Tree- or Chain-of-Thought? Exploring Complex Reasoning in Multi-Hop Question Answering

Zhenyu Bi
Virginia Tech
zhenyub@vt.edu

Daniel Hajialigol
Virginia Tech
danielhajialigol@vt.edu

Priya Pitre
Virginia Tech
priyapitre@vt.edu

Zhongkai Sun
Amazon Alexa AI
zhongkas@amazon.com

Jie Hao
Amazon Alexa AI
jieha@amazon.com

Xuan Wang
Virginia Tech
xuanw@vt.edu

Abstract

Multi-hop question answering (MHQA) requires a model to retrieve and integrate information from multiple passages to answer a complex question. Recent systems leverage the power of large language models and integrate evidence retrieval with reasoning prompts for the MHQA task. However, the complexities in the question types as well as the reasoning types require more novel and fine-grained prompting methods to enhance the performance of MHQA under the zero-shot setting. In this paper, we propose a tree-of-thought reasoning prompting method for MHQA and conduct a detailed comparison with chain-of-thought reasoning on different question types and reasoning types. Specifically, we construct a tree-like reasoning structure, by prompting the model to break down the original question into smaller sub-questions at each reasoning step. To the best of our knowledge, our work is the first to adapt the tree-of-thought reasoning prompting to natural language tasks such as MHQA. Experiments on HotpotQA showed that our method with tree-of-thought prompting works better in general, with an advantage in forming intermediate reasoning lines. A detailed comparison showed that tree-of-thought prompting is better at comparison and parallel questions, while chain-of-thought prompting is better at bridge and sequential questions.

1 Introduction

The question answering (QA) task is a fundamental problem in natural language processing (NLP) that involves designing systems capable of understanding human language questions and providing accurate and relevant answers. With the recent advancement of Large Language Models (LLMs) that demonstrated superior reasoning ability (Brown et al., 2020), researchers have been focusing more on complex benchmarks, such as Multi-hop Question answering (MHQA). MHQA is more challenging as it requires models to understand complicated questions, perform multiple reasoning steps, and gather evidence across different documents.

Inspired by the Tree-of-Thought (ToT) prompting method (Yao et al., 2023) that performs well on tasks including Mathematical Reasoning and Creative Writing, we propose a ToT-based MHQA method that allows the model to generate different reasoning paths from the same question, thus effectively avoiding reasoning dead-ends (Figure 1). To the best of our knowledge, our work is the first to adapt the Tree-of-Thought reasoning prompting to natural language tasks such as MHQA. We propose an effective ToT-based MHQA method with two modules of sub-question generation and evidence-based reasoning. We also conducted an in-depth analysis of different question types (bridge v.s. comparison questions) and reasoning types (sequential v.s. parallel reasonings) in HotpotQA (Yang et al., 2018), a benchmark MHQA dataset.

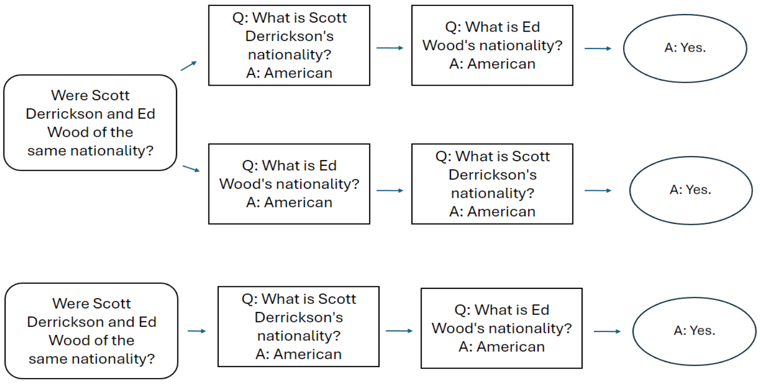
2 Method

2.1 Task Formation

Given a multi-hop question Q and background corpus of evidence P, the goal of our framework is to output the answer A to question Q, drawing its reasoning with the support of multiple evidence passages $p_1, p_2, ...$ retrieved from corpus P.

