based methods for general multi-hop QA:

- DFGN [43] dynamically constructs an entity graph from the text and incrementally identifies supporting entities relevant to the query within the provided documents.
- SAE-large [52] utilizes a GNN where contextual sentence embeddings serve as nodes, bypassing the use of entities as nodes, thereby directly predicting supporting sentences alongside the answer.
- C2F Reader [47] integrates task-specific prior knowledge via graph structures and adjacency matrices, employing graph attention as a variant of self-attention.
- HGN [13] introduces a hierarchical graph with nodes representing varying granularities, such as questions, paragraphs, sentences, and entities, leveraging pre-trained contextual encoders for initialization.
- AMGN [31] incorporates GNN-based methodologies to asynchronously update multi-grained nodes by modeling relationships across different levels, reflecting the logical progression of multi-hop reasoning.
- S2G [60] implements a select-to-guide (S2G) strategy to retrieve evidence paragraphs in a coarse-to-fine manner, augmented by two novel attention mechanisms, which effectively align with the inherent nature of multi-hop reasoning.

These baselines utilize CoT prompting methods for general multi-hop QA as follows:

- iCAP [53] adopts an iterative prompting framework designed to incrementally extract pertinent knowledge from pre-trained language models (PLMs), enabling step-by-step inference.
- PCL [9] introduces a Prompt-based Conservation Learning (PCL) framework, wherein soft prompts are optimized to guide sub-networks in executing type-specific reasoning tasks.

The biomedical domain baselines are as follows:

[5] http://qangaroo.cs.ucl.ac.uk.

- FastQA [57] employs a single bi-directional recurrent neural network followed by an answer prediction layer that independently identifies the start and end points of the answer span.
- BiDAF [46] adopts a hierarchical approach, representing contextual information at varying levels of granularity and leveraging a bidirectional attention flow mechanism to construct a query-aware context representation without premature summarization.
- Document-cue [58] emphasizes the model's ability to leverage document-answer co-occurrence patterns to identify relevant information.
- BAG [6] introduces a bidirectional attention entity graph convolutional network that captures relationships between graph nodes and employs attention mechanisms to link the query with the entity graph.
- EPAr [23] (Explore-Propose-Assemble reader) mimics human-like reading strategies by adopting a coarse-to-fine approach for reasoning and answering QA tasks.
- NLProlog [55] integrates symbolic reasoning and rule learning with distributed representations of sentences and entities to perform rule-based multi-hop reasoning over natural language inputs.
- DrKIT [10] simulates traversals in a knowledge base (KB) constructed over a text corpus, providing ability to follow relations in the "virtual" KB over text for multi-hop questions.
- DILR-BERT [54] combines transformer-based model with inductive logic reasoning, first extracting query-relevant data and then applying rule induction to conduct logical reasoning across the filtered information.
- ClueReader [14] employs a heterogeneous graph attention network inspired by the grandmother cell concept, aggregating semantic features at multiple levels and dynamically focusing or suppressing information for reasoning.
- MedKGQA [15] integrates external biomedical knowledge bases, including "drug-protein" triplets, with graph neural networks (GNNs) to facilitate effective navigation and answer retrieval within biomedical pathways, particularly for drug-drug interaction (DDI) prediction and medical reasoning.
- AEGCN [11] enhances entity graph representations with enriched edge relationships derived from supporting texts, employing adversarial training to improve resistance and interference.

4.3 Implementation details

ELECTRA-large [7] and Llama 3.1 8B [12] serve as the foundation LM for both Type Prompter and Unified Prompter modules. In Type Prompter, p-tuning v2 [34] is used for the prompt tuning to acquiring the weights of specific-type information of the given questions. In Unified Prompter, we conduct two training settings depends on various foundation LM: 1) for ELECTRA, we employ joint p-tuning and full parameter fine-tuning to the whole module, and 2) for Llama, we use p-tuning to optimize the trainable Pu and freeze the Llama. Inspired by the studies of Wang et al [53], we adopt BART-large [30] as the foundation LM in the Knowledge Prompter module. Specially, BioELECTRA [26] and BioBART [65] are employed as the foundation LMs for biomedical domain extension.

Our implementation is built upon the Huggingface platform [59]. For model optimization, we employ the AdamW optimizer [35] along with a linear learning rate scheduler with a warmup ratio of 0.05.

In terms of hyperparameters, we conduct a search for the optimal batch size. We explored batch sizes of $\{4, 8, 12, 16, 32\}$ respectively. Additionally, we performed a tuning process for the learning rate, considering values from $\{2e-5, 4e-5, 8e-5, 2e-4, 4e-4, 8e-4, 2e-3, 4e-3, 8e-3, 2e-2, 4e-2, 8e-2\}$. Moreover, we conducted tuning experiments for the length of the encoder/decoder prompts P_k type prompts P_t and the unified prompt P_u, exploring values from $\{20, 40, 60, 80, 100, 120, 150\}$.

