Prompting Explicit and Implicit Knowledge for Multi-hop Question Answering Based on Human Reading Process

Guangming Huang[1], Yunfei Long[1], Cunjin Luo[1], Jiaxing Shen[2], Xia Sun[3]
[1]School of Computer Science and Electronic Engineering, University of Essex, United Kingdom
[2]Department of Computing and Decision Sciences, Lingnan University, Hong Kong, China
[3]School of Information Science and Technology, Northwest University, China
{gh22231, yl20051, cunjin.luo}@essex.ac.uk, raindy@nwu.edu.cn, jiaxingshen@ln.edu.hk

Abstract

Pre-trained language models (PLMs) leverage chains-of-thought (CoT) to simulate human reasoning and inference processes, achieving proficient performance in multi-hop QA. However, a gap persists between PLMs' reasoning abilities and those of humans when tackling complex problems. Psychological studies suggest a vital connection between explicit information in passages and human prior knowledge during reading. Nevertheless, current research has given insufficient attention to linking input passages and PLMs' pre-training-based knowledge from the perspective of human cognition studies. In this study, we introduce a **P**rompting **E**xplicit and **I**mplicit knowledge (PEI) framework, which uses prompts to connect explicit and implicit knowledge, aligning with human reading process for multi-hop QA. We consider the input passages as explicit knowledge, employing them to elicit implicit knowledge through unified prompt reasoning. Furthermore, our model incorporates type-specific reasoning via prompts, a form of implicit knowledge. Experimental results show that PEI performs comparably to the state-of-the-art on HotpotQA. Ablation studies confirm the efficacy of our model in bridging and integrating explicit and implicit knowledge.

Keywords: Multi-hop QA, Prompt, Implicit Knowledge

1. Introduction

Multi-hop question answering (QA) represents a challenging task that demands intricate reasoning and inference across diverse sources to predict a coherent and precise answer (Yang et al., 2018; Cao et al., 2023). Chain-of-thought (CoT) emulates the process of human reasoning and inference to generate a sequence of intermediate natural language reasoning steps that lead to the final results for complex reasoning problems. By employing prompt-based learning with CoT on pre-trained language models (PLMs), several studies have demonstrated proficient performance for multi-hop QA (Trivedi et al., 2022a; Deng et al., 2022; Zhang et al., 2023).

However, a substantial disparity remains evident between the reasoning abilities of PLMs and human cognition when it comes to addressing complex problems. Current studies overlook the establishment of a direct link between the input passages and the knowledge assimilated by PLMs during the pre-training phase, considering the cognitive framework of humans (Wang et al., 2022; Deng et al., 2022; Atif et al., 2023; Cao et al., 2023).

In human reading comprehension studies, Smith (1971) believes that information sources are often repeated during reading comprehension, leading to redundancy at linguistic levels, including letter-to-letter, word-to-word, sentence-to-sentence, and text-to-text. Consequently, readers can reduce their reliance on explicit information within the reading text by incorporating alternative sources of information, such as world knowledge (Hagoort et al., 2004). According to Clarke and Silberstein (1977), readers engage in the comprehension and question-answering process while reading, drawing upon both the explicit information conveyed in the text and their 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 usually referred to as prior knowledge (Yin, 1985; Baldwin et al., 1985; Abdelaal and Sase, 2014). The experimental results demonstrate a significantly high correlation between the high prior knowledge and reading comprehension (Abdelaal and Sase, 2014).

As an example shown in Figure 1, given a question "*Was Morris Lee born in the capital of Democratic Republic of the Congo ?*", human retrieves the information in the given passages to make a prediction. According to the auxiliary verb "*was*" in the yes-no question, human can predict the answer as "*yes*" or "*no*" beforehand, drawing upon linguistic knowledge (as part of implicit knowledge), even in the absence of information regarding the capital of Congo.

Consequently, an inseparable and inherent linkage prevails between the explicit information within the passages and the pre-existing prior knowledge. The prior knowledge diminishes the reliance on explicit information, thereby reducing the need for

