Enhancing Knowledge through Revisable Chain-of-Thought for Commonsense Question Answering

Anonymous ACL submission

Abstract

Large Language Models (LLMs) are effective at natural language reasoning, but still struggle with answering commonsense questions that require implicit knowledge of the world. LLMs rely on knowledge learned through training, which can be limited to specific domains and may lack inductive abstraction, resulting in hallucinations and inaccurate knowledge. To alleviate these, recent research integrates external knowledge sources (e.g., fine-tuning, self-correction, retrieval enhancement, and chain-of-thought (CoT)). While CoT reveals specific incorrect knowledge in LLMs, it lacks abstraction and is uneasy to be revised. In this paper, we propose a revisable three-step CoT framework, categorizing knowledge into abstract meta-knowledge and concrete instantiated knowledge. Meanwhile, we use transfer knowledge to address the logical form sensitivity of LLMs. Furthermore, we propose online revision by teacher models and offline revision with knowledge base. We propose an antisense retrieval method to check if the newly generated knowledge contradicts any existing knowledge in the knowledge base to avoid retrieving meta-knowledge that is not relevant to the problem. The experimental results on the Winogrande dataset have corroborated the efficacy of our proposed method. We revised the meta-knowledge of GPT-3.5 with GPT-4, which enhanced the accuracy from 68.11% to 73.64%, an improvement of 5.53 percentage points.

1 Introduction

Large language models (LLMs) (e.g. GPT-4 OpenAI (2023)) have demonstrated strong capabilities in dealing with natural language reasoning (NLR) Yu et al. (2023); Lin et al. (2023) problems, where reasoning refers to the process of drawing logical inferences or conclusions from given information. Commonsense Question Answering (CQA) Zhang et al. (2024); Talmor et al. (2021); Huang et al. (2019) is a subfield of NLR that requires the understanding and application of implicit world knowledge(e.g., spatial relations, social conventions and scientific facts, etc.) Branco et al. (2021); Zhou et al. (2021) .

Effective utilization of knowledge in LLMs is crucial Yin et al. (2023). The knowledge embedded in LLMs is known as parameterised knowledge Luo et al. (2023), acquired through extensive data training within the neural network's weights. Such parameterised knowledge in LLMs includes widely accepted fundamental facts and concepts. When answering commonsense questions, parameterized knowledge faces two challenges: constrained training corpus leading to domain limitations; and a lack of inductive abstraction, where models fail to lift and drop. These limitations can lead to hallucinations, resulting in the generation of erroneous or fictitious information without a clear basis.

To alleviate hallucinations, recent developments focus on integrating external knowledge sources but are limited in commonsense question answering: (a) Fine-tuning Shen et al. (2023); Fu et al. (2023) improves model performance through additional training on specific datasets, but identifying effective knowledge is still challenging (i.e. incorrect parameterised knowledge and unseen knowledge). (b) Self-revision Madaan et al. (2023); Guo et al. (2023) allows models to self-assess and correct potential errors before output, but it is ineffective when the model is uncertain of the correct answer Huang et al. (2023). (c) Retrieval augmentation Asai et al. (2023); Shuster et al. (2021) strengthens models by retrieving from external information sources. Representing the knowledge and its availability to LLM are the key challenges. (d) Chain-of-thought(CoT) Wei et al. (2022); Zhou et al. (2023) guides models to generate reasonable results through a linear chain pattern of reasoning and interpretation. By observing solving how LLMs solve commonsense questions through CoT,

Question: Jessica lost to Jennifer at darts at the fair, so _ won the goldfish in a bag. (A) Jessica (B) Jennifer
Regular CoT: To solve this problem, we need to identify who won the goldfish in a bag based on the outcome of a darts game between Jessica and Jennifer at the fair. 1. We are given that Jessica and Jennifer played darts. 2. The sentence states that Jessica lost to Jennifer. 3. Since Jennifer won the game of darts, she was the victor. 4. The winner of the darts game won the goldfish in a bag. 5. Therefore, since Jennifer won the darts game, she also won the goldfish. Based on this chain of thought, the answer is (B) Jennifer.
Revisable CoT: • **Meta-knowledge:** If X wins against Y, then X gets a reward. • **Transfer knowledge:** X wins against Y, so X gets a reward. (If P, then Q. $\rightarrow$ P, so Q.) • **Instantiated knowledge:** X=Jennifer, Y=Jessica, X wins against Y = Y lost to X, reward = goldfish, _=X=Jennifer

Table 1: Regular CoT, meta-knowledge, transfer knowledge, instantiated knowledge of an example.

we find that some commonsense knowledge of LLMs is inaccurate. However, CoT can reveal specific wrong knowledge which is revisable.