4.4 Main results and analysis

Initially, we evaluate PEI framework on the test set of HotpotQA in the distractor setting comparing with peer-reviewed baselines, including the baseline model of HotpotQA [63], Beam Retrieval [66] that is the state-of-art model on the leaderboard, iCAP [53] and PCL [9] which we inspired by, and other baselines. For a reminder, we will refer to the PEI(ELECTRA$_{PT+FT}$) framework that employs ELECTRA with full parameter fine-tuning and prompt tuning as PEI to be more brief.

As depicted in Table 1, PEI framework outperforms all baseline models across the evaluated metrics, with the sole exception of Beam Retrieval. Notably, PEI achieves performance comparable to that of Beam Retrieval [66] on the HotpotQA, demonstrating the substantial advancements facilitated by PEI in addressing multi-hop QA.

More specifically, comparing with Beam Retrieval, PEI achieves an improvement of 0.20/0.28/0.30 in answer EM, answer F1 score and joint F1 score, respectively. In contrast, Beam Retrieval exhibits superior results with an improvement of 1.22 in supporting EM, 0.28 in supporting F1 score, and 0.62 in joint EM compared to PEI.

The difference in performance between PEI and Beam Retrieval in answer prediction versus supporting prediction could be attributed to the distinct methodologies employed by the two approaches. Beam Retrieval preserves multiple partial hypotheses of relevant passages at each step,

Table 1 Results on the blind test set of HotpotQA in the distractor setting. “-” denotes the case where no results are available. † denotes that we implement the codes. Other results are derived from [9] and [66]. “Ans” represents the metrics for answer; “Sup” denotes the metrics for supporting facts; “Joint” is the joint metrics that combine the evaluation of answer spans and supporting facts. “FT” is full fine-tuning and “PT” is the prompt tuning

Models	Ans		Sup		Joint	
	EM	F1	EM	F1	EM	F1
Baseline Model [63]	45.60	59.02	20.32	64.49	10.83	40.16
DecompRC [37]	55.20	69.63	–	–	–	–
OUNS [41]	66.33	79.34	–	–	–	–
QFE [39]	53.86	68.06	57.75	84.49	34.63	59.61
DFGN [43]	56.31	69.69	51.50	81.62	33.62	59.82
SAE-large [52]	66.92	66.92	61.53	86.86	45.36	71.45
C2F Reader [47]	67.98	81.24	60.81	87.63	44.67	72.73
Longformer [3]	68.00	81.25	63.09	88.34	45.91	73.16
HGN [13]	69.22	82.19	62.76	88.47	47.11	74.21
AMGN [31]	70.53	83.37	63.57	88.83	47.77	75.24
S2G [60]	70.72	83.53	64.30	88.72	48.60	75.45
iCAP † [53]	68.61	81.82	62.80	88.51	47.02	74.11
PCL [9]	71.76	84.39	64.61	89.20	49.27	76.56
Beam Retrieval [66]	72.69	85.04	**66.25**	**90.09**	**50.53**	77.54
PEI(ELECTRA$_{PT+FT}$)	**72.89**	**85.32**	65.03	89.81	49.91	**77.84**
PEI(Llama$_{PT}$)	71.45	85.05	64.00	88.97	48.89	76.41

expanding the search space (albeit at the expense of an exponentially complex retrieval process) and reducing the risk of missing relevant passages. Consequently, it excels in supporting prediction. On the other hand, PEI draws inspiration from human reading processes by integrating implicit knowledge and type-specific information, which enhances its accuracy in answer prediction. However, this design may limit its efficacy in supporting prediction compared to Beam Retrieval, owing to the differences in their retrieval strategies.

Though PCL and PEI adopts the same backbone LM (i.e., ELECTRA) and prediction module, PEI framework demonstrates a significant improvement of 0.64/1.28 in the Joint EM/F1 score compare to PCL. Compare with the question classification that PCL trains a PLM to acquired the reasoning type, PEI leverages the prompt tuning to learn the type-specific knowledge and transfers the trained weight to the Unifier Prompter, which effectively reduces computational cost.

Moreover, the proposed PEI framework demonstrates a notable improvement over iCAP, achieving a 2.89/3.73 increase in joint EM/F1 scores, despite both models utilizing the same encoder-decoder architecture (BART) as their foundational LMs. When compared to the graph-based AMGN model, PEI achieves even greater gains, with a 2.14/2.6 enhancement in joint EM/F1 scores.

4.4.1 Comparison on LM architecture and training paradigms

In Table 1, the performance of PEI(Llama$_{PT}$) framework based on Llama with prompt tuning is slightly lower that that of the standard PEI in both answer and supporting facts prediction. Because Unified Prompter actually acts as a context encoder, the performance of PEI depends on the architecture of foundation LMs and training paradigms. Extractive multi-hop QA (e.g., HotpotQA) often requires token-level attention over the input representation to have an edge in span-based extraction, which is a native strength of encoder-only architectures such as ELECTRA [44]. In contrast, decoder-only models (such as Llama) are not inherently designed to output spans of text, instead, they excel at generative QA tasks, where their ability to synthesize and produce text is advantageous [38].