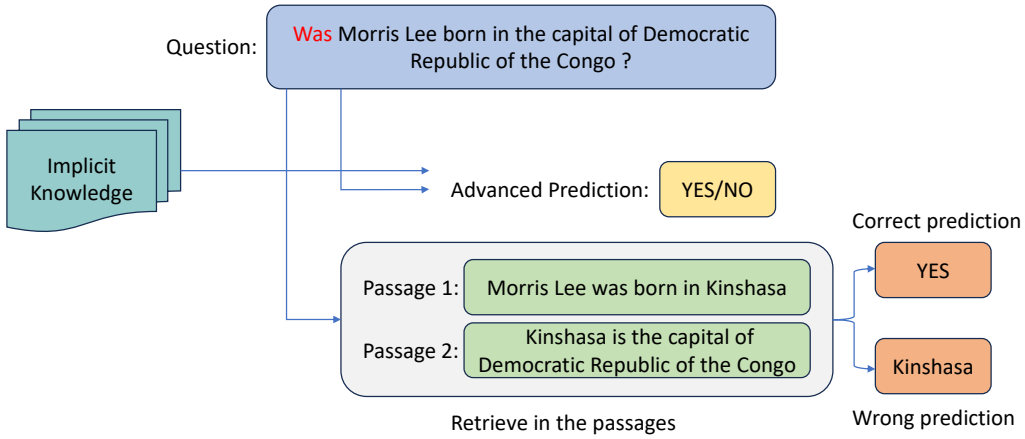

Figure 1: An example of the significance of implicit knowledge in reading comprehension.

its extensive utilization. Moreover, the harmonious integration of prior knowledge and explicit information enhances the efficacy of the reading process, leading to improved comprehension and engagement.

Incorporating insights from the aforementioned human cognition theories, we propose a novel framework called **P**rompting **E**xplicit and **I**mplicit knowledge (PEI) to address the challenges in multi-hop QA. Within this framework, readers are analogized to PLMs, their prior knowledge represents implicit knowledge acquired during the pre-training, and the explicit information in the passages serves as the input context conveying explicit knowledge. While acknowledging existing disparities between language models and human, and recognizing potential biases in considering readers as language models, Jin and Rinard (2023) believe that language models transcend being mere "stochastic parrots" (Bender et al., 2021), possessing the capability to acquire semantics during the pre-training.

To effectively utilize these knowledge sources, we employ prompts to capture explicit knowledge and invoke implicit knowledge. This facilitates the bridging of these two types of knowledge, thereby improving the performance of our proposed PEI model for multi-hop QA tasks. Moreover, this approach reduces the reliance on explicit knowledge present in the input passages by allowing selective filtering of irrelevant information or "redundancy" unrelated to the corresponding questions, as per Smith (1971)'s theory. To further substantiate the contribution of implicit knowledge to our model's performance, we conduct an ablation study, the results of which affirm our hypothesis (see Section 4.5).

Illustrated in Figure 2, our proposed PEI framework consists of three components: 1) a type prompter for identifying and learning the weights of reasoning types for multi-hop questions; 2) an encoder-decoder PLM is employed to acquire implicit knowledge by leveraging explicit knowledge, mirroring the cognition process of human reading process.; 3) a unified prompt-based PLM integrating both explicit, implicit knowledge and question types for multi-hop QA.

Our contributions are summarized as follows:

- The proposed PEI framework provides an effective approach for multi-hop QA based on human reading process, by modeling the input passages or context as explicit knowledge and PLMs mirroring with human prior knowledge as implicit knowledge.
- Our proposed PEI model achieves comparable performance with state-of-the-art on HotpotQA that is a benchmark dataset for multi-hop QA. Notably, our approach also maintains robust performance on corresponding single-hop subquestions and other multi-hop datasets.
- Ablation studies confirm that implicit knowledge enhances the model's reasoning ability, supporting our hypothesis for the PEI model, grounded in the human reading process.

2. Related Work

2.1. Chain-of-Thought Prompting

Prompt tuning has been recognized as an effective mechanism for conditioning PLMs to utilize relevant knowledge for specific downstream tasks (Lester et al., 2021; Liu et al., 2023). CoT prompting, a prompt-based approach, has emerged as a method to recall implicit knowledge from large-scale language models (LLMs) for complex reasoning. It emulates the sequential and coherent reasoning process of human thinking by generating intermediate natural language reasoning steps that lead to the final outcome (Chu et al., 2023). Manual-CoT