Revising regular CoT knowledge is challenging because it lacks abstraction and transferability (see reguar CoT in Table 1). Structuring the CoT and categorising the knowledge it contains, in turn, can help to identify and revise specific types of knowledge. Inspired by the discursive logic of "rising from the abstract to the concrete", we believe that we need to identify the basic principles or laws of the problem in various scenarios, and then concretise the abstract concepts into context-specific instances. In addition, since LLMs are sensitive to logical forms of knowledge, such as the "curse of reversal", they also need to deal with the restatement or reversal of basic principles.

In this paper, we propose a revisable three-step CoT framework on commonsense question answering tasks for enhancing knowledge. And we conduct an empirical study on revising knowledge. We emphasize the following research questions:

- RQ1. Can we improve the performance of commonsense question answering by revising the knowledge in the chain of thought? We classify the knowledge in commonsense question answering into meta-knowledge, transfer knowledge and instantiated knowledge, and revise them progressively. Experiments prove that revising meta-knowledge is the most critical.

- RQ2. Can LLMs revise themselves without external help? If not who can revise them and how? We find that model self-revision fails to deliver gains while performance can be enhanced by using more powerful models or humans as teacher models. When teacher models are unavailable or expensive, we use a self-revision method via a knowledge base (KB). Correcting models online on a case-by-case basis is better than using offline models for KB retrieval.

- RQ3. Which is more effective for knowledge revision with KB: (a) determine whether newly generated knowledge conflicts with KB (i.e., antonymic retrieval), or (b) retrieve knowledge from KB without new knowledge generation? We find (a) is better than (b) because direct retrieval may introduce irrelevant knowledge. Furthermore, it demonstrates that LLMs perform better in determining knowledge contradictions than in selecting appropriate knowledge.

To sum up, our contributions are three-fold:

(1) We propose a revisable three-step CoT framework for enhancing knowledge in commonsense question answering tasks. We categorise knowledge into meta-knowledge (abstract) and instantiated knowledge (concrete), enhancing knowledge transferability and ease of revision. Transfer knowledge enables flexible application, reducing sensitivity of LLMs' logical form.

(2) We further propose online revision by teacher models and offline revision with knowledge bases, and our antonymic retrieval outperforms conventional retrieval.

(3) Experimental results on Winogrande show that our method is effective in correcting commonsense knowledge and improve the accuracy.

2 Related Work

Chain-of-thought (CoT) reasoning Chu et al. (2023) involves models explicitly outputting intermediate reasoning steps before the final answer. It enhances LLMs' performance on complex reasoning tasks and interpretability. We introduce constructing, structuring, and enhancing the CoT.

CoT construction are categorised into three main methods: manual, automatic and semi-automatic. Manual construction Wei et al. (2022); Gao et al. (2023) relies on complete manual annotation, which yields high-quality results and is particularly beneficial for learning with fewer samples but faces larger labour costs and cross-task migration challenges. In contrast, automatic construction eliminates human intervention. It generates inference chains via both Zero-shot CoT Kojima et al. (2022) and Auto CoT Zhang et al. (2022), which reduces labour costs and facilitates cross-task migration. Still, its performance may be limited by the lack of high-quality annotation and is prone to logical or factual errors. The semi-automatic construction Shum et al. (2023) method uses a few high-quality manually labelled "seed samples" Pitis et al. (2023) to generate reasoning chains through automatic expansion, balancing human cost and reasoning performance.

CoT structures are varied, with the most primitive structure being a chain that describes intermediate reasoning steps in natural language Wei et al. (2022). (Gao et al., 2023) uses procedural language instead of natural language, while Long (2023) introduces a tree structure to tackle complex tasks. Graph structures Besta et al. (2023), on the other hand, can handle complex tasks efficiently due to their complex topology and ring structures. ResPrompt Jiang et al. (2023) connects reasoning steps with residual connections in the prompt text, building graph structures.

