MindMap: Constructing Evidence Chains for Multi-Step Reasoning in Large Language Models

Yangyu Wu[1], **Xu Han**[1*], **Wei Song**[1], **Miaomiao Cheng**[1], **Fei Li**[2]
[1]Beijing Key Laboratory of Electronic System Reliability Technology,
College of Information Engineering, Capital Normal University, China
[2]School of Cyber Science and Engineering, Wuhan University, China
{wu, hanxu, wsong, miaomiao}@cnu.edu.cn
lifei_csnlp@whu.edu.cn

Abstract

Large language models (LLMs) have demonstrated remarkable performance across a range of natural language processing (NLP) tasks. However, they encounter significant challenges in automated reasoning, especially in multi-step reasoning scenarios. In order to solve complex reasoning problems, LLMs need to perform faithful multi-step reasoning based on a given set of facts and rules. A lot of work has focused on guiding LLMs to think logically by generating reasoning paths, but ignores the relationship among available facts. In this paper, we introduce MindMap, a straightforward yet powerful approach for constructing evidence chains to support reasoning in LLMs. An evidence chain refers to a set of facts that are associated with the same subject. In this way, we can organize related facts together to avoid missing relevant information. MindMap can seamlessly integrate with existing reasoning frameworks, such as Chain-of-Thought (CoT) and Selection-Inference (SI), by enabling the model to generate and select relevant evidence chains from independent facts. The experimental results on the bAbI and ProofWriterOWA datasets demonstrate the effectiveness of MindMap. Our approach can significantly enhance the performance of CoT and SI, particularly in multi-step reasoning tasks.

Introduction

The pursuit of general artificial intelligence has remained a central objective within the realm of artificial intelligence research (Silver et al. 2021; Goertzel and Pennachin 2007). Recent years have witnessed remarkable advancements in Natural Language Processing (NLP), largely attributing to the emergence of large language models (LLMs) (Ouyang et al. 2022). These models have exhibited exceptional efficacy across diverse tasks including machine translation, question answering, and reading comprehension (Yang et al. 2023; Bang et al. 2023). The strategic expansion of language model scale yields tangible improvements across various problem domains, with task performance exhibiting positive correlations with model size (Creswell, Shanahan, and Higgins 2023). However, a study pointed that the benefits of scaling up are significantly reduced when dealing with complex problems (Rae et al. 2021). Particularly, the enhanced advantages of these models in tasks that involve sophisticated logical reasoning are less evident compared to other tasks (Wei et al. 2022).

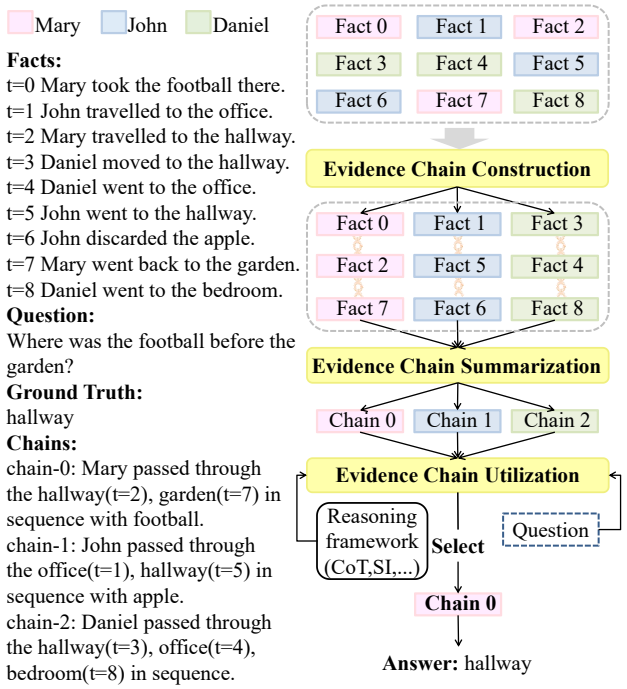

Figure 1: Illustrating the working flow of the proposed MindMap approach based on an example from the bAbI dataset.

Logical reasoning is essential for advancing various scientific fields (Liu et al. 2020). It involves deducing new conclusions from existing facts and rules. For instance, with facts like "David picked up the apple" and "David went to the bedroom,", deducing the apple's location being in the bedroom is a logical process. Such reasoning challenges often require multiple steps to be executed effectively to complete the reasoning process (Saparov and He 2023). Although LLMs shows good ability in learning from instructions and demonstrations in context to answer questions (Brown et al. 2020; Dong et al. 2023; Min et al. 2022), they struggle with complex logical reasoning, espe-

*Corresponding author.

cially multi-step reasoning (Wei et al. 2022; Liu et al. 2023; Kazemi et al. 2023).

Recent approaches have focused on guiding LLMs to think step-by-step to improve the performance in logical reasoning. For example, Chain of Thought (CoT) (Wei et al. 2022) and Selection-Inference (SI) (Creswell, Shanahan, and Higgins 2023) frameworks try to construct reasoning paths or formulate reason procedures and obtain large improvements in many reasoning tasks.