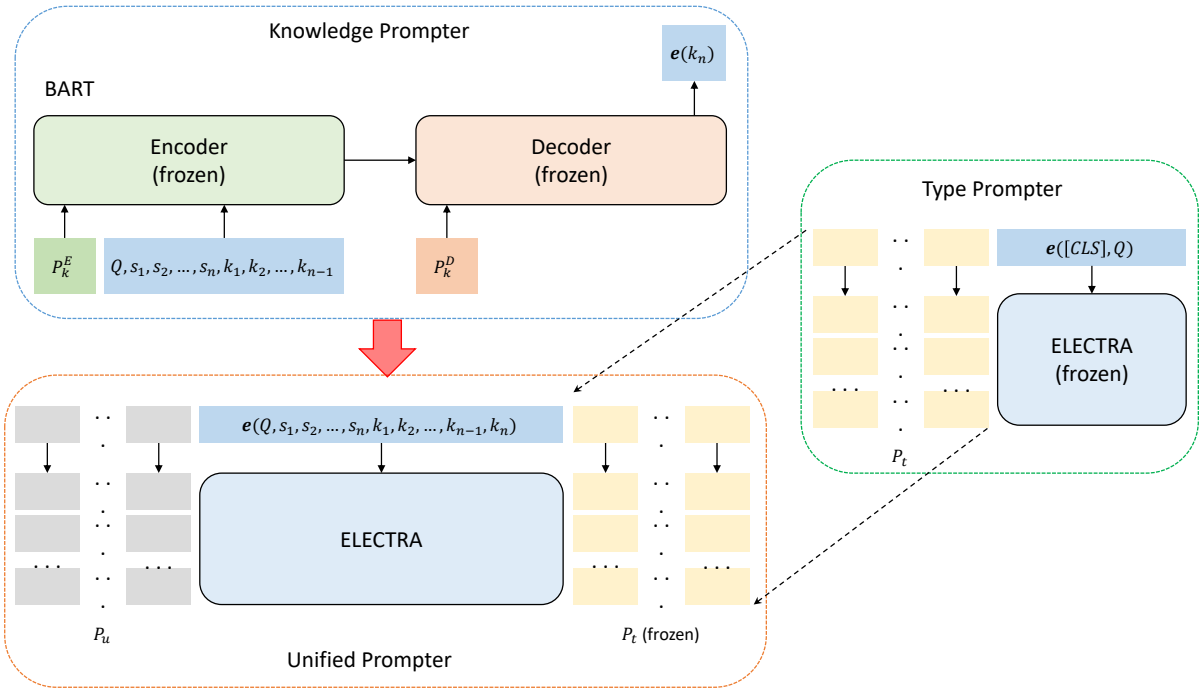

Figure 2: The overview of our proposed PEI framework for multi-hop QA. The right green dashed block is the type prompter; the top blue dashed block refers to the knowledge prompter; and the bottom orange dashed block is the unified prompter.

(Wei et al., 2022) is designed to elicit the CoT reasoning ability through manual demonstrations. Subsequently, Kojima et al. (2022) demonstrated that LLMs can serve as competent zero-shot reasoners, as they generate rationales that already embody CoT reasoning, by incorporating the incantation "Let's think step by step" to facilitate a step-by-step thinking process before answering questions. An automatic CoT prompting, AutoCoT (Zhang et al., 2022), leverages diverse question sampling and reasoning chain generation to construct demonstrations, reducing the needs for human effort. Some studies explored CoT prompt learning for multihop QA (Wang et al., 2022; Trivedi et al., 2022a). Inspired by these advancements, our work explores the utilization of CoT prompting to elicit implicit knowledge from PLMs. In contrast to CoT's generation of intermediate natural language steps, our model generates continuous embeddings as representations of implicit knowledge.

2.2. Prompt-based Learning for Multi-hop QA

Recent studies have made significant progress by employing prompts for multi-hop QA. For example, PromptRank (Khalifa et al., 2023) constructs an instruction-based prompt that includes a candidate document path and then computes the relevance score between a given question and the path based on the conditional likelihood of the question given the path prompt according to a language model. IRCoT (Trivedi et al., 2022a) interleaves CoT generation and knowledge retrieval steps to guide the retrieval using CoT prompting. Wang et al. (2022) introduced an iterative CoT prompting framework, which elicits knowledge from PLMs step by step utilizing a sequence-to-sequence BART-large model to recall a sequence of natural language statements for multi-hop QA. They employed a simple template to convert each triplet (subject entity, relation, object entity) in the evidence path into a natural language statement, which collectively formed the final statement. Inspired by their work, our approach also utilizes an encoder-decoder architecture PLM, i.e. BART-large, to recall implicit knowledge through iterative prompting. However, our proposed method differs distinctly from their study in three aspects: 1) our method does not necessitate the conversion of triple evidence paths into natural language statements; 2) we leverage input passages as explicit knowledge to elicit implicit knowledge from the PLM, which was not explored in their approach; 3) the recalled implicit knowledge in our model is represented as continuous embeddings rather than natural language statements or lexical knowledge (Huang et al., 2023). Hence, our model does not rely on the natural language statements derived from the evidence paths.

3. Methodology

3.1. Overview

Shown in Figure 2, our PEI framework comprises three components: 1) the type prompter identifies and acquires the weights associated with specific reasoning types for multi-hop questions; 2) the knowledge prompter elicits implicit knowledge through harnessing of explicit knowledge; 3) the unified prompt-based PLM integrates explicit and implicit knowledge, as well as question types, providing a comprehensive approach for multi-hop QA.