2.2 Tree of Thought Prompting

For each of the question Q, there could exist multiple reasoning lines, and thus multiple ways of breaking down the question. To avoid reasoning dead-ends, we prompt the model to output different possible reasoning lines at each reasoning step, by proposing different possible sub-questions to solve. Formally, given a question Q, we instruct the model to break down the question into sub-questions $q_1, q_2, ..., q_i$ and get their corresponding answer $a_1, a_2, ..., a_n$. We build a tree structure with the original question Q as the root node, and each (q_i, a_i) pair as subsequent nodes. The reasoning



Were Scott Derrickson and Ed Wood of the same nationality?
Q: What is Scott Derrickson's nationality? A: American
Q: What is Ed Wood's nationality? A: American
A: Yes.
Q: What is Ed Wood's nationality? A: American
Q: What is Scott Derrickson's nationality? A: American
A: Yes.
Were Scott Derrickson and Ed Wood of the same nationality?
Q: What is Scott Derrickson's nationality? A: American
Q: What is Ed Wood's nationality? A: American
A: Yes.

Figure 1: Example of the reasoning paths. Top: Tree-of-Thought Reasoning Path; Down: Chain-of-Thought Reasoning Path.

paths are thus represented as branches in the tree structure. We use majority voting on answers from different reasoning paths to decide the final answer to the original question.

2.3 ToT-Based MHQA Framework

We propose to solve the MHQA task in a two-module, self-interactive way. One part of the module focuses on question breakdown and forming reasoning paths, while the other focuses on evidence retrieval, reasoning, and resolving answers to the specific sub-question. By specializing in the task and avoiding providing the model with excessive information, the model will be able to focus on one task at a time and avoid confusion.

3 Experiment

3.1 Experimental Setup

Dataset We experiment our framework with the HotpotQA dataset (Yang et al., 2018), which is widely used across different MHQA baselines. The experiments are done in the distractor setting, where we provide the model with an evidence pool that contains both golden evidence and irrelevant evidence. We randomly selected 200 examples from the dataset and reported the results in terms of EM and F1.

Baselines We included two baselines, the Vanilla prompting and the Chain-of-Thought (CoT) prompting. For the Vanilla prompting, no examples are provided and we only present the model with the question and evidence. For CoT prompting, we use a standard input-output (IO) prompt with 1 in-context example, which presents the whole reasoning chain, including all intermediate steps, to the model.

Model Variant	**GPT3.5**	
Eval Metric	EM	F1
Zero-Shot Vanilla	32.0	44.0
Chain-of-Thought	35.5	47.3
Ours (Tree-of-Thought)	36.5	49.5
Ours (ToT, Upper Bound)	**40.5**	**53.5**

Table 1: Overall Results

LLM We experiment the baselines and our model utilizing the the gpt-3.5-turbo variant of InstructGPT.

3.2 Results

Results are presented in Table 1. Our ToT-based framework has an EM score of 36.5, which outperforms the Chain-of-Thought framework, which has an EM score of 35.5, by a small margin. The upper bound performance of our framework is a possible upper bound if we select the golden answer instead of majority voting when calculating the answer to the original question. It hints that the model's reasoning ability can be further improved if we provide a more complex metric that guides the model along its reasoning path. We provide a more in-depth analysis in the appendix section.

4 Conclusion

We explored tree-of-thought prompting in the task of Multi-hop Question Answering and compared its performance with chain-of-thought prompting. Experiments showed that the question type and the reasoning type jointly influence the LLM's reasoning ability. With our results, we hope to provide insights for future prompting methods with fine-grained considerations of questions and reasoning types in complex natural language reasoning tasks.

Limitations

We use a naive method, majority voting on the reasoning lines, to arrive at the final answer. The tree-of-thought prompting method can be further improved by implementing proper evaluators with solid evaluating metrics, as suggested by the upper bound performance in table 1. Other factors, such as how we designed our framework such as the sub-question generation module, might affect the tree-of-thought performance as well. More extensive experiments, including experimenting on other different datasets and case studies, should be provided.

Acknowledgements

This work is sponsored by the Commonwealth Cyber Initiative, Children's National Hospital, Fralin Biomedical Research Institute (Virginia Tech), Sanghani Center for AI and Data Analytics (Virginia Tech), Virginia Tech Innovation Campus, and a generous gift from the Amazon + Virginia Tech Center for Efficient and Robust Machine Learning. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Amazon + VT Center for Efficient and Robust ML.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *CoRR*, abs/2305.10601.