CoT enhancement approach is a key strategy for addressing LLMs' hallucinatory. Validation and refinement-based approaches (e.g. Verify-CoT Ling et al. (2023) and DIVERSE Li et al. (2023b)) ensure consistency through calibration of reasoning steps and deductive reasoning while introducing knowledge from internal and external sources to reinforce factual accuracy. Least-to-Most Zhou et al. (2022) and Successive Prompting Dua et al. (2022) decompose complex problems into manageable sub-problems. Chain-of-Knowledge Li et al. (2023a) introduces exogenous knowledge to provide up-to-date information for the model. Sorting or voting-based methods (e.g., Self-Consistency Wang et al. (2022)) optimise the inference process by multiple sampling and result integration to reduce errors due to randomness.

Ours is semi-automatically constructed through a three-step revisable CoT framework. It progressively specifies the meta-knowledge, transfer knowledge, and instantiated knowledge used in new problems. It also self-revises by introducing a knowledge base that can be either a larger model or a human construct.

3 Methodology

3.1 Design of Revisable Chain-of-Thought (RCoT)

We classify the knowledge in commonsense question answering into meta-knowledge, transfer knowledge and instantiated knowledge, and revise them progressively.

3.1.1 Meta-knowledge(MK) and Instantiated knowledge(IK)

Meta-Knowledge(MK) is the abstract, simple and correct general knowledge that you need to master when answering questions, and many questions may be solved by the same Meta-Knowledge. Instantiated Knowledge(IK) is the knowledge that corresponds the abstract elements of meta-knowledge to the concrete content of the problem to solve the concrete problem.

We design a Meta-Knowledge pattern in the form of "If P, then Q," where P and Q represent the premise and conclusion, respectively. Table 2 presents several typical instances of meta-knowledge. Some symbols and concepts within P and Q need to be instantiated, which we refer to as slots. For example, in meta-knowledge "If X wins against Y, then X gets a reward," X and Y could be two individuals, two teams, two companies, or two countries. The term "win" could refer to victory in a game, a sports competition, a business rivalry, or a war, while "reward" could signify a prize, market share, honour, or war spoils, among other things.

The evaluation of meta-knowledge includes correctness, relevance and abstractness. Correctness indicates whether the meta-knowledge is correct or

not. If the meta-knowledge is wrong, it doesn't matter whether the result is correct or not. Relevance indicates whether meta-knowledge is applicable to answering the question that needs to be addressed. Meta-knowledge is of no value if it cannot answer the question. Abstractness indicates whether the meta-knowledge is reasonably abstract, meaning that it can be used to solve similar problems and can also be effectively instantiated for specific problems.

3.1.2 Transfer knowledge

Transfer Knowledge(TK) is used to transform meta-knowledge into another form that is more suitable for the problem at hand, requiring the use of logical knowledge and linguistic expertise. The purpose of transforming linguistic knowledge is to better adapt to specific problems, thereby more effectively mapping the slots in the meta-knowledge to the actual issues.

There are three aspects in which the various forms of meta-knowledge differ: first, the sequence of the premise P and the conclusion Q in the sentence. The premise P can precede the conclusion Q, or the conclusion Q can come before the premise P. Second, whether there is a negation in the premise P and the conclusion Q, which combines to create four possibilities. Third, the sentence components that connect the premise P and the conclusion Q. Table 3 shows typical examples of transfer knowledge.

The evaluation metrics for transfer knowledge encompass correctness and applicability. Correctness pertains to the assessment of whether the transformation of meta-knowledge maintains equivalence. For example, given meta-knowledge in the form of "If P, then Q", a correct transformation would be "not Q, so not P", while "not Q, so P" would be incorrect. Applicability refers to the degree to which the transformed meta-knowledge aligns with the syntactic structure of the target problem.

3.2 Knowledge Revision Method

If Model M_T performs significantly better than Model M on a Commonsense Question Answering task, this paper speculates that M_T performs better than M on at least one, or all three, of the revisable chain-of-thought solutions in terms of meta-knowledge, transfer knowledge, and instantiated knowledge. The chain-of-thought of M can be modified with M_T, which we call the teacher model.

3.2.1 Online Revision by Teacher Models

The *Online Revision by Teacher Models* (RTM) method employs a teacher model M_T to iteratively refine the chain-of-thought in model M, specifically targeting the knowledge components Meta-Knowledge(MK), Transfer Knowledge(TK), and Instantiated knowledge(IK). The teacher model M_T can be a more capable language model or even a human. For instance, GPT-4 serves as the teacher model for GPT-3.5, while humans act as the teacher model for GPT-4.