Pre-training on Single-hop QA. In order to understand the performance characteristics of current QA models at each step of the reasoning process, we built our works on a PLM foundation model, i.e. ELECTRA [1] (Clark et al., 2020) on SQuAD (Rajpurkar et al., 2016), which is a single-hop QA dataset. Subsequently, the pre-trained ELECTRA model is employed as the backbone PLM for our type prompter module. By leveraging the trained ELECTRA model on single-hop, we aim to explore the interaction between the model's behaviour and the reasoning process across various hops.

3.2. Type Prompter

The type prompter is crafted to facilitate the training process for the acquired weights of soft prompts, enabling them to effectively capture the distinctive features associated with various question types. As shown in Figure 2, yellow blocks P_t refer to trainable prompt embeddings; blue blocks are embeddings stored and frozen PLM.

Given question Q, we construct an input sequence consisting of soft prompts denoted as $\{P_t, [CLS], Q\}$, where P_t represents the trainable soft prompts of the type prompter module. P-tuning v2 (Liu et al., 2021) is employed to train soft prompts P_t acquiring the weights of specific-type information of the given questions. As shown in Type Prompter of Figure 3, we initially freeze the PLM and optimize the trainable soft prompt P_t. After training, we obtain the trained P_t and establish a connection between it and the unified prompter, ensuring its fixed nature throughout subsequent operations.

We denote d as the embedding dimension of the language model ELECTRA, h as the number of layers within the PLM, and l as the length of the prompt tokens. In this component, the total number of trainable parameters can be calculated as $\Theta(d \cdot h \cdot l)$. Compared with the fine-tuning paradigm, the type prompter module based on p-tuning can reduce training parameters while ensuring that it can capture the type-specific information. Secondly, it facilitates to transfer the trained weights of P_t with the type-specific information to unified prompter module. Moreover, through the application of the p-tuning v2, we can effectively capture and learn a broader range of features than prompt tuning (Lester et al., 2021).

[1]The Pre-trained language models can also be replaced by more competent models. In line with previous works on prompt learning, we choose ELECTRA.

3.3. Knowledge Prompter

By capitalizing on the textual information input, the knowledge prompter seeks to activate and engage the innate prior knowledge possessed by individuals, thereby facilitating the assimilation of these two information sources for comprehensive reading comprehension. As illustrated in Figure 2, the knowledge prompter, an iterative encoder-decoder PLM, retrieves implicit knowledge through prefix tuning (Li and Liang, 2021). The trainable prompt embeddings, denoted as P_k^E and P_k^D, are integrated into each layer of the encoder and decoder of the PLM, respectively. This incorporation of prompt embeddings facilitates the efficient retrieval and utilization of explicit knowledge throughout the iterative encoding and decoding phases.

Given a multi-hop query Q and a sequence of supporting sentences $S_n = [s_1, s_2, ..., s_i, ..., s_n]$, our objective is to retrieve 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 prompting methods, where we aim to construct prompts P_k^E and P_k^D to guide the encoder-decoder language model in recalling the desired knowledge K_n. Notably, we maintain fixed parameters for the encoder-decoder PLM, allowing us to direct its retrieval process through the strategic construction of prompts.

Motivated by the sequential nature observed in multi-step reasoning tasks (Wang et al., 2022), we adopt an iterative approach as below:

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

$$decoder(k_j) = encoder(Q, S_j, K_{j-1}) \quad (2)$$

where at each step j, PLM 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 (1) and (2):

$$decoder(k_1) = encoder(Q, s_1) \quad (3)$$

3.4. Unified Prompter

As described in Figure 2, we concatenate the unified prompter module and P_t, which is with the weights of type-specific reasoning. Moreover, the implicit knowledge K_n from the knowledge prompter module is the additional inputs. Intuitively, this amalgamation of two information is capable to enhances the model's reasoning ability.

In the unified prompter, the trainable prompt embeddings are denoted as P_u, while we freeze the trained promopt embeddings P_t. To preserve the learned weights of P_t from the type prompter, we adopt the identical architecture of the language model (ELECTRA), allowing for seamless concatenation of P_t with the unified prompter. Subsequently, we perform joint fine-tuning and prompt-tuning of both the language model and the prompt P_u to optimize the overall performance of the model in the context of multi-hop question-answering (QA) prediction.

4. Experiments

4.1. Dataset and Metrics

HotpotQA (Yang et al., 2018) contains a collection of 113k question-answer pairs drawn from Wikipedia. Additionally, HotpotQA offers sentence-level supporting facts that are essential for the process of reasoning, which allows QA systems to infer with robust supervision and explain the predictions.