However, these methods general treat each individual fact as an isolated evidence, overlooking the inherent interconnections among these pieces of evidence. In this paper, we make a focused contribution in organizing available facts for supporting reasoning. When we deal with a specific problem, only part of the available facts are relevant, while others may be even noise. Therefore, it is important to group related facts together to prompt us to think more comprehensively and deeply. Motivated by the procedure of managing chains of custody in disclosure of crimes, we propose the **MindMap**, which aims to construct evidence chains for supporting logical reasoning.

Figure 1 shows the workflow of the proposed MindMap for responding to a question from the bAbI dataset. The framework consists of 3 modules: *evidence chain construction, evidence chain summarization* and *evidence chain utilization* for reasoning. Specifically, an evidence chain is defined as a group of facts associated with the same subject, such as a series of events involving a person. Motivated by the work of narrative event chains (Chambers and Jurafsky 2008), we extract subjects from facts using NLP tools to form a subject set and construct an evidence chain for every subject. To obtain more concise and coherent information, the evidence chain summarization module provides a summary that covers the main content and entities in the evidence chain.

In this manner, MindMap utilizes a set of constructed evidence chains or their summaries, rather than a collection of independent facts, to support further reasoning. MindMap can be integrated into existing reasoning frameworks such as CoT and SI. Instead of selecting facts, MindMap enhances CoT and SI by selecting and updating evidence chains.

We conduct evaluation on the bAbI (Weston et al. 2016) and ProofWriterOWA (Tafjord, Dalvi, and Clark 2021) datasets based on a LLM with 13B parameters. The experimental results show that MindMap can significantly improve the performance of CoT and SI, especially in the multi-step reasoning setting. We observe that MindMap can help cover many more supporting facts. Our method is straightforward, and its effectiveness underscores the necessity of organizing available facts. It also indicates that the integration of traditional NLP tools with LLMs has potential in addressing complex reasoning problems.

Related Work

Large Language Models

Due to the continuous advancement of deep learning technology and the increase in computing power, remarkable progress has been made in the development of LLMs (Radford et al. 2019; Chowdhery et al. 2023; Muennighoff et al. 2023). Notably, GPT4, which was released recently, has achieved excellent results in various tasks (Katz et al. 2023; Peng et al. 2023; Nori et al. 2023). However, these large language models possess numerous parameters and consume substantial resources, leading to the emergence of many smaller open-source models in response to current demands. For instance, Llama(13B) (Touvron et al. 2023) has demonstrated superior performance on most benchmarks when compared to GPT-3(175B) (Brown et al. 2020). Stanford's Alpaca (Touvron et al. 2023) and Vicuna (Chiang et al. 2023) models, which are supervised fine-tuned versions based on LLaMa, exhibit even stronger dialogue capabilities. Specifically, Vicuna utilizes GPT-4 for scoring and evaluation, boasting 13B parameters, and achieves up to 90% of ChatGPT's effectiveness (Chiang et al. 2023). Due to computation resource constraints, our experiments are conducted based on Vicuna.

Reasoning with LLMs

Automated reasoning has been a challenging task in NLP. Before the era of LLMs, the prevalent approaches to logical reasoning were based on fine-tuning pre-trained models (Clark, Tafjord, and Richardson 2020; Betz, Voigt, and Richardson 2021; Han et al. 2022). However, these methods often led to unrealistic inferences due to implicit label-data correlations (Zhang et al. 2023).

Recently, LLMs have shown stronger reasoning abilities compared to previous approaches (Dong et al. 2023; Min et al. 2022). The strength of LLMs lies in their ability to automatically learn from context through in-context learning (Dai et al. 2023; Min et al. 2022), enabling them to make accurate inferences by understanding specific contextual situations with just a few examples. Consequently, LLMs become more flexible in handling various tasks by only modifying the contextual hints (Dong et al. 2023).

However, LLMs face significant challenges when it comes to multi-step reasoning tasks (Zhou et al. 2023). Consequently, multi-step reasoning has emerged as a key area that LLMs need to address. A representative work in this direction is the Chain-of-Thought (CoT) approach (Wei et al. 2022), which aids the model in making correct inferences by outputting a series of reasoning paths, thereby also enhancing the interpretability of the model's outputs. Despite its advantages, CoT has its limitations, including instances of incorrect reasoning and a tendency to fabricate facts. Several subsequent improvements have been made to CoT, such as the Tree-of-Thought approach (Yao et al. 2023) and Graph-of-Thought approach (Besta et al. 2023). Recently, the Selection-Inference (SI) algorithm has proposed introducing a separation between the inference and selection steps. This allows the model to reason based on the selected relevant facts, generate new conclusions, and iteratively update the facts. SI is reported to alleviate the problem of fabricating facts (Creswell, Shanahan, and Higgins 2023).