Compared to standard PEI, which employs full parameter fine-tuning, the Llama-based PEI utilizing prompt tuning significantly reduces the number of trainable parameters while maintaining comparable reasoning performance. However, it still incurs higher computational cost and longer inference time due to the large-scale parameters of Llama and the associated memory requirements.

4.4.2 Comparison with LLMs under zero-shot

Table 2 presents the performance of open-source LLMs in a zero-shot setting for multi-hop QA. The results clearly indicate that zero-shot LLMs perform significantly inferior than PEI and other baseline models (see Tables 1 and 3) across the HotpotQA, 2WikiMultihopQA, and MuSiQue datasets. This demonstrates that extractive multi-hop QA still remains a challenge to LLMs in zero-shot setting. Notably, Llama-based PEI surpass all LLM baselines, including Llama 3.1-8B, across diverse datasets. Compare to zero-shot Llama 3.1-8B, the PEI (Llama 3.1-8B$_{PT}$) attains a

Table 2 Performance of Llama-based PEI compared to open source LLMs with a zero-shot setting on HotpotQA, 2WikiMultihopQA and MuSiQue. † denotes that is our implement using the prompt template of [61]. Other results come from [61]

Models	HoptpotQA		2WikiMultihopQA		MuSiQue	
	EM	F1	EM	F1	EM	F1
Mistral-7B	30.6	37.2	27.4	29.8	25.2	28.9
Qwen 2-7B	36.2	43.5	31.7	35.8	28.2	31.2
Llama 2-7B	34.5	41.3	30.6	34.7	31.7	35.6
Llama 3.1-8B †	39.4	45.9	33.1	37.5	32.9	36.5
PEI (Llama 3.1-8B$_{PT}$)	71.4	85.0	45.9	73.5	40.5	67.2

Table 3 Results of our proposed PEI compared to PCL, HGN and iCAP on 2WikiMultihopQA and MuSiQue multi-hop QA test set. "-" denotes the case where no results are available. PEI is version of PEI(ELECTRA$_{PT+FT}$)

Models	2WikiMultihopQA		MuSiQue	
	EM	F1	EM	F1
iCAP	42.80	47.90	–	–
HGN	38.74	68.69	39.42	65.12
PCL	46.03	73.42	41.28	67.34
PEI (Ours)	**47.32**	**74.56**	**41.97**	**67.85**

notable performance boost with EM and F1 improvements of 32.0/39.1 on HotpotQA and 12.8/36.0 on 2WikiMultihopQA, respectively. On MuSiQue, while the performance gains are less pronounced (an improvement of 7.6/30.7 in EM/F1 compared to Llama 3.1-8B), PEI still leads with an EM of 40.5 and F1 of 67.2, demonstrating consistent performance across diverse datasets. These findings demonstrate the effectiveness of leveraging task-specific prompt tuning to enhance the reasoning capabilities of LLMs. By enabling task-specific fine-tuning, PEI enhances the model's ability to fuse explicit and implicit knowledge effectively for complex reasoning tasks.

4.5 Evaluation of robustness

To further evaluate the robustness of proposed PEI framework, we conduct three aspects experiments: (1) assessing PEI on other multi-hop QA datasets; (2) evaluation on sub-question dataset in composing answers from solved sub-questions; and (3) effect of foundation LMs in the same training paradigm.

4.5.1 Evaluation on other multi-hop datasets

To evaluate generalization, we validate the PEI framework on the 2WikiMultihopQA and MuSiQue datasets. As presented in Table 3, PEI consistently outperforms all baseline models across both EM and F1 metrics. Notably, although both PEI and iCAP utilize the same encoder-decoder architecture (BART), PEI achieves a significant improvement of 4.52/26.66 in answer EM/F1 scores on the 2WikiMultihopQA dataset. Additionally, PEI demonstrates superior performance over PCL, with gains of 1.29/1.14 and 0.69/0.51 in answer EM/F1 scores on the 2WikiMultihopQA and MuSiQue, respectively.

4.5.2 Evaluation on sub-question dataset

To assess the efficacy of the PEI model in multi-hop reasoning, particularly in synthesizing answers from resolved sub-questions, we conduct an evaluation on the sub-question QA dataset [49]. Each parent question, denoted as q, is associated with two corresponding sub-questions, q_{sub1} and q_{sub2}. As presented in Table 4, the PEI model achieves a success rate of 97.62% in correctly answering the parent multi-hop question q while both sub-questions q_{sub1} and q_{sub2} are answered correctly.[6] This highlights the proficiency of PEI in retaining acquired knowledge through the integration of explicit and implicit knowledge, surpassing other baseline models. Interestingly, PEI also demonstrates a notable success rate of 36.55% in correctly answering the parent multi-hop question even when only one of the sub-questions is answered correctly.[7] Fig. 3 further illustrates the sub-question-dependent success rates for various multi-hop QA models, showing that these models frequently predict the correct parent question answer even when only one sub-question is answered correctly. This observation highlights a persistent challenge in multi-hop QA: models often exploit unreliable reasoning shortcuts for answer prediction, a phenomenon that deviates from the expected logical reasoning process [9].