A Prompt Templates

Question Generation Module Template Given a question, break it into sub-questions that are easier to answer. Here are two example as guidelines: "Question: Are Tokyo and Busan in the same country? Thought 1: I could either find which country Tokyo is located in, or which country Busan is located in. Sub Question 1-1: Which country is Tokyo located in? Sub Question 1-2: Which country is Busan located in?" "Question: Tokyo is located in the country that has what colors present on its national flag? Thought 1: I need to first find out which country Tokyo is located in. Sub Question 1-1: Which country is Tokyo located in?" Only give out your thought process and current-level sub-questions. Do not give out answers to your questions. Question: *Given Question*. Thought 1:

Evidence Reasoning and Answering Template Given a question, and a list of evidence that may of help, find evidences that could help answer the question, and give out your answer. Here is an example as guideline: "Question: Which country is Tokyo located in? Evidence as reference: 1: Tokyo is the capitol of Japan. 2: Egypt is an African country. 3: Shinzo Abe was a Japanese politician and statesman who served as Prime Minister of Japan from 2012 to 2020. 4: The United States has a military base in Yokohama, Japan. 5: Kyoto was the capitol of Japan before Tokyo. 6: Japan is an Asian Country. Supporting Evidence: 1,5. Answer: Japan" Question: *Given Question*. Evidence as reference: *Given Evidence*. Supporting Evidence:

Chain of Thought Template Solve a question answering task, thinking step by step. Here is an example as a guideline: 'Question: In 2015, who is the prime minister of the country that the city Tokyo is located in? Evidence as reference: Tokyo is the capitol of Japan. Shinzo Abe was a Japanese politician and statesman who served as Prime Minister of Japan from 2012 to 2020. Answer: Think step by step. I need to first find out where Tokyo is located. First sentence in evidence suggests that Tokyo is located in Japan. Then, I need to find out the Prime Minister of Japan in 2015. Second sentence in evidence suggests that in 2015, the president of Japan is Shinzo Abe. Thus the final answer to the question is: Shinzo Abe. Final Answer: Shinzo Abe.' Now, answer the following question: *Given Question*. Evidence as reference: *Given Evidence*. Answer:

Model Variant	Comparison	Bridge
Chain-of-Thought	**52.9**	31.9
Ours (ToT)	41.2	**35.5**

Table 2: Performance Based on Question Type, reported in EM metric.

Model Variant	Sequential	Parallel
Chain-of-Thought	36.4	**48.8**
Ours (ToT)	**38.6**	43.9

Table 3: Performance Based on Reasoning Type, reported in EM metric.

Question Examples See table 4 and 5.

B Ablation Studies

We examine our ToT-based framework and Chain-of-Thought prompting by comparing their performance under different question-type settings proposed by HotPotQA. The "Bridge" question contains a "bridge entity” that connects the question and the final answer; while the "Comparison" question requires the model to compare two entities of the same type. Examples are provided in Table 4. Out of the 200 questions, 34 questions are Comparison Questions and 166 questions are Bridge Questions. The performance of our framework and that of Chain-of-Thought (CoT) prompting on the sampled questions are shown in Table 2.

Reasoning Type We also do an in-depth analysis of the reasoning types in the existing MHQA datasets, by randomly selecting 100 questions from our testing set above. The questions are roughly divided into three categories: questions that contain Tree-like Parallel Reasoning, Chain-like Sequential Questions, or single-hop questions that only require one level of reasoning. The Sequential questions follow a strict reasoning chain, and all the sub-questions must be solved to form the correct reasoning process. The Parallel Questions contain two or more reasoning paths that can be solved in arbitrary order. All Comparison Questions are Parallel Reasoning, but some Bridge Questions contain Parallel Reasoning. Examples are provided in Table 5. Out of the selected 100 questions, 44 questions were Sequential, 41 questions were Parallel, and 15 questions were single-hop, which was excluded from the comparison. The performance of our framework and that of Chain-of-Thought (CoT) prompting on the sampled questions are shown in Table 3.

Analysis Our ToT-based framework performs better at Bridge Questions and Sequential Questions, suggesting that our framework can avoid reasoning dead-ends and is better at forming intermediate reasoning lines. On the other hand, the Chain-of-Thought method performs parallel questions, suggesting that forming a coherent and fluent reasoning line is still important when answering multi-hop questions.

Bridge Question	**Comparison Question**
What distinction is held by the former NBA player who was a member of the Charlotte Hornets during their 1992-93 season and was head coach for the WNBA team Charlotte Sting?	Were Scott Derrickson and Ed Wood of the same nationality?

Table 4: Question Type Examples

Sequential Reasoning	**Parallel Reasoning**
The football manager who recruited David Beckham managed Manchester United during what timeframe?	What distinction is held by the former NBA player who was a member of the Charlotte Hornets during their 1992-93 season and was head coach for the WNBA team Charlotte Sting?

Table 5: Reasoning Type Examples