However, these methods often fail to consider the relationships among available facts, leading to the loss of important and relevant information during reasoning. In this paper, we introduce evidence chains to connect related facts and use

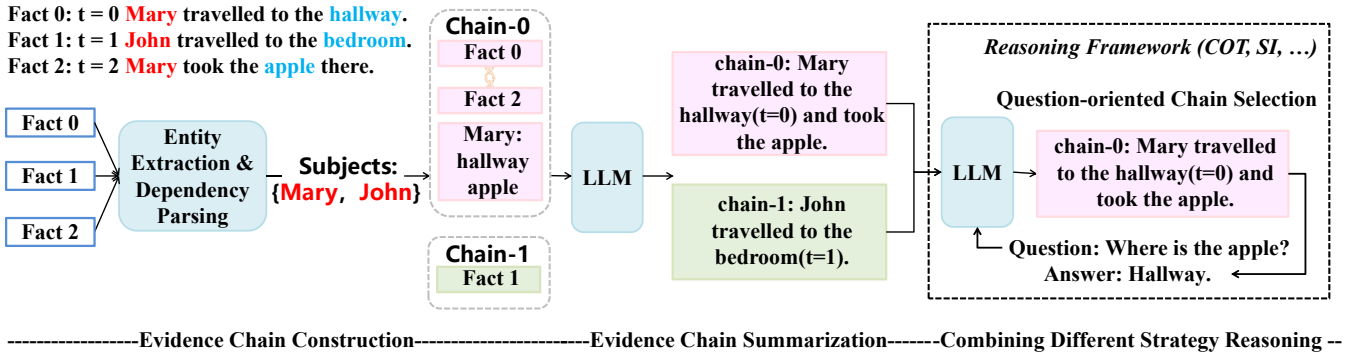

Figure 2: The main framework of the proposed MindMap approach.

evidence chains for supporting reasoning. Our method can be integrated with existing reasoning frameworks as a valuable plug-in component.

Methodology

Overview

We aim to solve natural language reasoning problems based on LLMs. Formally, given a set of facts $\mathcal{F} = \{f_1, f_2, ..., f_n\}$, a set of rules $\mathcal{R} = \{r_1, r_2, ..., r_m\}$ and a question q, we need to perform reasoning based on $\mathcal{F}$ and $\mathcal{R}$ to figure out an answer a to respond q, i.e., $a = \texttt{reasoning}(\mathcal{F}, \mathcal{R}, q)$, where `reasoning` represents specific reasoning framework. The rule set $\mathcal{R}$ can be empty in the task like reading comprehension via question answering.

We propose a framework called MindMap. The key idea is to introduce the concept of evidence chain to explore the structure within the given facts to support reasoning.

An **evidence chain** is defined as a group of facts, i.e., $c_s = \{f_1^s, f_2^s, ..., f_n^s\}$, where the facts $f_1^s, f_2^s, ..., f_n^s$ are all associated with the same subject s. For example, s can be a person, and c_s can be a sequence of events that involve the person s. Suppose there are k subjects $\mathcal{S} = \{s_1, ..., s_k\}$ in the set of facts $\mathcal{F}$. So $\mathcal{F}$ can be re-organized as a set of evidence chains $\mathcal{C} = \{c_1, c_2, ..., c_k\}$. Accordingly, the reasoning task can be represented as $a = \texttt{reasoning}(\mathcal{C}, \mathcal{R}, q)$.

Figure 2 illustrates the main framework of MindMap. It has 3 core components: *evidence chain construction*, *evidence chain summarization*, and *evidence chain utilization for inference*.

Evidence Chain Construction

Subject extraction Each evidence chain is associated with a subject. We first construct a set of subjects. Given a set of facts $\mathcal{F}$, we use the entity extraction and dependency parsing modules in the Stanford CoreNLP toolkit (Manning et al. 2014) to extract the entities which are subjects in the facts and group all of them into a set of subjects $\mathcal{S}$.

Subject-centric evidence chain After extracting the subjects, we build an evidence chain for every subject simply by grouping all facts containing a specific subject. If there exists temporal information, the facts in an evidence chain will be temporally sorted. Otherwise they would be sorted in the order as the original context.

Notice that a fact may contain multiple entities, so it can be involved in multiple evidence chains.

Evidence Chain Summarization

Given a context, there may be multiple evidence chains, the length of which could be short or long. To obtain a more concise description of an evidence chain, we introduce the evidence chain summarization module.

We propose an entity-centric summarization manner based on LLMs in a few-shot setting. We use instructions to guide a LLM to generate a summary covering main entities in each evidence chain. Below is an one-shot learning example of the prompt that is used for summarization.

```
TASK Instruction: Below are some
stories about people moving objects
between rooms. The facts are organized
as evidence chains, each of which
involves the same subject. Please Write
a summary for each chain and cover main
entities in each chain.
chain-0:
at t=0 Daniel went to the kitchen.
at t=1 Daniel picked up the apple
there.
at t=3 Daniel journeyed to the garden.
at t=9 Daniel travelled to the office.
at t=12 Daniel left the apple.
at t=13 Daniel went back to the
bedroom.
Entities about Daniel:
kitchen,apple,garden,
office,apple,bedroom
Summary:chain-0: Daniel passed through
the kitchen(t=0),garden(t=3),
office(t=9) in sequence with
apple.Then,he went to bedroom(t=13).
...
chain-to-be-summarized:
at t=2 Sandra travelled to the
```

```
bathroom.
at t=3 Sandra went to the bedroom.
at t=8 Sandra journeyed to the office.
at t=9 Sandra went back the bedroom.
Entities about Sandra:
bathroom,bedroom,office, bedroom
Summary:[Let the model generate]
```

In this way, given an evidence chain c, we can get its summary $\texttt{summarize}(c)$.

