Solving Commonsense Question Answering via Large language Models
CPSC 532V 2023W2 Assignment #2

Juntai Cao[♠]
Student #: 50171404
jtcao7@cs.ubc.ca

Yilin Yang[♠]
Student #: 24754350
yangyl17@cs.ubc.ca

Yuwei Yin[♠]
Student #: 36211928
yuweiyin@cs.ubc.ca

Department of Computer Science, University of British Columbia
2366 Main Mall, Vancouver, BC V6T 1Z4, Canada

Abstract

In this paper, we report our solution to Assignment 2, which explores applying large language models (LLMs) on a multiple choice commonsense QA task. We conducted a comprehensive evaluation of three distinct methodologies for enhancing language model performance: traditional in-context learning (ICL), integrating Chain-of-Thought (CoT) prompting to improve logical reasoning, and an retrieval-augmented generation (RAG) approach by leveraging information from knowledge bases (KBs). For ICL, we randomly select 5 entries from the training set as the few-shot examples. In the CoT prompting, we provide each ICL example with a human-written rationale to guide the model. In the KB-augmented method, the rationale comes from the paths in ConceptNet, where the source word and target word in the path are keywords extracted from the original questions and choices. This involves identifying keywords from the context, followed by generating a description based on the best path between the keywords. This description is then presented to the LLM as a rationale to enrich its decision-making process. The experimental results on the COPA dataset show that the basic ICL method of relatively small LLM performs poorly, and so does the naïve KB-augmented method. The CoT method usually works better, while it may fail occasionally. The code and results are available on GitHub. [1]

1 Introduction

In this Section, we introduce the task & dataset in § 1.1 and related work in § 1.2.

1.1 Task and Dataset

In this assignment, we are dealing with a commonsense causal reasoning task named Choice of Plausible Alternatives (COPA) (Roemmele et al., 2011) [2]. COPA has of 500, 100, and 400 questions in the training, validation, and test set respectively, where each data entry consists of the following parts:

- **premise**: the context of the question, e.g., "*My body cast a shadow over the grass.*"
- **question**: the question type. It is either "cause" or "effect".
  "Cause" represents the question "*What could have caused this?*"
  "Effect" represents the question "*What might have happened as a result?*"
- **choice1** and **choice2**: two choices for answering the question, e.g., choice1 is "*The sun was rising.*" and choice2 is "The grass was cut."
- **label**: the correct answer to the question. It is either 0 or 1.
  0 means the correct answer is choice1, while 1 means the correct answer is choice2.

Evaluation Metrics We make the option selection and calculate the accuracy score for evaluation. Given the answer in the example, we can evaluate the performance of our method on the multi-choice QA task using the accuracy metric.:

$$\alpha = \sum_i \mathbb{1}(\alpha_i = a_i) \tag{1.1}$$

where a_i is the correct answer (golden reference) presented in the i-th example $\mathbf{e_i}$, and $\mathbb{1}(\cdot)$ is the indicator function:

$$\mathbb{1}(x) = \begin{cases} 1 & \text{if x is true} \\ 0 & \text{otherwise} \end{cases} \tag{1.2}$$

1.2 Related Work

Large Language Models Large language models (LLMs) (OpenAI, 2022, 2023) have revolutionized the field of natural language processing (NLP) in recent years. Most LLMs are based on the Transformer (Vaswani et al., 2017) architect and are trained on many corpora. The training objective is to predict the next token as accurately or reasonably as possible based on the generated ones.

[♠] Authors contributed equally and listed alphabetically.
[1] https://github.com/YuweiYin/UBC_CPSC_532V
[2] https://people.ict.usc.edu/~gordon/copa.html

In-context Learning We leverage the excellent text-completion and in-context learning (ICL) (Brown et al., 2020; Dong et al., 2023) ability of LLMs to solve commonsense question-answering tasks by providing a few examples. The few-shot ICL approach provides LLMs with the input-output format and thus enhances the generated answer.

Chain-of-Thought Prompting Wei et al. (2022) propose chain-of-thought prompting (CoT) as an intuitive method to guide LLMs to generate the output step by step. This idea is further developed and applied into various reasoning and generation tasks (Kojima et al., 2022; Li et al., 2023; Chai et al., 2024).

External Knowledge Using external knowledge to improve the model performance on commonsense QA tasks is reasonable and promising. Retrieval-augmented generation (RAG) (Lewis et al., 2020; Gao et al., 2023) methods are developed to incorporate external knowledge into LLMs for handling NLP tasks.

2 Step 0: Environment

In this Section, we introduce the setup of the development environment.

Python We use Python3 as the programming language. Refer to the GitHub repository [3] for the detailed development environment setup.