2WikiMultiHopQA (Ho et al., 2020) has over 192k samples, including 167k training, 12.5k evaluation, and 12.5 test samples. The format of the dataset primarily follows HotpotQA (Yang et al., 2018), but it provides additional enhancements, including a broader range of reasoning types for questions and comprehensive annotations of evidence paths associated with each question.

MuSiQue (Trivedi et al., 2022b) contains 25k 2-4 hop questions samples. This dataset involves the systematic selection of composable pairs of single-hop questions demonstrating logical connections to generate a set of multi-hop questions.

Sub-question QA dataset (Tang et al., 2021) was created to facilitate the analysis of the reasoning capabilities of multi-hop QA models at each step of the reasoning process. To evaluate the performance of these models, the authors curated a specialized dataset consisting of single-hop sub-questions. This dataset comprises 1000 samples manually verified from the development set of HotpotQA, ensuring high-quality evaluation resource for the study.

In order to ensure consistency and comparability across the datasets used in our experimental evaluations, we categorize the question types of the three datasets under investigation into the broader categories of comparison and bridge. This categorization of question types facilitates a standardized approach to handling diverse question structures across the datasets subjected to our evaluation.

Metrics. We employ 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.

4.2. Implementation Details

Inspired by the studies of Wang et al. (2022) and Deng et al. (2022), we adopt BART-large (Lewis et al., 2020) as the language model in the knowledge prompter module. ELECTRA-large (Clark et al., 2020) serves as the foundation language model for both the type prompter and unified prompter modules. Our implementation is built upon the Huggingface platform (Wolf et al., 2020). For model optimization, we employ the AdamW optimizer (Loshchilov and Hutter, 2018) 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. For the HotpotQA, 2WikiMultiHopQA, and MuSiQue datasets, 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 and the type prompts P_t, exploring values from $\{15, 30, 45, 60, 75, 90, 100\}$.

4.3. Main Results and Analysis

We compare our PEI model with other published baselines on the test set of HotpotQA in the distractor setting, including the baseline model of HotpotQA (Yang et al., 2018), SOTA model on the leaderboard (Zhang et al., 2023), iCAP (Wang et al., 2022) and PCL (Deng et al., 2022) which we inspired by, and other baselines. As shown in Table 1, our proposed PEI model demonstrates superior performance across all evaluation metrics compared to all other baselines (except Beam Retrieval), and achieves comparable performance with Beam Retrieval (Zhang et al., 2023) on the HotpotQA dataset, highlighting the significant progress made by PEI for multi-hop QA.

Specifically, our PEI achieves an improvement of 0.20 in answer EM, 0.28 in answer F1 score, and 0.30 in join F1 score when compared to Beam Retrieval. Conversely, Beam Retrieval shows an improvement of 1.22 in supporting EM, 0.28 in sup-

Models	Ans		Sup		Joint	
	EM	F1	EM	F1	EM	F1
Baseline Model (Yang et al., 2018)	45.60	59.02	20.32	64.49	10.83	40.16
DecompRC (Min et al., 2019)	55.20	69.63	-	-	-	-
OUNS (Perez et al., 2020)	66.33	79.34	-	-	-	-
QFE (Nishida et al., 2019)	53.86	68.06	57.75	84.49	34.63	59.61
DFGN (Qiu et al., 2019)	56.31	69.69	51.50	81.62	33.62	59.82
SAE-large (Tu et al., 2020)	66.92	66.92	61.53	86.86	45.36	71.45
C2F Reader (Shao et al., 2020)	67.98	81.24	60.81	87.63	44.67	72.73
Longformer (Beltagy et al., 2020)	68.00	81.25	63.09	88.34	45.91	73.16
HGN (Fang et al., 2020)	69.22	82.19	62.76	88.47	47.11	74.21
AMGN (Li et al., 2021)	70.53	83.37	63.57	88.83	47.77	75.24
S2G (Wu et al., 2021)	70.72	83.53	64.30	88.72	48.60	75.45
iCAP † (Wang et al., 2022)	68.61	81.82	62.80	88.51	47.02	74.11
PCL (Deng et al., 2022)	71.76	84.39	64.61	89.20	49.27	76.56
Beam Retrieval (Zhang et al., 2023)	72.69	85.04	**66.25**	**90.09**	**50.53**	77.54
PEI (Ours)	**72.89**	**85.32**	65.03	89.81	49.91	**77.84**

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. "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. HotpotQA Leaderboard: https://hotpotqa.github.io/.

porting F1 score, and 0.62 in join EM when compared to PEI.