Evidence Chain Utilization for Inference

Utilizing evidence chains to organize facts for supporting reasoning changes the reasoning formulation from $a = \texttt{reasoning}(\mathcal{F}, \mathcal{R}, q)$ to $a = \texttt{reasoning}(\mathcal{C}, \mathcal{R}, q)$ or $a = \texttt{reasoning}(\{\texttt{summarize}(c), c \in \mathcal{C}\}, \mathcal{R}, q)$. During inference, the facts are organized as evidence chains rather than a sequence of facts.

The change of the formulation does not affect the reasoning framework, indicating that we can integrate MindMap in any reasoning framework.

MindMap enhanced CoT In the CoT setup, besides answering questions, explanations for the answers are also included to inspire the model to reason.

MindMap enhances CoT based on evidence chains. The key step is **question-oriented evidence chain selection** that we let LLMs choose evidence chains that can help answer the question. We use a few-demonstrations to guide the model to learn to integrate proper evidence chains for reasoning.

We show an example below with only one demonstration.

```
Demonstration 0
The evidence chains:
chain-0: ...
chain-1: ...
...
chain-t: ...
Question-oriented chain selection: the
question q0 can be inferred based on
the chain-0 and chain-3. Given the
summary that summarize(chain-0) and
summarize(chain-3), the answer should
be office.
Test 1
The evidence chains:
chain-0: ...
chain-1: ...
...
Question-oriented chain selection: the
question q1 can be inferred based on
[Let the model generate]
```

Here, q_0 and q_1 are specific questions which are included in the prompts to guide the model to select proper evidence chains for covering relevant and complete facts for reasoning, and filtering out irrelevant facts.

MindMap enhanced SI We also try to integrate MindMap with the selection-inference (SI) reasoning framework. SI divides the CoT reasoning framework into two parts: selection and inference. The selection module picks the most relevant facts and rules for the given question. The inference module feeds the selected facts and rules to the LLMs for deriving new conclusions. These two modules iterate, with each iteration adding a new conclusion to the evolving reasoning.

With MindMap, we select relevant evidence chains instead of choosing isolated facts in the selection stage. For the newly generated conclusions, we extract subjects from them. This procedure is similar to that of evichence chain construction. Then the newly generated conclusions are used for updating the evidence chains for supporting later SI iterations.

Experimental Settings

In this section, we will introduce the datasets, the baseline methods, and backbone model settings for evaluation.

Datasets

Our experiments are conducted using two challenging multi-step logical reasoning datasets.

- **bAbI** (Weston et al. 2016): Originating from the QA bAbI task, this dataset comprises a series of 20 tasks designed to evaluate reading comprehension via question answering. These tasks gauge understanding in various dimensions, including systems' ability to deduce answers through reasoning. We conduct experiments using the tasks 1-3 of bAbI, where to logically answer a question, 1-3 facts among a set of facts with temporal information are required.
- **ProofWriterOWA** (Tafjord, Dalvi, and Clark 2021): This synthetic dataset serves as a common benchmark for assessing logical reasoning, particularly when facts and rules are presented in natural language. The dataset is divided into subsets based on the steps of inferences, including 0, 1, 2, 3, and 5. Following previous work (Creswell, Shanahan, and Higgins 2023), we divide this dataset into two parts: ProofWriter-PUD, the answer set of which include *True*, *False*, and *Unknown*, and ProofWriter-PD, which excludes data labeled as *Unknown*.

Regarding the bAbI dataset, we use the full test set. For the ProofWriterOWA dataset, due to the heavy inference cost, we only use the first 1,000 samples in the test set.

Baseline Settings

We use the Vicuna-13B model (Chiang et al. 2023) in a few-shot setting as the **Standard** backbone model. We also consider two more advanced reasoning frameworks.

- **Chain-of-Thought (CoT)** (Wei et al. 2022): We also consider a CoT inspired baseline. It involves having LLMs generate a reasoning chain and leveraging its own reasoning capacity to provide interpretable model generation. A few examples in the prompt are used as CoT demonstrations.