4.5.3 Effect of foundation LMs

To evaluate the effects of foundation LMs, we conduct a comparative analysis of PEI against PCL and HGN under identical experimental conditions, including the same datasets and foundation models. As shown in Table 5, PEI consistently exceeds both PCL and HGN across all evaluation

[6] The calculation process: $49.2/(49.2 + 1.2) = 97.62\%$

[7] The calculation process: $(7.1 + 22.1)/(49.2 + 7.1 + 22.1 + 1.5) = 36.55\%$

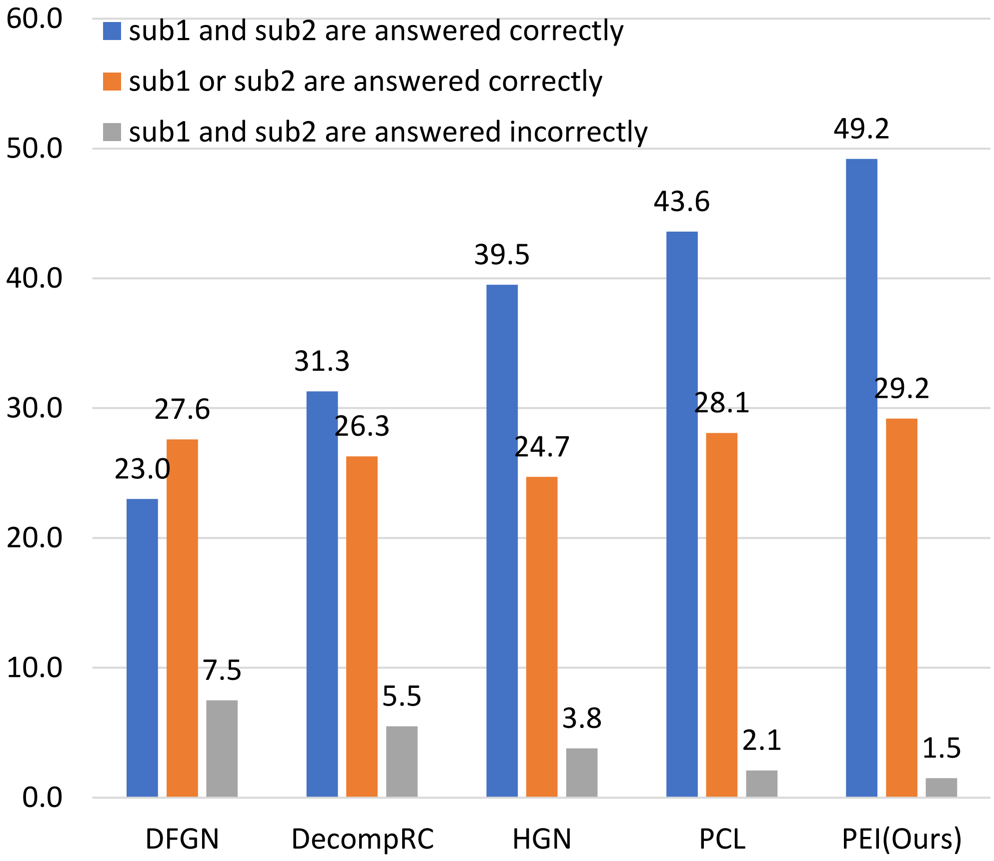

Fig. 3 The success rate (%) of five multi-hop QA models. *sub*1 denotes the first sub-question and *sub*2 is the second sub-question of corresponding question q. The results of DFGN, DecompRC, HGN and PCL are from [9]

Table 4 Results on sub-question dataset. *c*/*w* denotes that the question is answered correctly/wrongly. *sub*1 denotes the first sub-question and *sub*2 is the second sub-question of corresponding question q. The results of DFGN, DecompRC, HGN and PCL are derived from [9]

q	q_{sub1}	q_{sub2}	DFGN	DecompRC	HGN	PCL	PEI(Ours)
c	c	c	23.0	31.3	39.5	43.6	49.2
c	c	w	9.7	7.2	5.1	6.8	7.1
c	w	c	17.9	19.1	19.6	21.3	22.1
c	w	w	7.5	5.5	3.8	2.1	1.5
w	c	c	4.9	3.0	2.8	1.7	1.2
w	c	w	17.0	18.6	16.7	16.3	13.4
w	w	c	3.5	3.4	2.6	1.1	1.0
w	w	w	16.5	11.9	9.9	7.1	4.5

metrics, demonstrating its effectiveness and robustness across various foundation LMs. Furthermore, PEI that implemented with ALBERT achieves an improvement of 0.62/0.23 in Answer/Support F1 scores compared to its implementation with RoBERTa. This aligns with prior findings that ALBERT surpasses RoBERTa on the GLUE benchmark under a single-model configuration.[8] These results affirm that incorporating a more advanced foundation LM can significantly enhance the performance of the PEI framework.