The difference in performance between PEI and Beam Retrieval in answer prediction versus supporting prediction can be attributed to their respective approaches. Beam Retrieval maintains multiple partial hypotheses of relevant passages at each step, 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. In contrast, our PEI model leverages insights from the human reading process, incorporating implicit knowledge and type-specific information. This approach contributes to the model's improvement in answer prediction. However, it may not exhibit the same level of performance in supporting prediction as Beam Retrieval due to differences in retrieval strategies.

Moreover, our PEI framework demonstrates a significant improvement of $0.64/1.28$ in the Joint EM/F1 score compared to the PCL model, which PEI and PCL both use the same backbone PLM (i.e., ELECTRA). Although PCL also identified the reasoning type of multi-hop question as a soft prompt via a transformer-based question classifier, our proposed model not only consider the type-specific knowledge but also incorporate the implicit knowledge through iteratively eliciting it from an encoder-decoder PLM. Additionally, our proposed PEI outperforms iCAP with $2.89/3.73$ in the joint EM/F1 score, despite both PEI and iCAP utilizing the same encoder-decoder skeleton PLM (i.e.,BART). Compared to the graph-based model AMGN, our PEI framework exhibits substantial gains, with an improvement of $2.14/2.6$ in the Joint EM/F1 score. This indicates that our framework achieves better performance employing the same backbone PLM.

4.4. Evaluation of Robustness

Since our proposed model employs identical backbone models (i.e., ELECTRA and BART) and similar prompt learning framework as PCL (Deng et al., 2022) and iCAP (Wang et al., 2022), we further evaluate the robustness of our model, particularly in comparison to PCL and iCAP. Additionally, we extend our evaluation to include a graph-based model, HGN (Fang et al., 2020), to ensure a thorough assessment of its robustness.

Evaluation on Other Multi-hop Datasets. To verify the generalization, we evaluate PEI model on 2WikiMultihopQA and MuSiQue. In Table 2, the results show that our PEI model surpasses all comparison baselines in terms of both EM and F1 metrics. Notably, despite both PEI and iCAP employ the same encoder-decoder BART model, PEI achieves a substantial improvement of $4.52/26.66$ in the answer EM/F1 score compared to iCAP on the 2WikiMultihopQA dataset. Furthermore, our model exhibits superior performance compared to PCL, yielding improvements of $1.29/1.14$ and $0.69/0.51$ in the answer EM/F1 score on 2WikiMultihopQA and MuSiQue, respectively.

Evaluation on Sub-question Dataset. To evaluate the PEI model's efficacy in the multi-hop rea-

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**

Table 2: Results of our proposed PEI compared to PCL, HGN and iCAP on 2WikiMultihopQA and MuSiQue multi-hop QA test set.

soning process, specifically in composing answers from solved sub-questions, we conducted an evaluation on sub-question QA dataset (Tang et al., 2021). This dataset consists of 1000 multi-hop questions q along with their corresponding sub-questions q_{sub1} and q_{sub2}. Table 3 shows that our PEI model achieves a 97.62% success rate in correctly answering the parent multi-hop question q when both sub-questions q_{sub1} and q_{sub2} are answered correctly [2]. This highlights the proficiency of PEI in retaining acquired knowledge through the integration of explicit and implicit knowledge, surpassing other multi-hop QA models. However, it is noteworthy that PEI also demonstrates a significant likelihood (36.55%) of correctly answering the parent multi-hop question even when only one of the sub-questions is answered accurately [3]. To further illustrate the sub-question dependent success rate of various multi-hop QA models, Figure 3 shows that these models exhibit a high probability (exceeding 20%) of correctly answering parent multi-hop questions even when only one sub-question is answered correctly. This indicates that the utilization of potentially unsound reasoning shortcuts in predicting answers is a prevalent and challenging phenomenon in multi-hop QA tasks.

4.5. Ablation Studies

To evaluate the contributions of distinct components within our PEI model, we conducted a series of ablation studies using the development set of HotpotQA as the experimental platform.

Effect of Implicit Knowledge. To verify the hypothesis that implicit knowledge enhances reasoning for multi-hop QA, we conduct an ablation study between the performance of the ELECTRA model with and without implicit knowledge. In Table 4, the experimental results show that the ELECTRA model with implicit knowledge improves $3.10/2.05/2.30$ in Ans F1/Sup F1/Joint F1 compared to the model without the implicit knowledge. These findings support the notion that implicit knowledge enhances the reasoning ability of the model thereby proving the hypothesis underlying our proposed PEI model, inspired by human reading comprehension.