Strategy	proofWriter-PD depth-0	depth-1	depth-2	depth-3	depth-5	proofWriter-PUD depth-0	depth-1	depth-2	depth-3	depth-5
standard	0.224	0.192	0.146	0.115	0.086	0.533	0.518	0.497	0.487	0.487
CoT	0.545	0.559	0.552	0.520	0.540	0.417	0.382	0.436	0.396	0.419
CoT+MindMap	0.711	0.700	0.683	0.646	0.589	0.469	0.450	0.476	0.443	0.441
relative impr.	30.5%	25.2%	23.7%	24.2%	9.07%	12.5%	17.8%	9.17%	11.9%	5.25%
SI	0.546	0.535	0.535	0.550	0.553	0.416	0.396	0.406	0.390	0.376
SI+MindMap	0.723	0.694	0.666	0.604	0.537	0.466	0.458	0.465	0.445	0.43
relative impr.	32.4%	29.7%	24.5%	9.82%	-2.89%	12.0%	15.7%	14.5%	14.1%	14.4%

Table 1: Prediction accuracy on the proofWriter-PD and proofWriter-PUD datasets.

Strategy	bAbI task-1	task-2	task-3
standard	0.675	0.369	0.181
CoT	0.607	0.467	0.281
CoT+MindMap	0.881	0.473	0.340
relative impr.	45.1%	1.3%	21.0%
SI	0.767	0.356	0.253
SI+MindMap	0.803	0.392	0.318
relative impr.	4.7%	10.1%	25.7%

Table 2: Prediction accuracy on the bAbI dataset.

- **Selection-Inference (SI)** (Creswell, Shanahan, and Higgins 2023): The SI framework alternates between selection and inference to generate a sequence of interpretable, casual reasoning steps leading to the final answer. During iterations, new conclusions can be generated and used for updating the fact set. SI runs 3 iterations for bAbI and 5 iterations for proofwriterOWA.

The standard, CoT and SI frameworks all adopt 5-shot setting, and use the same examples for constructing demonstration prompts. Details about prompt construction for these frameworks are based on the settings described in the appendix of the SI paper.

Results and Analysis

Main Results

Results on ProofWriterOWA Table 1 shows the ProofWriterOWA dataset results. In the ProofWriter-PD subset, both CoT and SI models did much better than the usual baseline. MindMap also improved results compared to many standard CoT and SI baselines. At depth-5, the benefit of MindMap was not as clear, possibly because the problems were too complex for the model to figure out the evidence chain accurately. But, we saw big improvements from depth-0 to 4. This is mainly because MindMap can combine and sum up various facts well, helping the model make more direct and precise decisions. The improvements in CoT and SI with MindMap have shown significant performance boosts, highlighting MindMap's effectiveness.

On the ProofWriter-PUD subset, MindMap also increases the performance of CoT and SI, but the relative improvements become smaller compared with that on the ProofWriter-PD subset. We can also see that the advanced reasoning frameworks cannot beat the standard baseline. We find that the standard baseline prefers to predict the Unknown label, while the 'Unknown' label accounts for about 46% in the ProofWriter-PUD dataset. So the standard baseline may benefit from the imbalanced label distribution and outperforms CoT, SI with or without MindMap. Moreover, predicting Unknown should be more difficult, since in the ProofWriterOWA dataset, Unknown means it is not possible to prove something is True or False. For True and False examples, there would be a reasoning trace leading to the final answer, while there would be no trace for the Unknown examples. A model with a larger parameter scale may have a better ability in handling this complex problem.

Results on bAbI dataset Table 2 illustrates the significant performance improvement on the bAbI dataset achieved through MindMap. Specifically, MindMap facilitated average relative improvements of 22.5% in CoT and 13.5% in SI. The notable advancements in both CoT and SI, particularly within the bAbI-3 subset, are particularly impressive, showing MindMap's strength in multi-step reasoning. This highlights how much MindMap helps in improving reasoning and understanding skills.

In summary, across both datasets, MindMap has proven to effectively enhance the performance of existing reasoning frameworks, such as CoT and SI, particularly in scenarios involving multi-step reasoning.

Ablation Study

We conduct analyses on the bAbI dataset to quantify the impact of the manipulating evidence chains within MindMap. Table 3 shows the ablation study results within two reasoning framework CoT and SI.

SEC represents the proposed subject-centric evidence chain construction, which are then directly used for inference without evidence summarization. MindMap combines SEC and evidence chain summarization for inference. SEC shows big improvements in simple reasoning tasks, like

COT based on the original context:
at t=0 Daniel moved to the office.
at t=1 Daniel travelled to the bedroom.
at t=2 John took the milk there.
at t=3 John travelled to the garden.
at t=4 Sandra moved to the office.
at t=5 Daniel went back to the office.
Question:Where is the milk?
Inference: at t=2 John took the milk there.We know that **at t=5 Daniel went back to the office.**Therefore, milk is in the **office**.

MindMap based on the evidence chain summaries:
chain-0: Daniel passed through the office(t=0),bedroom(t=1) in sequence.
chain-1: John passed through the bedroom(t=2),garden(t=3) with milk in sequence.
chain-2: Sandra passed through the office(t=4) in sequence.
Question:Where is the milk?
Inference: According to **chain-1**: John passed through the bedroom(t=2),garden(t=3) with milk in sequence.Therefore,milk was in garden(t=3).

(a) An example from the bAbI dataset.