4.6 Evaluation on biomedical inference

4.6.1 Results and discussion

To comprehensively evaluate PEI framework on biomedical domain, we compare PEI with diverse baselines including biomedical knowledge-based and non-biomedical knowledge-based models. As shown in Table 6, the vanilla PEI,

[8] https://github.com/google-research/albert.

Table 5 Results with different LMs on the development set of HotpotQA. The results of HGN and PCL are derived from [9]

Model	Ans F1	Sup F1	Joint F1
HGN (RoBERTa)	82.22	88.58	74.37
HGN (ELECTRA)	82.24	88.63	74.51
HGN (ALBERT)	83.46	89.20	75.79
PCL (RoBERTa)	84.33	90.75	77.12
PCL (ELECTRA)	84.42	91.15	77.76
PCL (ALBERT)	85.47	91.28	78.76
PEI (RoBERTa)	85.61	92.02	78.95
PEI (ELECTRA)	85.68	92.11	79.02
PEI (ALBERT)	**86.23**	**92.25**	**79.11**
PEI (Llama$_{PT}$)	85.05	88.97	76.41

which employs BART and ELECTRA foundation models without biomedical pretraining, exceeds all the baselines on MEDHOP dataset and outperforms the state-of-art model, AEGCN [11] with 0.75% in accuracy. This shows that even without using a biomedical LMs, the PEI framework exhibits its significant reasoning ability in biomedical multi-hop QA.

Although general capabilities of foundation models have become increasingly apparent, there remain open questions regarding whether exceptional performance can be achieved on specialized tasks, such as those in the medical field, without extensive domain-specific training or fine-tuning of these general models [22, 40]. Much of the research on the application of foundation models in biomedicine has been heavily reliant on fine-tuning tailored to specific domains and tasks. With the advent of first-generation foundation models, the benefits of domain-specific pretraining became evident, as demonstrated by widely used models in the biomedical domain, such as PubMedBERT [16] and BioGPT [36]. In Table 6, the experimental results show that the biomedical PEI, which employs both BioBART and BioELECTRA as fundamental models to make fully use of pretrained biomedical domain knowledge, achieve a significant 3.34% improvement in accuracy compared with AEGCN. Compared with the 0.75% accuracy improvement of vanilla PEI, it is obvious that biomedical PEI has better performance and stronger reasoning ability in the biomedical domain.

Table 6 Results on testset of MEDHOP. PEI (vanilla) denotes that PEI model employs BART and ELECTRA as foundation models without biomedical pre-trained knowledge. PEI (biomedical) denotes that bioBART and bioELECTRA serve as foundation models for PEI, which inject pre-trained biomedical knowledge

Model	Accuracy
FastQA [57]	23.10
BiDAF [46]	47.80
Document-cue [58]	44.90
BAG [6]	64.50
EPAr [23]	64.90
NLProlog [55]	65.78
DrKIT [10]	67.25
DILR-BERT [54]	71.35
ClueReader [14]	46.00
MedKGQA [15]	64.80
AEGCN [11]	<u>72.28</u>
PEI (Vanilla)	73.03 $\uparrow_{0.75}$
PEI (Biomedical)	**75.62** $\uparrow_{3.34}$

4.6.2 Effect of biomedical knowledge

To further analyze the effect of specialty knowledge of PEI framework for biomedical domain, we explore the various settings of foundation models (see Table 7). As illustrated in Table 7, the experimental results show that biomedical knowledge boost performance of PEI. The PEI with bioBART shows an improvement of 0.96% in accuracy compared with the vanilla PEI, demonstrating the Knowledge Prompter module is able to elicit the implicit specialty knowledge via CoT prompting. Moreover, the Unified Prompter makes fully use of the implicit knowledge $K_n = [k_1, k_2, .., k_i, ..., k_n]$ and fuse with explicit knowledge $S_n = [s_1, s_2, ..., s_i, ..., s_n]$ to enhance the inference ability. These results provide evidence that implicit knowledge plays a crucial role in improving the model's reasoning capabilities, thereby comfirming the hypothesis that underpins our proposed PEI framework, which is inspired by human reading process.

Additionally, the PEI with BioELECTRA outperforms the vanilla PEI with an improvement of 1.27% in accuracy, further emphasizing the reliance of most investigations into the efficacy of foundation models in biomedical tasks on extensive domain- and task-specific fine-tuning.

4.7 Ablation studies

To investigate the contributions of each components within PEI framework, we perform a series of ablation studies on the validation set of HotpotQA.