The teacher model possesses the capability either to revise the knowledge embedded within the model or to regard the model's inherent knowledge as accurate, thus not requiring revision. Algorithm 1 provides a simplified description of the RTM method, omitting the details of revisions to TK and IK. The revision process for TK and IK is identical.

In the algorithm 1, IsCorrect(mk) and IsMatch(q, mk) respectively indicate whether mk is correct and whether mk matches the question q. These can be determined by the M_T model or by a specialized model.

Algorithm 1 Online Revision by Teacher Models

Input:
the question q, model M, teacher model M_T.
Output:
The output is the revision sequence $S0, S1, S2$.

$S0 : M(q) = (mk, tk, ik, a)$
if IsCorrect(mk) and IsMatch(q, mk) **then**
  $S1 : mk' = mk$
else
  $S1 : mk' = M_T(q, mk)$
end if
$S2 : M(q, mk') = (tk', ik', a')$
Output the sequence $S0, S1, S2$.

3.2.2 Offline Revision with Knowledge Base

When the teacher model is not available, or is expensive to use, such as when the teacher model is human, we use a modified method of using the teacher model knowledge offline, which is called *Offline Revision with Knowledge Base*(RKB) in this paper.

As mentioned in Section 3.1, since multiple problems may rely on the same meta-knowledge for resolution, the meta-knowledge required for a problem

Question	**Options**
Jessica lost to Jennifer at darts at the fair, so _ won the goldfish in a bag.	(A) Jessica (B) Jennifer
Meta-knowledge	**Slot**
If X wins against Y, then X gets a reward.	X,Y,win, reward
Transfer knowledge	**Form**
X wins against Y, so X gets a reward.	If P, then Q. → P, so Q.
Instantiated knowledge	
X=Jennifer, Y=Jessica, X wins against Y = Y lost to X, reward = goldfish, _=X=Jennifer	
Question	**Options**
Michael had a cat but Nelson didn't have any pets because _ had little allergies.	(A) Michael (B) Nelson
Meta-knowledge	**Slot**
If X is allergic, then X does not have a pet.	X,pet
Transfer knowledge	**Form**
X have a pet because X is not allergic.	If P, then Q. → not Q because not P.
Instantiated knowledge	
X=Michael, pet=cat, _=X=Michael	

Table 2: Examples of meta-knowledge, transfer knowledge,and knstantiated knowledge

Category		**Sentence Form**		
P	**Q**	Because P, so Q.	P; therefore, Q.	Q, as a result of P.
P	**not Q**	P, but not Q.	Even though P, not Q.	not Q, although P.
not P	**Q**	Although not P, Q.	Even though not P, Q.	Q, even though not p.
not P	**not Q**	not Q, because not P.	not Q, not P.	Since not P, then not Q.

Table 3: Hierarchical Classification of Transfer knowledge. The "Sentence Form" in the table represents an incomplete list of examples.

might have already been provided by the teacher model when solving similar problems in the past and may exist within the meta-knowledge base.

Despite the accuracy of the knowledge in the meta-knowledge base being ensured by the teacher model, finding the appropriate meta-knowledge for new questions from the vast meta-knowledge base is challenging. To reduce errors caused by irrelevant meta-knowledge, we adopt the most conservative strategy: if there is meta-knowledge in the knowledge base that contradicts the model's meta-knowledge, we can ascertain that the model's meta-knowledge is incorrect, while also ensuring relevance. For details, see Algorithm 2.

In Algorithm 2, NegateP and NegateQ represent the negations of the premise and conclusion, respectively, of the meta-knowledge. This process produces the two antonymous meta-knowledge mk_{n1} and mk_{n2}. The generation and retrieval of antonymous meta-knowledge can be accomplished by model M itself or by a dedicated model designed

Algorithm 2 Offline Revision with Knowledge Base

Input:
the question q, model M, Meta-Knowledge Base MKB.

Output:
The output is the revision sequence $S0, S1, S2$.

$S0 : M(q) = (mk, tk, ik, a)$
$NegateP(mk) = mk_{n1}, NegateQ(mk) = mk_{n2}$
if $\exists mk_b \in$MKB, $mk_b \approx mk_{n1} \vee mk_b \approx mk_{n2}$ **then**
  $S1 : mk' = mk_b$
else
  $S1 : mk' = mk$
end if
$S2 : M(q, mk') = (tk', ik', a')$
Output the sequence $S0, S1, S2$.

for this purpose.