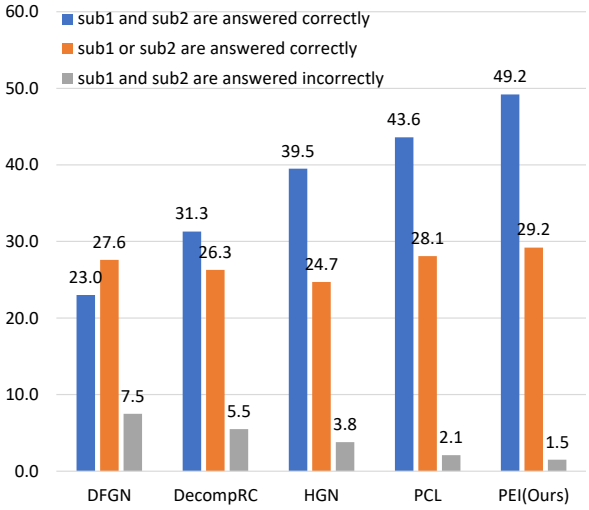

Figure 3: The success rate (%) of five multi-hop QA models. $sub1$ denotes the first sub-question and $sub2$ is the second sub-question of corresponding question q.

Effect of Type Prompts. To verify the effects of the type prompts and perform type-specific reasoning on multihop QA, we conduct a comparative analysis between the performance of the ELECTRA language model with and without the type prompter. As illustrated in Table 4, the language model combined with the type prompter achieves a substantial improvement of $3.02/1.17/2.18$ in Ans/Sup/Joint F1 compared to the model without the type prompter component. These findings demonstrate that integrating question type information through the type prompter module effectively enhances the overall performance of the model and enables type-specific reasoning for multi-hop QA. Furthermore, these results validate the alignment of our model design with the cognitive processes observed in human reasoning. Because type information could be consider as one form of implicit knowledge.

Effect of Pre-training on Single-hop. Initially, we trained an ELECTRA-based QA model on the single-hop QA dataset SQuAD (Rajpurkar et al., 2016), and subsequently retrained it on the HotpotQA dataset. Although conservation learning (Deng et al., 2022) is not employed in our model, we evaluate the performance of our model with and without pre-training in order to verify the effect of the pre-training on single-hop. As shown in Table 4, the language model combined with the pre-training improves $0.70/0.62/1.04$ in Ans F1/Sup F1/Joint F1 compared to the model without the pre-training.

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

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

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

Table 3: Categorical EM statistics (%) of sub-question evaluation for five multi-hop QA models. c/w denotes that the question is answered correctly/wrongly. $sub1$ denotes the first sub-question and $sub2$ is the second sub-question of corresponding question q. For example, the first four rows show the percentage of multi-hop questions that can be correctly answered.

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** $\uparrow_{5.52}$

Table 4: Ablation Study of PEI on the development set of HotpotQA. Ans F1 stands for answer F1; Sup F1 is supporting F1.

This indicates that pre-training in the single-hop QA task enable the model to acquire valuable information, enhancing its overall performance. However, it is noteworthy that this improvement is relatively limited without conservation learning.

Effect of Foundation PLMs. To assess the effects of foundation PLMs, we compare PEI with PCL and HGN based on the same data and backbone. As shown in Table 5, PEI consistently outperforms PCL and HGN across all metrics. This indicates the effectiveness and robustness of PEI across PLMs. Moreover, PEI with ALBERT achieves an improvement of $0.62/0.23$ in Ans/Sup F1 compared to PEI with RoBERTa. ALBERT outperforms RoBERTa on GLUE benchmark using a single-model setup [4]. These results confirm that adopting a more competent foundation PLM can improve the performance of our model.

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**

Table 5: Results with different PLMs on the development set of HotpotQA. RoBERTa, ELECTRA and ALBERT denote that we use RoBERTa-large, ELECTRA-large and ALBERT-xxlargev2 as the PLM respectively

5. Conclusions and Future Work

In this paper, we introduce a novel framework that mimics human cognitive reading processes, employing prompts to bridge explicit and implicit knowledge. Our framework leverages prompts to elicit implicit knowledge from PLMs within the input context. Additionally, we integrate question type information to enhance model performance. Experimental results show that PEI performs comparably to the state-of-the-art on HotpotQA. Furthermore ablation studies confirm the effectiveness and robustness of our model in emulating human reading processes. In the future, we aim to extend and apply human reading cognition theories to diverse reasoning tasks, with the hope of enabling stronger, complex reasoning capabilities.

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

6. Limitations

Human reading and cognition theory for reasoning. While our experimental results demonstrate that our model outperforms all but one baseline model for multi-hop QA, we have solely validated the efficacy of these cognition theories within this specific domain. Future research opportunities include extending these principles to diverse reasoning tasks, such as mathematical reasoning. Additionally, exploring alternative theories of human cognition and their potential applications in reasoning tasks would be valuable.