Table 7 Ablation study of comparison on vanilla PEI and biomedical PEI with different components

Components		Accuracy
BioBART	Bio-ELEC-TRA	
✗	✗	73.03
✓	✗	73.99 $\uparrow_{0.96}$
✗	✓	74.30 $\uparrow_{1.27}$
✓	✓	**75.62** $\uparrow_{2.59}$

Table 8 Ablation Study of PEI on the development set of HotpotQA. Ans F1 stands for answer F1; Sup F1 is supporting F1. The "Implicit Knowledge" specifically refers to the implicit knowledge K_n from Knowledge Prompter

Model	Ans F1	Sup F1	Joint F1
ELECTRA	78.12	88.20	73.50
- Type Prompter	81.14 $\uparrow_{3.02}$	89.37 $\uparrow_{1.17}$	75.68 $\uparrow_{2.18}$
- Pre-trained	78.82 $\uparrow_{0.70}$	88.82 $\uparrow_{0.62}$	74.54 $\uparrow_{1.04}$
- Implicit Knowledge	81.22 $\uparrow_{3.10}$	90.25 $\uparrow_{2.05}$	75.80 $\uparrow_{2.30}$
PEI	**85.68** $\uparrow_{7.56}$	**92.11** $\uparrow_{3.91}$	**79.02** $\uparrow_{5.52}$

4.7.1 Effect of implicit knowledge

To validate the hypothesis that implicit knowledge enhances reasoning capabilities in multi-hop QA, we conduct an ablation study comparing the ELECTRA model with and without implicit knowledge integration (specifically referring to the implicit knowledge K_n derived from the Knowledge Prompter). Table 8 demonstrates that incorporating implicit knowledge yielded significant improvements of 3.10/2.05/2.30 in Answer F1, Supporting F1, and Joint F1 scores, respectively, compared to the baseline model without implicit knowledge. These findings empirically confirm that implicit knowledge substantially augments models' reasoning abilities, aligning with the cognitive theories underpinning the PEI framework, which draws inspiration from the human reading process.

4.7.2 Effect of type prompts

To assess the influence of type-specific prompts and the model's capacity for type-driven reasoning in multi-hop question answering, we perform a comparative analysis of the ELECTRA language model, both with and without the Type Prompter module. As presented in Table 8, integrating the Type Prompter with the language model results in significant improvements of 3.02/1.17/2.18 in Answer F1, Supporting F1, and Joint F1 scores, respectively, over the model without Type Prompter. These results highlight the effectiveness of incorporating question type information through the Type Prompter in enhancing the model's overall performance and facilitating type-specific reasoning. Furthermore, the findings support the alignment of PEI framework architecture with cognitive process observed in human reasoning, as type information can be regarded as a form of implicit knowledge.

4.7.3 Effect of pre-training on single-hop

We initially trained an ELECTRA-based QA model on the single-hop QA dataset, SQuAD [45], and subsequently fine-tuned it on the HotpotQA dataset. While conservation learning [9] is not utilized in our approach, we assess the model's performance both with and without pre-training to evaluate its impact on single-hop QA. As indicated in Table 8, incorporating pre-training resulted in improvements of 0.70/0.62/1.04 in Answer F1, Supporting F1, and Joint F1 scores, respectively, compared to the model without pre-training. These results suggest that pre-training in the single-hop QA task helps the model capture valuable information, thereby boosting its performance. However, it is important to note that the observed improvements are relatively modest in the absence of conservation learning.

5 Conclusions and future work

In this study, we present a novel framework inspired by human cognitive theories, which utilizes prompts to connect explicit and implicit knowledge. Our approach incorporates chain-of-thought prompts to extract implicit knowledge from LMs within the given input context, while also integrating question type information to improve the overall performance of the model. Moreover, we propose two training paradigms for PEI framework, and extend PEI on biomedical domain QA to further explore the fusion and relation of explicit and implicit biomedical knowledge and analyze the consistency of the domain knowledge fusion. Experimental results show that PEI performs comparably to the state-of-the-art on HotpotQA, and excels all baselines on MEDHOP.

Additionally, ablation studies affirm the efficacy and resilience of PEI framework in mirroring human reading comprehension. Moving forward, we intend to expand and apply cognitive theories of human reading to a wider range of reasoning tasks, with the goal of fostering more advanced and intricate reasoning capabilities.

Acknowledgements This work is supported by the Alan Turning Institute/DSO grant: *Improving multimodality misinformation detection with affective analysis*. Yunfei Long, Guangming Huang and Cunjin Luo acknowledge the financial support of the School of Computer science and Electrical Engineering, University of Essex.

Author Contributions **Guangming Huang:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization. **Yunfei Long and Cunjin Luo:** Writing - Review & Editing, Supervision, Project administration, Funding acquisition.

Funding This study was funded by the Alan Turning Institute/DSO grant:*Improving multimodality misinformation detection with affective analysis.*

Data Availability Data will be made available on request.

Declarations

Conflict of interest **Non-financial interests:** none.