4 Experiments

4.1 Winogrande

To validate our approach, we conducted relevant experiments on the Winogrande Sakaguchi et al. (2021) dataset. Winogrande takes inspiration from winograd schemas Levesque et al. (2012) to create a large-scale dataset of coreference resolution problems requiring both physical and social common sense. Each question presents a sentence with a blank where a pronoun might be and two options to fill it. The Winogrande dataset is divided into training, development, and test sets, containing 9,248, 1,267, and 1,767 examples. Since the test set does not provide answers, we carried out our experiments on the development set.

For examples from the Winogrande dataset, refer to the three questions in Table 2.

4.2 Experimental Settings

In this paper, we employ GPT-3.5 and GPT-4 as the instruction-following models for our study, with the model names designated as gpt-3.5-turbo-16k and gpt-4-1106-preview, respectively. All other parameters are maintained at their default settings. Due to the high cost of human experts as a teacher model and knowledge base source, we use GPT-4 to revise the response of GPT-3.5.

In the experiments on Online Revision by Teacher Models(RTM), GPT-4 is utilized as the teacher model for GPT-3.5. In the experiments on Offline Revision with Knowledge Base(RKB), this paper has a subset of instances from the Winogrande training set answered by GPT-4 in an RCoT method, from which 5,000 meta-knowledge entries are extracted to form a database. We test two meta-knowledge retrieval models: the all-mpnet-base-v2 vectorized retrieval Reimers and Gurevych (2019) and the GPT-4 batch retrieval. The all-mpnet-base-v2 is a language representation model that vectorizes the meta-knowledge of GPT-4 and the counter-knowledge of GPT-3.5, and then retrieves them using cosine similarity. We input meta-knowledge into GPT-3.5, utilizing instructions and eight examples to prompt GPT-3.5 to generate two sets of meta-knowledge, one with negation applied solely to the premise and the other with negation applied solely to the conclusion. We then input the two negated forms of meta-knowledge into the all-mpnet-base-v2 model for vector retrieval.

We input meta-knowledge into GPT-3.5 by employing directives and eight examples, enabling GPT-3.5 to generate two antisense meta-knowledge representations: one that negates the premise and another that negates the conclusion separately. Subsequently, we input these two antisense meta-knowledge into the all-mpnet-base-v2 model for vector-based retrieval.

We employ a directive approach combined with a four-shot learning technique to guide the GPT model to respond to queries in accordance with our specified intentions.

4.3 Experimental results

Table 4 shows the performance of the Enhancing Knowledge through Revisable Chain-of-Thought on the Winogrande development set.

The numbers in Table 4 all omit %, indicating the accuracy rate. We employ GPT-4 to evaluate the meta-knowledge provided by GPT-3.5 for problem-solving, determining its correctness and suitability for the current issue. The last two columns, C0 and C1, represent whether the evaluated meta-knowledge is inapplicable or applicable, with 414 and 853 instances respectively, accounting for 32.68% and 67.32% of the total. The content within the angle brackets "[]" following the model in the first column indicates the method used. A blank space indicates that no chain-of-thought is used. RCoT denotes the use of a revisable COT, that is, the Revisable Chain-of-Thought method proposed in this paper. RTM$_{GPT-4_{MK}}$ and RTM$_{GPT-4_{MK,TK}}$ represent revising Meta-Knowledge(MK) and Transfer Knowledge(TK) in GPT-3.5 with the MK and TK of GPT-4. RKB$_{GPT-4}$ represents a method for offline revision based on a meta-knowledge database from GPT-4. With and without the use of chain-of-thought, GPT-4's accuracy surpasses that of GPT-3.5 by 17.28% to 19.10%, indicating that GPT-4 possesses the fundamental qualifications to serve as a teacher model for GPT-3.5.