COT based on the original context:
The bald eagle is cold. The bald eagle sees the bear. The dog is rough. The lion likes the bald eagle. The lion sees the bear. The bear is big. The bear is rough. The dog visits the bear. The lion is young. The lion sees the bald eagle. The lion visits the bear. The lion visits the dog..
Rule: If something likes the bald eagle and sees the bear then it sees the dog. red things are rough. If someone sees the bear and visits the bear then it sees the tiger.
Question:The bald eagle sees the dog
Inference: The bald eagle sees the bear. We know that If someone sees the bear and visits the bear then it sees the tiger. Therefore, the bald eagle sees the tiger, which means answer is Unknown.

MindMap based on the evidence chain summaries:
chain-0: The bald eagle is cold,sees the bear,likes the bald eagle,sees the bald eagle.
chain-1: The bear is big,rough,visited by the dog,lion.
chain-2: The dog is rough,visits the bear.
chain-3: The lion is young,likes the bald eagle,sees the bald eagle,sees the bear,visits the bear,visits the dog
Rule: If something likes the bald eagle and sees the bear then it sees the dog. red things are rough. If someone sees the bear and visits the bear then it sees the tiger.
Question:The bald eagle sees the dog
Inference: According to **chain-0:** The bald eagle is cold,sees the bear,likes the bald eagle,sees the bald eagle.And according the rules: If something likes the bald eagle and sees the bear then it sees the dog. Therefore, The bald eagle sees the dog, which means answer is True.

(b) An example from the ProofWriterOWA dataset.

Figure 3: The traces of CoT and MindMap enhanced CoT for solving examples from the bAbI and ProofWriterOWA dataset for the qualitative analysis.

Strategy	bAbI task-1	bAbI task-2	bAbI task-3
standard	0.675	0.369	0.181
CoT	0.607	0.467	0.281
CoT+SEC	0.789	0.462	0.26
CoT+MindMap	**0.881**	**0.473**	**0.340**
CoT+Auto-MindMap	0.372	0.233	0.183
SI	0.767	0.356	0.253
SI+SEC	**0.860**	**0.451**	0.269
SI+MindMap	0.803	0.392	**0.318**
SI+Auto-MindMap	0.364	0.218	0.166

Table 3: The ablation study results. SEC: subject-centric evidence chain construction; Auto-MindMap: using the model to build and summarize evidence chains via instruction tuning; MindMap: using the summaries of evidence chains for inference.

task-1, and it's even more effective in the SI framework. This means that creating evidence chains helps the SI model pick better facts for reasoning. Also, MindMap significantly improves complex tasks, like task-3, involving multi-step reasoning. This indicates that evidence chain summarization can enhance inference by compressing information and emphasizing important parts in each evidence chain, which is important for multi-step reasoning, when multiple facts are involved and both useful and noisy information are mixed. For simple reasoning tasks under the SI framework, evidence chain summarization seems unnecessary.

We also try to let the model automatically build evidence chain through instruction learning, called Auto-MindMap. However, Auto-MindMap often results in a decrease in performance, as this automatic process may add an extra burden to the model. The comparison between MindMap and Auto-MindMap also indicates that combining traditional NLP tools with LLMs is a simple and effective way to incorporate linguistic motivated structures.

Qualitative Analysis

We further conduct a qualitative analysis by comparing the behaviors of CoT and MindMap enhanced CoT. Figure 3 shows two comparison examples on the bAbI and ProofWriterOWA datasets respectively.

As shown in Figure 3a, the example from the bABI dataset shows a reasoning problem, requiring to integrate two facts to infer the correct answer. CoT selects two unrelated facts and one key fact involving John is missed during inference. So it makes a wrong prediction. In contrast, MindMap correctly selects an evidence chain, which summarizes the series of activities of John. This summary leads the model to make a correct prediction.

Figure 3b shows an example from the ProofWriterOWA dataset. We can also see that the MindMap can provide con-

cise summaries for related facts, which have a better matching with the rules and help improve the inference performance.

Based on the examples, we can see that one advantage of evidences is grouping related facts. This advantage can often avoid missing important and relevant information. We conduct an analysis of the coverage of supporting facts on the task-2 and task-3 in bAbI dataset by matching the temporal id, e.g., $t = 0$, $t = 1$, between the predictions and the gold supporting fact reference. MindMap can cover 69% and 47% supporting facts on task-2 and task-3 resepctively, while the same statistics for CoT are 48% and 23%. The analysis confirms that the advantage of MindMap may be not so important for simple reasoning but should be helpful for multi-step reasoning.

In this work, we use subjects or entities to group facts. This is consistent with our intuition, since people also often infer some conclusions by tracking and analyzing someone's behavior or related events.