Dataset We load the COPA dataset (Roemmele et al., 2011) from the SuperGLUE benchmark (Wang et al., 2019) via the Hugging Face API [4]. In this assignment, we use the validation set (100 questions) to evaluate the performance of different methods.

```
from datasets import load_dataset

copa = load_dataset("super_glue",
↪  "copa")
val_set = list(copa["validation"])
```

3 Step 1: Large Language Models

In this Section, we introduce the method implementation in § 3.1 and error analysis in § 3.2 at Step 1.

[3] https://github.com/YuweiYin/UBC_CPSC_532V/tree/master/Assignment_2

[4] https://huggingface.co/datasets/super_glue

3.1 Implementation

Large Langauge Model We use GPT-NEO (Black et al., 2022) (`gpt-neo-125m` [5]) as the LLM and load it to the CPU. GPT-NEO is the backbone model for all experiments.

In-context Learning For a test case in the validation set, we construct an in-context learning prompt for the LLM by combining the premise, question, choices, and 5 examples. These examples are randomly sampled from the training set. The random seed is set as 28, so the examples are determined. The prompt template is shown in Table 1.

After setting up the in-context learning examples, we apply a random seed of 0 to all the modules using the `set_seed` method of the Transformers (Wolf et al., 2020) toolkit.

```
from transformers import set_seed

set_seed(0)
```

COPA Evaluation Results The in-context learning method achieves an accuracy of 51% on the validation set.

3.2 Error Analysis

In the first section, we opted for `gpt-neo-125m` to conduct in-context or few-shot learning experiments. The model achieved a prediction accuracy of 51%. We analyzed 20 question-answer (QA) pairs sampled from the examples that the model predicted incorrectly, with 14 categorized as *cause* questions (where the task was to identify the most likely cause of a given premise from a set of options) and 6 as *effect* questions (where the task was to determine the most probable effect of a given premise from a set of options). The findings indicate that, on the whole, the model demonstrates a stronger capability in deducing effects rather than pinpointing causes. We categorize the 20 error examples into three main types: **missing knowledge** (14 examples), **sentiment misinterpretation** (4 examples), and **ambiguous QA pair** (1 example). In our analyses, the correct choice will be highlighted in *teal*, while the incorrect option will be marked in *purple*.

[5] https://huggingface.co/EleutherAI/gpt-neo-125m

Q: The woman felt lonely. What might have happened as a result?
1) She renovated her kitchen.
2) She adopted a cat.
A: 2
Q: The mother needed help looking after her children. What might have happened as a result?
1) She sent the children to daycare.
2) She gave up custody of the children.
A: 1
Q: I learned how to play the board game. What could have caused this?
1) My friend explained the rules to me.
2) My friend got the rules wrong.
A: 1
Q: The woman's eyeglasses fogged up. What could have caused this?
1) She reclined by the pool.
2) She entered the sauna.
A: 2
Q: I ran out of breath. What could have caused this?
1) I climbed several flights of stairs.
2) I read several chapters of the book.
A: 1
Q: {Premise} {Question}
1) {Choice1}
2) {Choice2}

Table 1: Prompt template for LLM generation with in-context learning.

(1) Missing knowledge. A significant portion of the errors stems from the model's failure to connect knowledge between terms (**missing knowledge**). For example, in QA pair 58, to predict the cause of the premise "*The man fainted.*", the model chooses "*He took a nap.*" instead of "*He ran a marathon.*". In QA pair 31, to predict the cause of the premise "*The woman walked with crutches.*", the model chooses "*She shaved her legs.*" instead of "*She broke her leg.*" The choices are very similar, only diverging on a critical detail. These examples highlight the model's inability to grasp causal relationships between terms—such as the link between fainting and running a marathon, or using crutches and having a broken leg—relationships that are generally straightforward for humans to understand.

(2) Sentiment misinterpretation. Some errors can be attributed to the **model's inability to accurately interpret the sentiment of the premise**. For instance, QA pair 37 presents the premise, "*The man felt obligated to attend the event.*" with the task of determining its effect. The options provided are "*He turned down his friend's invitation to go.*" and "*He promised his friend that he would go.*" The model incorrectly chose the first option, demonstrating a misunderstanding of the positive sentiment implied. Similarly, in QA pair 69, the premise "*The couple was happy to see each other.*" led to a choice between a neutral effect "*They rested.*" and a more positive effect "*They kissed.*" The model's selection of the neutral option over the positive one further illustrates its struggle with recognizing and interpreting sentiments accurately.

(3) Ambiguous QA pairs. Certain questions present **a challenge due to their ambiguous nature**. In QA pair 35, the task here is to determine the cause based on the premise, "*The police searched the offender's car.*" The options given for this scenario are "*They were trying to elicit a confession.*" and "*They were looking for illegal drugs.*" Determining the correct answer is difficult, even from a human perspective, showing ambiguity involved in some of the question-answer pairs.