From the experimental results in Table 4, we can find the following observations and conclusions:

(1) In the role of a teacher model, GPT-4 can assess the correctness and applicability of the meta-knowledge possessed by GPT-3.5. We approach this evaluation as a binary classification task, where C1 denotes meta-knowledge that is correct and applicable, while C0 indicates otherwise. Examination of the data reveals that, across all rows, the values for C1 consistently exceed those for C0,

Method	Acc	C0	C1
GPT-4	86.03	83.57	87.22
GPT-4 [Regular CoT]	86.58	85.02	87.34
GPT-4 [Revisable CoT]	87.21	85.75	87.92
GPT-3.5	68.75	65.70	70.22
GPT-3.5 [Regular CoT]	68.35	64.00	70.46
GPT-3.5 [Revisable CoT]	68.11	62.80	70.70
GPT-3.5 [RTM$_{GPT-4_{MK}}$]	73.64	71.74	74.56
GPT-3.5 [RTM$_{GPT-4_{MK,TK}}$]	74.59	69.81	76.91
GPT-3.5 [RKB$_{GPT-4}$] Retrieval:all-mpnet-base-v2	68.67	64.49	70.70
GPT-3.5 [RKB$_{GPT-4}$] Retrieval Model:GPT-4	70.80	67.39	72.45
GPT-4 [RKB$_{GPT-4}$] Retrieval Model:GPT-4	86.98	84.78	88.04

Table 4: Results for the *Enhancing Knowledge through Revisable Chain-of-Thought* on the Winogrande development set.

with a range spanning from 2.32% to 7.9%. This discrepancy reflects an inherent imbalance in the meta-knowledge of GPT-3.5 and GPT-4 and suggests a positive correlation between the quality of meta-knowledge and the accuracy of responses.

(2) In the third section of the table, we revised the meta-knowledge and transfer knowledge of GPT-3.5 with that of GPT-4, resulting in a performance improvement of to 5.53% to 6.48% for GPT-3.5. This demonstrates the effectiveness of GPT-4 as a teacher model for GPT-3.5.

(3) In addressing the issue of inappropriate meta-knowledge discernment by GPT-4, GPT-3.5 offline revises the meta-knowledge through the meta-knowledge base of GPT-4, resulting in a marginal improvement of 0.56%. The slight enhancement is due to using the most conservative strategy for offline revision, which is only to revise meta-knowledge when its antonymous meta-knowledge exists within the knowledge base. Owing to the antonymy of meta-knowledge and the deficiencies of the semantic retrieval model, we set the correlation coefficient to 0.8, leading to only 16% of the meta-knowledge being offline revised.

(4) To verify the coverage capability of the knowledge base, we ignore the ability to retrieve the model. We directly used GPT-4 as the retrieval model of GPT-4 knowledge base, and the results showed that the performance of the model improved by 2.69%, which was higher than that of the conservative strategy (0.56%) and lower than that of the online teacher model (5.53%). It shows that the conservative correction strategy needs to be improved, and the knowledge base of the teacher model can play a more significant role. The purpose of our experiment is to illustrate the importance of retrieval models. If the teacher model is available, online revision is better than offline revision.

(5) The last line in Table 4 shows that GPT-4 uses its own past unprocessed knowledge base for offline revision without benefit, indicating that the model cannot revise faulty knowledge in the chain-of-thought without external help.

GPT-3.5's accuracy improved from 68.11% (GPT-3.5 [Revisable CoT]) to 74.59% (GPT-3.5[RTM$_{GPT-4_{MK,TK}}$]). However, it is still significantly smaller than the 87.21% (GPT-4 [Revisable CoT]) used directly with GPT-4. We consider the reason lies in the difference in the knowledge representations of language models. Although the accuracy after knowledge revision does not surpass the accuracy of the teacher model, the goal of our study was not to surpass the performance of the teacher model but to explore the potential of knowledge revision as a viable approach to improve large models with the help of teacher models like human expertise, in scenarios such as education, health, and law, where the expertise of human professionals is paramount. In the experiments, GPT-4 plays the role of teacher model to help GPT-3.5, as getting human expertise in the experiment is costly.

4.4 Case Study

By revising the chain-of-thought, we can obtain the correct answer, as shown in Table 5.

Block 1 of Table 5 presents an example of Meta-Knowledge of GPT-3.5 revised by GPT-4. In the cognition of GPT-3.5, a good doctor should handle simple cases, whereas in reality, a good doctor needs to take on difficult cases. GPT-4 revises it. This case shows that large models may have meta-knowledge contrary to reality and can be revised by other large models.