Error Analysis

In our study, we conducted an error analysis on a sample of outputs generated by MindMap. The identified errors broadly fall into the following categories:

- **Summarization Errors**: These errors originate from the evidence chain summarization module and manifest in two forms: the omission of crucial information and the introduction of inaccurate or fabricated details (hallucinations) during the summarization process. In our analysis of the bAbI dataset following LLMs' summarization, correct evidence chains were identified 68.9% of the time on average, with a reduced accuracy of 63.8% specifically for task-2. This lower rate of accurate evidence chain identification in task-2 accounts for the less pronounced improvements observed with our method in this particular task.
- **Evidence Chain Selection Errors**: Challenges may arise in selecting pertinent evidence chains for reasoning, even when LLMs produce informative and accurate summaries. Situations that require consideration of multiple evidence chains pose a particular challenge, often leading to errors in chain selection.
- **Hallucination errors during inference**: When the correct evidence chain is selected, LLMs may also hallucinate during paraphrasing the summary of an evidence chain, thereby introducing information bias and inaccurate inferences. This includes modifying or fabricating evidence, ignoring key facts, and incorrectly aligning questions and answers, all of which ultimately lead to incorrect conclusions. For example, the question in this example is the squirrel visits the cow, and the answer given by LLM is *chain-3: The squirrel is big. And according to the rules: If something is round then it visits the cow. We know that the squirrel is round . Therefore, the squirrel visits the cow.* LLM selects the correct chain. The model needs to use two rules to get the final answer. However, in the subsequent reasoning process, LLMs fabricated *the squirrel is round* in order to take shortcuts.
- **Model inference errors** : The remaining errors can be attributed to the model's inference capabilities. These errors still occur despite choosing the correct chain of evidence. Even after the model has reached the correct conclusion, errors still occur when obtaining the final answer. For example, *Therefore, Dave is nice. The question is "Dave is nice."*, which means answer is false. This is due to insufficient capabilities of the model.

In summary, our error analysis identified key factors that contribute to reasoning errors. To reduce these errors and improve overall performance, some solutions are proposed. These include enhancements to *evidence chain summarization* to ensure more accurate summaries to significantly reduce error rates and using larger models to increase the LLM's efficiency in complex reasoning tasks.

Conclusion

In this paper, we explore how structuring available facts can enhance reasoning capabilities. The proposed MindMap approach organizes these facts into evidence chains, seamlessly integrating with existing reasoning frameworks such as CoT and SI. Experiments are conducted on two complex multi-step reasoning datasets. As shown in the results, both CoT and SI, when augmented with MindMap, can achieve significant improvements, particularly in multi-step reasoning tasks. Underscoring the importance of systematic organization of available facts, our approach demonstrates superior performance to CoT and SI in recalling supporting facts. The simplicity of constructing evidence chains suggests that integrating traditional natural language processing tools with LLMs could effectively tackle complex reasoning challenges by leveraging linguistically motivated structures.

However, our work still has some limitations. First, due to computational resource constraints, our evaluation was conducted using a LLM with 13B parameters. In future research, exploring a diverse range of models with varying sizes of parameters could prove beneficial. Second, though intuitive and heuristic, our approach to constructing subject-centric evidence chains has been successfully used for evaluations on synthetic datasets. Nevertheless, practical applications could encounter challenges such as pronoun resolution, which we aim to address by enhancing the model's capabilities in autonomously constructing evidence chains. Third, although constructing evidence chains has proven effective for organizing facts, exploring alternative methods remains a valuable avenue. Employing semantic information or knowledge graphs could offer a more comprehensive approach to organizing facts, thereby improving the reasoning and judgment capabilities of LLMs.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.62376166, No.62306188), the National Key Research and Development Programme of China (No.2022YFC3303504), and the Science Research & Development Program of Beijing Municipal Education Commission (No.KM202010028004).

References

Bang, Y.; Cahyawijaya, S.; Lee, N.; Dai, W.; Su, D.; Wilie, B.; Lovenia, H.; Ji, Z.; Yu, T.; Chung, W.; et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *ArXiv preprint*, abs/2302.04023.

Besta, M.; Blach, N.; Kubicek, A.; Gerstenberger, R.; Gianinazzi, L.; Gajda, J.; Lehmann, T.; Podstawski, M.; Niewiadomski, H.; Nyczyk, P.; et al. 2023. Graph of thoughts: Solving elaborate problems with large language models. *ArXiv preprint*, abs/2308.09687.

Betz, G.; Voigt, C.; and Richardson, K. 2021. Critical Thinking for Language Models. In *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*, 63–75. Groningen, The Netherlands (online): Association for Computational Linguistics.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Chambers, N.; and Jurafsky, D. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, 789–797.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H. W.; Sutton, C.; Gehrmann, S.; et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240): 1–113.

Clark, P.; Tafjord, O.; and Richardson, K. 2020. Transformers as Soft Reasoners over Language. In Bessiere, C., ed., *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, 3882–3890. ijcai.org.

Creswell, A.; Shanahan, M.; and Higgins, I. 2023. Selection-Inference: Exploiting Large Language Models for Interpretable Logical Reasoning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Dai, D.; Sun, Y.; Dong, L.; Hao, Y.; Ma, S.; Sui, Z.; and Wei, F. 2023. Why Can GPT Learn In-Context? Language Models Secretly Perform Gradient Descent as Meta-Optimizers. In Rogers, A.; Boyd-Graber, J. L.; and Okazaki, N., eds., *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, 4005–4019. Association for Computational Linguistics.