Interpretability of implicit knowledge. Although leveraging implicit knowledge benefits the model's reasoning process, the soft prompting we elicit are challenging to explain. Currently, it is a lack of insights into the specific knowledge acquired and how it contributes to the model's reasoning processes and decision-making capabilities.

Experiments with larger-scale models. While we currently conduct our experiments with BART and ELECTRA, the availability of large-scale language models like GPT-3 presents an opportunity to enhance the model's capabilities by incorporating more powerful models with richer prior knowledge. Nevertheless, it remains imperative to strike a balance between computational cost and model performance.

7. 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.

8. Bibliographical References

Noureldin Mohamed Abdelaal and Amal Saleh Sase. 2014. Relationship between prior knowledge and reading comprehension. *Advances in Language and Literary Studies*, 5(6):125–131.

Farah Atif, Ola El Khatib, and Djellel Difallah. 2023. Beamqa: Multi-hop knowledge graph question answering with sequence-to-sequence prediction and beam search. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 781–790.

R Scott Baldwin, Ziva Peleg-Bruckner, and Ann H McClintock. 1985. Effects of topic interest and prior knowledge on reading comprehension. *Reading research quarterly*, pages 497–504.

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

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 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*, pages 610–623.

Ziyi Cao, Bingquan Liu, and Shaobo Li. 2023. Rpa: reasoning path augmentation in iterative retrieving for multi-hop qa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12598–12606.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2023. A survey of chain of thought reasoning: Advances, frontiers and future. *arXiv preprint arXiv:2309.15402*.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Mark A Clarke and Sandra Silberstein. 1977. Toward a realization of psycholinguistic principles in the esl reading class 1. *Language learning*, 27(1):135–154.

Zhenyun Deng, Yonghua Zhu, Yang Chen, Qianqian Qi, Michael J Witbrock, and Patricia Riddle. 2022. Prompt-based conservation learning for multi-hop question answering. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1791–1800.

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

Peter Hagoort, Lea Hald, Marcel Bastiaansen, and Karl Magnus Petersson. 2004. Integration of word meaning and world knowledge in language comprehension. *science*, 304(5669):438–441.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625.

Guangming Huang, Yunfei Long, Cunjin Luo, and Yingya Li. 2023. LIDA: Lexical-based imbalanced data augmentation for content moderation. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 59–69, Hong Kong, China. Association for Computational Linguistics.

Charles Jin and Martin Rinard. 2023. Evidence of meaning in language models trained on programs. *arXiv preprint arXiv:2305.11169*.

Muhammad Khalifa, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, and Lu Wang. 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)*, pages 15882–15897.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 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*, pages 7871–7880.

Ronghan Li, Lifang Wang, Shengli Wang, and Zejun Jiang. 2021. Asynchronous multi-grained graph network for interpretable multi-hop reading comprehension. In *IJCAI*, pages 3857–3863.

Xiang Lisa Li and Percy Liang. 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)*, pages 4582–4597.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2021. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

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

Kosuke Nishida, Kyosuke Nishida, Masaaki Nagata, Atsushi Otsuka, Itsumi Saito, Hisako Asano, and Junji Tomita. 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*, pages 2335–2345.

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

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

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Nan Shao, Yiming Cui, Ting Liu, Shijin Wang, and Guoping Hu. 2020. Is graph structure necessary for multi-hop question answering? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7187–7192.

Frank Smith. 1971. *Understanding reading: A psycholinguistic analysis of reading and learning to read*. Holt, Rinehart and Winston, New York.

Yixuan Tang, Hwee Tou Ng, and Anthony Tung. 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*, pages 3244–3249.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022a. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. *arXiv preprint arXiv:2212.10509*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022b. Musique: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Ming Tu, Kevin Huang, Guangtao Wang, Jing Huang, Xiaodong He, and Bowen Zhou. 2020. Select, answer and explain: Interpretable multi-hop reading comprehension over multiple documents. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9073–9080.

Boshi Wang, Xiang Deng, and Huan Sun. 2022. Iteratively prompt pre-trained language models for chain of thought. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2714–2730.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, 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*, pages 38–45.

Bohong Wu, Zhuosheng Zhang, and Hai Zhao. 2021. Graph-free multi-hop reading comprehension: A select-to-guide strategy. *arXiv preprint arXiv:2107.11823*.

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

Koh Moy Yin. 1985. The role of prior knowledge in reading comprehension.

Jiahao Zhang, Haiyang Zhang, Dongmei Zhang, Yong Liu, and Shen Huang. 2023. Beam retrieval: General end-to-end retrieval for multi-hop question answering. *arXiv preprint arXiv:2308.08973*.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*.