Block 2 of Table 5 presents an example of a Transfer Knowledge of GPT-4 revised by a human. In the cognition of GPT-4, it understands that if a person is allergic, they will not keep pets. However, the question in the table requires the knowledge that if a person has a pet, then they are not allergic. This necessitates the use of the transfer knowledge that the contrapositive of a statement is logically equivalent to the original statement in order to trans-

Question	**Options**
Sarah was a much better surgeon than Maria so _ always got the easier cases.	(A) Sarah (B) Maria
Meta-knowledge of GPT-3.5	**Evaluation**
If X is a better surgeon than Y, then X always gets the easier cases.	Incorrect, Applicable
Online Revision by GPT-4	**Evaluation**:
If X is a better surgeon than Y, then Y always gets the easier cases.	Correct, Applicable
Question	**Options**
Michael had a cat as a pet but Nelson didn't have any pets because _ had little allergies in their system.	(A) Michael (B) Nelson
Meta-knowledge of GPT-4	**Evaluation**
If X has allergies, especially to pets,then X is less likely to have pets.	Correct, Applicable
Transfer knowledge of GPT-4	**Evaluation**
If P, then Q. → Q, due to not P.	Incorrect, Inapplicable
Online Revision by a human	**Evaluation**
If P, then Q. → not Q because not P.	Correct, Applicable
Question	**Options**
Felicia wanted to be pampered by Emily, so _ went to the jewelry store and bought an expensive ring.	(A) Felicia (B) Emily
Meta-knowledge of GPT-3.5	**Evaluation**
If X wants to be pampered by Y, then X will buy something expensive.	Incorrect,Applicable
Offline Revision with Knowledge Base of GPT-4	**Evaluation**
If X treats Y to something, then X is the one who spends money for it.	Correct, Applicable

Table 5: Three examples of Revision chain-of-thoughts. Text in red indicates errors, while text in blue represents the revistion made.

form the form of the meta-knowledge. However, GPT-4 lacks this capability and has to be corrected by a human.

Block 3 of Table 5 presents an example of offline revision of GPT-3.5 using the knowledge base from GPT-4. The meta-knowledge possessed by GPT-3.5 is not sufficiently abstract and is sometimes contrary to the facts. In contrast, the meta-knowledge abstracted by GPT-4, when addressing similar problems in the past, can be demonstrated by its ability to recognize that 'pamper' can be instantiated as a 'treat.'

5 Conclusion

In this paper, we identify a category of commonsense question answering problems that can be addressed by utilizing the same abstract knowledge and its variations. Through the structured design of chain-of-thought patterns, we propose a revisable chain-of-thought approach that allows for the modification of steps within the chain-of-thought. We introduce two revision methods: 1) specific revisions made by a teacher model for individual problems, and 2) offline revision using a teacher's knowledge base when the teacher model is unavailable or too costly to use. We analyze the difficulty of offline revision, which lies in the potential introduction of correct but irrelevant knowledge. To address this, we propose a method of antonym retrieval that only corrects meta-knowledge conflicting with the meta-knowledge base. Our empirical studies validate the feasibility of correcting thought chains in large language models and highlight the challenges of revision based on offline knowledge bases. This paper suggests that how a model can detect conflicts between its knowledge and external knowledge bases is a question worthy of further investigation.

6 Limitations

In this paper, we only conducted experiments on the Winogrande dataset, given its clear and straightforward problem patterns, which facilitate the demonstration of our proposed revisable chain-of-thought method. Although we did not perform experiments on other datasets, we expect that the underlying principles of our proposed method remain valid.

References

Akari Asai, Sewon Min, Zexuan Zhong, and Danqi Chen. 2023. Retrieval-based language models and applications. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts, ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 41–46. Association for Computational Linguistics.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Michal Podstawski, Hubert Niewiadomski, Piotr Nyczyk, et al. 2023. Graph of thoughts: Solving elaborate problems with large language models. *arXiv preprint arXiv:2308.09687*.

Ruben Branco, António Branco, João António Rodrigues, and João Ricardo Silva. 2021. Shortcutted commonsense: Data spuriousness in deep learning of commonsense reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

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*.

Dheeru Dua, Shivanshu Gupta, Sameer Singh, and Matt Gardner. 2022. Successive prompting for decomposing complex questions. *arXiv preprint arXiv:2212.04092*.

Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. 2023. Specializing smaller language models towards multi-step reasoning. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 10421–10430. PMLR.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Pal: Program-aided language models. In *International Conference on Machine Learning*, pages 10764–10799. PMLR.