Dong, Q.; Li, L.; Dai, D.; Zheng, C.; Wu, Z.; Chang, B.; Sun, X.; Xu, J.; and Sui, Z. 2023. A survey for in-context learning. *ArXiv preprint*, abs/2301.00234.

Goertzel, B.; and Pennachin, C. 2007. *Artificial general intelligence*, volume 2. Springer.

Han, S.; Schoelkopf, H.; Zhao, Y.; Qi, Z.; Riddell, M.; Benson, L.; Sun, L.; Zubova, E.; Qiao, Y.; Burtell, M.; et al. 2022. Folio: Natural language reasoning with first-order logic. *ArXiv preprint*, abs/2209.00840.

Katz, D. M.; Bommarito, M. J.; Gao, S.; and Arredondo, P. 2023. Gpt-4 passes the bar exam. *Available at SSRN 4389233*.

Kazemi, M.; Kim, N.; Bhatia, D.; Xu, X.; and Ramachandran, D. 2023. LAMBADA: Backward Chaining for Automated Reasoning in Natural Language. In Rogers, A.; Boyd-Graber, J. L.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, 6547–6568. Association for Computational Linguistics.

Liu, H.; Ning, R.; Teng, Z.; Liu, J.; Zhou, Q.; and Zhang, Y. 2023. Evaluating the logical reasoning ability of chatgpt and gpt-4. *ArXiv preprint*, abs/2304.03439.

Liu, J.; Cui, L.; Liu, H.; Huang, D.; Wang, Y.; and Zhang, Y. 2020. LogiQA: A Challenge Dataset for Machine Reading Comprehension with Logical Reasoning. In Bessiere, C., ed., *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, 3622–3628. ijcai.org.

Manning, C. D.; Surdeanu, M.; Bauer, J.; Finkel, J. R.; Bethard, S.; and McClosky, D. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, 55–60.

Min, S.; Lyu, X.; Holtzman, A.; Artetxe, M.; Lewis, M.; Hajishirzi, H.; and Zettlemoyer, L. 2022. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 11048–11064. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Muennighoff, N.; Wang, T.; Sutawika, L.; Roberts, A.; Biderman, S.; Scao, T. L.; Bari, M. S.; Shen, S.; Yong, Z. X.; Schoelkopf, H.; Tang, X.; Radev, D.; Aji, A. F.; Almubarak, K.; Albanie, S.; Alyafeai, Z.; Webson, A.; Raff, E.; and Raffel, C. 2023. Crosslingual Generalization through Multitask Finetuning. In Rogers, A.; Boyd-Graber, J. L.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, 15991–16111. Association for Computational Linguistics.

Nori, H.; King, N.; McKinney, S. M.; Carignan, D.; and Horvitz, E. 2023. Capabilities of gpt-4 on medical challenge problems. *ArXiv preprint*, abs/2303.13375.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744.

Peng, B.; Li, C.; He, P.; Galley, M.; and Gao, J. 2023. Instruction tuning with gpt-4. *ArXiv preprint*, abs/2304.03277.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Rae, J. W.; Borgeaud, S.; Cai, T.; Millican, K.; Hoffmann, J.; Song, F.; Aslanides, J.; Henderson, S.; Ring, R.; Young, S.; et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *ArXiv preprint*, abs/2112.11446.

Saparov, A.; and He, H. 2023. Language Models Are Greedy Reasoners: A Systematic Formal Analysis of Chain-of-Thought. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Silver, D.; Singh, S.; Precup, D.; and Sutton, R. S. 2021. Reward is enough. *Artificial Intelligence*, 299: 103535.

Tafjord, O.; Dalvi, B.; and Clark, P. 2021. ProofWriter: Generating Implications, Proofs, and Abductive Statements over Natural Language. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 3621–3634. Online: Association for Computational Linguistics.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *ArXiv preprint*, abs/2302.13971.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35: 24824–24837.

Weston, J.; Bordes, A.; Chopra, S.; and Mikolov, T. 2016. Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks. In Bengio, Y.; and LeCun, Y., eds., *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Yang, J.; Jin, H.; Tang, R.; Han, X.; Feng, Q.; Jiang, H.; Yin, B.; and Hu, X. 2023. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ArXiv preprint*, abs/2304.13712.

Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; Griffiths, T. L.; Cao, Y.; and Narasimhan, K. 2023. Tree of thoughts: Deliberate problem solving with large language models. *ArXiv preprint*, abs/2305.10601.

Zhang, H.; Li, L. H.; Meng, T.; Chang, K.; and den Broeck, G. V. 2023. On the Paradox of Learning to Reason from Data. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*, 3365–3373. ijcai.org.

Zhou, D.; Schärli, N.; Hou, L.; Wei, J.; Scales, N.; Wang, X.; Schuurmans, D.; Cui, C.; Bousquet, O.; Le, Q. V.; and Chi, E. H. 2023. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.