References

1. Abdelaal NM, Sase AS (2014) Relationship between prior knowledge and reading comprehension. Adv Lang Literary Stud 5(6):125–131
2. Baldwin RS, Peleg-Bruckner Z, McClintock AH (1985) Effects of topic interest and prior knowledge on reading comprehension. Read Res Quart pp 497–504
3. Beltagy I, Peters ME, Cohan A (2020) Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150
4. Bender EM, Gebru T, McMillan-Major A, et al (2021) On the dangers of stochastic parrots: Can language models be too big? In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, pp 610–623
5. Biancofiore GM, Deldjoo Y, Noia TD et al (2024) Interactive question answering systems: Literature review. ACM Comput Surv 56(9):1–38
6. Cao Y, Fang M, Tao D (2019) Bag: Bi-directional attention entity graph convolutional network for multi-hop reasoning question answering. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp 357–362
7. Clark K, Luong MT, Le QV, et al (2020) Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555
8. Clarke MA, Silberstein S (1977) Toward a realization of psycholinguistic principles in the esl reading class 1. Lang Learn 27(1):135–154
9. Deng Z, Zhu Y, Chen Y, et al (2022) Prompt-based conservation learning for multi-hop question answering. In: Proceedings of the 29th International Conference on Computational Linguistics, pp 1791–1800
10. Dhingra B, Zaheer M, Balachandran V, et al (2020) Differentiable reasoning over a virtual knowledge base. In: International Conference on Learning Representations, https://openreview.net/forum?id=SJxstlHFPH
11. Du Y, Yan R, Hou Y et al (2024) Adversarial entity graph convolutional networks for multi-hop inference question answering. Expert Syst Appl 258:125098
12. Dubey A, Jauhri A, Pandey A, et al (2024) The llama 3 herd of models. arXiv preprint arXiv:2407.21783
13. Fang Y, Sun S, Gan Z, et al (2020) Hierarchical graph network for multi-hop question answering. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp 8823–8838
14. Gao P, Gao F, Wang P et al (2023) Cluereader: Heterogeneous graph attention network for multi-hop machine reading comprehension. Electronics 12(14):3183
15. Gao P, Gao F, Ni JC, et al (2024) Medical knowledge graph question answering for drug-drug interaction prediction based on multi-hop machine reading comprehension. CAAI Trans Intell Technol
16. Gu Y, Tinn R, Cheng H et al (2021) Domain-specific language model pretraining for biomedical natural language processing. ACM Trans Comput Healthcare (HEALTH) 3(1):1–23
17. Hagoort P, Hald L, Bastiaansen M et al (2004) Integration of word meaning and world knowledge in language comprehension. Science 304(5669):438–441
18. Ho X, Nguyen AKD, Sugawara S, et al (2020) Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. In: Proceedings of the 28th International Conference on Computational Linguistics, pp 6609–6625
19. Huang G, Long Y, Luo C, et al (2023) LIDA: Lexical-based imbalanced data augmentation for content moderation. In: Huang CR, Harada Y, Kim JB, et al (eds) Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation. Association for Computational Linguistics, Hong Kong, China, pp 59–69, https://aclanthology.org/2023.paclic-1.6
20. Huang G, Li Y, Jameel S, et al (2024a) From explainable to interpretable deep learning for natural language processing in healthcare: How far from reality? Comput Struct Biotechnol J
21. Huang G, Long Y, Luo C, et al (2024b) Prompting explicit and implicit knowledge for multi-hop question answering based on human reading process. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pp 13179–13189
22. Huang G, Long Y, Luo C (2025) Similarity-dissimilarity loss for multi-label supervised contrastive learning. https://arxiv.org/abs/2410.13439, arXiv:2410.13439
23. Jiang Y, Joshi N, Chen YC, et al (2019) Explore, propose, and assemble: An interpretable model for multi-hop reading comprehension. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp 2714–2725
24. Jin C, Rinard M (2023) Evidence of meaning in language models trained on programs. arXiv preprint arXiv:2305.11169
25. Jin Q, Yuan Z, Xiong G et al (2022) Biomedical question answering: a survey of approaches and challenges. ACM Comput Surv (CSUR) 55(2):1–36
26. Kanakarajan KR, Kundumani B, Sankarasubbu M (2021) Bioelectra: pretrained biomedical text encoder using discriminators. In: Proceedings of the 20th workshop on biomedical language processing, pp 143–154
27. Khalifa M, Logeswaran L, Lee M, et al (2023) Few-shot reranking for multi-hop qa via language model prompting. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp 15882–15897
28. Kojima T, Gu SS, Reid M, et al (2022) Large language models are zero-shot reasoners. arXiv preprint arXiv:2205.11916
29. Lester B, Al-Rfou R, Constant N (2021) The power of scale for parameter-efficient prompt tuning. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp 3045–3059
30. Lewis M, Liu Y, Goyal N, et al (2020) Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp 7871–7880
31. Li R, Wang L, Wang S, et al (2021) Asynchronous multi-grained graph network for interpretable multi-hop reading comprehension. In: IJCAI, pp 3857–3863
32. Li XL, Liang P (2021) Prefix-tuning: Optimizing continuous prompts for generation. In: 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), pp 4582–4597