Chunxi Guo, Zhiliang Tian, Jintao Tang, Shasha Li, Zhihua Wen, Kaixuan Wang, and Ting Wang. 2023. Retrieval-augmented gpt-3.5-based text-to-sql framework with sample-aware prompting and dynamic revision chain. In *Neural Information Processing - 30th International Conference, ICONIP 2023, Changsha, China, November 20-23, 2023, Proceedings, Part VI*, volume 14452 of *Lecture Notes in Computer Science*, pages 341–356. Springer.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023. Large language models cannot self-correct reasoning yet. *CoRR*, abs/2310.01798.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2391–2401. Association for Computational Linguistics.

Song Jiang, Zahra Shakeri, Aaron Chan, Maziar Sanjabi, Hamed Firooz, Yinglong Xia, Bugra Akyildiz, Yizhou Sun, Jinchao Li, Qifan Wang, et al. 2023. Resprompt: Residual connection prompting advances multi-step reasoning in large language models. *arXiv preprint arXiv:2310.04743*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*.

Xingxuan Li, Ruochen Zhao, Yew Ken Chia, Bosheng Ding, Lidong Bing, Shafiq Joty, and Soujanya Poria. 2023a. Chain of knowledge: A framework for grounding large language models with structured knowledge bases. *arXiv preprint arXiv:2305.13269*.

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2023b. Making language models better reasoners with step-aware verifier. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5315–5333, Toronto, Canada. Association for Computational Linguistics.

Fangzhen Lin, Ziyi Shou, and Chengcai Chen. 2023. Using language models for knowledge acquisition in natural language reasoning problems. *CoRR*, abs/2304.01771.

Zhan Ling, Yunhao Fang, Xuanlin Li, Zhiao Huang, Mingu Lee, Roland Memisevic, and Hao Su. 2023. Deductive verification of chain-of-thought reasoning. *arXiv preprint arXiv:2306.03872*.

Jieyi Long. 2023. Large language model guided tree-of-thought. *arXiv preprint arXiv:2305.08291*.

Ziyang Luo, Can Xu, Pu Zhao, Xiubo Geng, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. 2023. Augmented large language models with parametric knowledge guiding. *CoRR*, abs/2305.04757.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Sean Welleck, Bodhisattwa Prasad Majumder, Shashank Gupta, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. *CoRR*, abs/2303.17651.

OpenAI. 2023. Gpt-4 technical report.

Silviu Pitis, Michael R Zhang, Andrew Wang, and Jimmy Ba. 2023. Boosted prompt ensembles for large language models. *arXiv preprint arXiv:2304.05970*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Sheng Shen, Le Hou, Yanqi Zhou, Nan Du, Shayne Longpre, Jason Wei, Hyung Won Chung, Barret Zoph, William Fedus, Xinyun Chen, Tu Vu, Yuexin Wu, Wuyang Chen, Albert Webson, Yunxuan Li, Vincent Zhao, Hongkun Yu, Kurt Keutzer, Trevor Darrell, and Denny Zhou. 2023. Flan-moe: Scaling instruction-finetuned language models with sparse mixture of experts. *CoRR*, abs/2305.14705.

KaShun Shum, Shizhe Diao, and Tong Zhang. 2023. Automatic prompt augmentation and selection with chain-of-thought from labeled data. *arXiv preprint arXiv:2302.12822*.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 3784–3803. Association for Computational Linguistics.

Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. 2021. Commonsenseqa 2.0: Exposing the limits of AI through gamification. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do large language models know what they don't know? In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 8653–8665. Association for Computational Linguistics.

Fei Yu, Hongbo Zhang, and Benyou Wang. 2023. Nature language reasoning, A survey. *CoRR*, abs/2303.14725.

Miao Zhang, Tingting He, and Ming Dong. 2024. Meta-path reasoning of knowledge graph for commonsense question answering. *Frontiers Comput. Sci.*, 18(1):181303.

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

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V. Le, and Ed H. Chi. 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.

Denny Zhou et al. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

Pei Zhou, Rahul Khanna, Seyeon Lee, Bill Yuchen Lin, Daniel Ho, Jay Pujara, and Xiang Ren. 2021. RICA: Evaluating robust inference capabilities based on commonsense axioms. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7560–7579, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.