33. Lin Z, Chen W, Song Y et al (2024) Prompting few-shot multi-hop question generation via comprehending type-aware semantics. Find Assoc Comput Linguist: NAACL 2024:3730–3740
34. Liu X, Ji K, Fu Y, et al (2021) P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. arXiv preprint arXiv:2110.07602
35. Loshchilov I, Hutter F (2018) Decoupled weight decay regularization. In: International Conference on Learning Representations
36. Luo R, Sun L, Xia Y et al (2022) Biogpt: generative pre-trained transformer for biomedical text generation and mining. Brief Bioinform 23(6):409
37. Min S, Zhong V, Zettlemoyer L, et al (2019) Multi-hop reading comprehension through question decomposition and rescoring. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp 6097–6109
38. Minaee S, Mikolov T, Nikzad N, et al (2024) Large language models: A survey. arXiv preprint arXiv:2402.06196
39. Nishida K, Nishida K, Nagata M, et al (2019) Answering while summarizing: Multi-task learning for multi-hop qa with evidence extraction. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp 2335–2345
40. Nori H, Lee YT, Zhang S, et al (2023) Can generalist foundation models outcompete special-purpose tuning? case study in medicine. arXiv preprint arXiv:2311.16452
41. Perez E, Lewis P, Yih Wt, et al (2020) Unsupervised question decomposition for question answering. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp 8864–8880
42. Plaat A, Wong A, Verberne S, et al (2024) Reasoning with large language models, a survey. arXiv preprint arXiv:2407.11511
43. Qiu L, Xiao Y, Qu Y, et al (2019) Dynamically fused graph network for multi-hop reasoning. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp 6140–6150
44. Qorib M, Moon G, Ng HT (2024) Are decoder-only language models better than encoder-only language models in understanding word meaning? Findings of the Association for Computational Linguistics ACL 2024:16339–16347
45. Rajpurkar P, Zhang J, Lopyrev K, et al (2016) Squad: 100,000+ questions for machine comprehension of text. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp 2383–2392
46. Seo M, Kembhavi A, Farhadi A, et al (2017) Bidirectional attention flow for machine comprehension. In: International Conference on Learning Representations, https://openreview.net/forum?id=HJ0UKP9ge
47. Shao N, Cui Y, Liu T, et al (2020) Is graph structure necessary for multi-hop question answering? In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp 7187–7192
48. Smith F (1971) Understanding reading: A psycholinguistic analysis of reading and learning to read. Holt, Rinehart and Winston, New York
49. Tang Y, Ng HT, Tung A (2021) Do multi-hop question answering systems know how to answer the single-hop sub-questions? In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pp 3244–3249
50. Trivedi H, Balasubramanian N, Khot T et al (2022) Musique: Multihop questions via single-hop question composition. Trans Assoc Comput Linguist 10:539–554
51. Trivedi H, Balasubramanian N, Khot T, et al (2023) Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp 10014–10037
52. Tu M, Huang K, Wang G, et al (2020) Select, answer and explain: Interpretable multi-hop reading comprehension over multiple documents. In: Proceedings of the AAAI conference on artificial intelligence, pp 9073–9080
53. Wang B, Deng X, Sun H (2022) Iteratively prompt pre-trained language models for chain of thought. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp 2714–2730
54. Wang W, Pan S (2022) Deep inductive logic reasoning for multi-hop reading comprehension. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp 4999–5009
55. Weber L, Minervini P, Münchmeyer J, et al (2019) Nlprolog: Reasoning with weak unification for question answering in natural language. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp 6151–6161
56. Wei J, Wang X, Schuurmans D, et al (2022) Chain of thought prompting elicits reasoning in large language models. arXiv preprint arXiv:2201.11903
57. Weissenborn D, Wiese G, Seiffe L (2017) Making neural qa as simple as possible but not simpler. In: Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), pp 271–280
58. Welbl J, Stenetorp P, Riedel S (2018) Constructing datasets for multi-hop reading comprehension across documents. Trans Assoc Comput Linguist 6:287–302
59. Wolf T, Debut L, Sanh V, et al (2020) Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations, pp 38–45
60. Wu B, Zhang Z, Zhao H (2021) Graph-free multi-hop reading comprehension: A select-to-guide strategy. arXiv preprint arXiv:2107.11823
61. Wu J, Yang L, Okumura M, et al (2024a) Cofca: A step-wise counterfactual multi-hop qa benchmark. arXiv preprint arXiv:2402.11924
62. Wu J, Yu T, Wang R, et al (2024b) Infoprompt: Information-theoretic soft prompt tuning for natural language understanding. Advances in Neural Information Processing Systems 36
63. Yang Z, Qi P, Zhang S, et al (2018) Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp 2369–2380
64. Yin KM (1985) The role of prior knowledge in reading comprehension
65. Yuan H, Yuan Z, Gan R, et al (2022) Biobart: Pretraining and evaluation of a biomedical generative language model. In: Proceedings of the 21st Workshop on Biomedical Language Processing, pp 97–109
66. Zhang J, Zhang H, Zhang D, et al (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 (Volume 1: Long Papers), pp 1718–1731
67. Zhang Z, Zhang A, Li M, et al (2023) Automatic chain of thought prompting in large language models. In: The Eleventh International Conference on Learning Representations, https://openreview.net/forum?id=5NTt8GFjUHkr
68. Zhu A, Hwang A, Dugan L, et al (2024) Fanoutqa: A multi-hop, multi-document question answering benchmark for large language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp 18–37

69. Zweigenbaum P (2003) Question answering in biomedicine. Proceedings Workshop on Natural Language Processing for Question Answering. EACL, Citeseer, pp 1–4