Our analysis suggests that the model may favour options containing more terms that appear closely related to keywords in the premise, instead of reasoning from their actual meaning. In other words, the model might focus more on counting synonyms instead of understanding the meaning. In QA pair 85, to find the effect of the premise "*The woman read the newspaper.*", the model chooses "*She cast a vote in the election.*" instead of "*She discovered the outcome of the election.*" In QA pair 19, to find the cause of the premise "*The woman ran her finger under cold water.*", the model chooses "*She put a diamond ring on her finger.*" instead of "*She burned her finger on the toaster.*" In QA pair 8, to find the cause of the premise "*The flame on the candle went out.*", the model chooses "*I put a match to the wick.*" instead of "*I blew on the wick.*" In all these cases, the chosen options contain terms that are superficially more related to keywords in the premise, though contextually incorrect. This pattern indicates that the model may prioritize familiarity with specific terms over a deeper comprehension of the text's true meaning.

4 Step 2: Chain-of-Thought

In this Section, we introduce the method implementation in § 4.1, result analysis in § 4.2, result

analysis in § 4.3, and error analysis in § 4.4 at Step 2.

4.1 Implementation

Chain-of-Thought Prompting For each of the five in-context learning examples, we manually write their corresponding rationales and insert each rationale between the choice2 ("2)") and the answer "A: ", as shown in Table 2). This CoT approach hints at the LLM to generate a reasonable rationale before answering. Specifically, the five rationales are as follows:

1. Rationale: "The answer is 2 because: the woman adopted a cat to alleviate her feelings of loneliness, seeking companionship and emotional support."
2. Rationale: "The answer is 1 because: The mother sent her children to daycare to receive assistance with childcare responsibilities, enabling her to fulfill other obligations or work commitments."
3. Rationale: "The answer is 1 because: My friend's explanation of the rules facilitated my learning of the board game, providing clarity and guidance on gameplay mechanics."
4. Rationale: "The answer is 2 because: As she entered the sauna, the temperature change caused condensation on the woman's eyeglasses, resulting in fogging up due to the heat and moisture."
5. Rationale: "The answer is 1 because: Climbing several flights of stairs increased my physical exertion, leading to a rapid depletion of oxygen and causing me to run out of breath."

4.2 Result Analysis

In COPA, the correct alternative is randomized so that the expected performance of random guessing is 50%, which means the basic model (51%) performs almost as badly as random guessing. With CoT, the same LLM achieves an accuracy of 58% on the validation set. Compared with the basic ICL method (51%), the CoT method does improve the performance by a large margin.

4.3 Case Study

With Chain-of-Thought (CoT) prompting, the model's performance boosts to an accuracy of 58%, which is better than the basic model (51%). Table 3 shows three examples that were incorrectly predicted by the basic model but correctly predicted by the CoT model.

```
Q: The woman felt lonely. What might have happened as a result?
1) She renovated her kitchen.
2) She adopted a cat.
Rationale: The answer is 2 because: ...
A: 2
Q: The mother needed help looking after her children. What might have happened as a result?
1) She sent the children to daycare.
2) She gave up custody of the children.
Rationale: The answer is 1 because: ...
A: 1
Q: I learned how to play the board game. What could have caused this?
1) My friend explained the rules to me.
2) My friend got the rules wrong.
Rationale: The answer is 1 because: ...
A: 1
Q: The woman's eyeglasses fogged up. What could have caused this?
1) She reclined by the pool.
2) She entered the sauna.
Rationale: The answer is 2 because: ...
A: 2
Q: I ran out of breath. What could have caused this?
1) I climbed several flights of stairs.
2) I read several chapters of the book.
Rationale: The answer is 1 because: ...
A: 1
Q: {Premise} {Question}
1) {Choice1}
2) {Choice2}
```

Table 2: Prompt template for LLM generation with chain-of-thought prompting. The LLM is prompted to generate a rationale before presenting the answer.

4.4 Error Analysis

From the 20 QA pairs sampled from incorrect prediction outcomes, 13 belong to ***cause*** questions, 7 belong to ***effect*** questions. This suggests that CoT prompting does not alter the model's performance dynamics: *cause* questions remain more challenging for the model to address compared to *effect* questions. We classified the sampled incorrect QA pairs into three categories: **flawed rationale** (7 examples), **irrelevant rationale** (11 examples), and **mismatched rationale & predictions** (2 examples).

Val idx	Fields	Basic Model	With CoT
1	Premise	The girl found a bug in her cereal.	
	Question	effect	
	Choice 1	She poured milk in the bowl.	
	Choice 2	She lost her appetite.	
	Rationale	The answer is 2 because: the milk spilled onto the woman's bed.	
	Prediction	w/o CoT: **1**	w/ CoT: **2**
7	Premise	My eyes became red and puffy.	
	Question	cause	
	Choice 1	I was sobbing.	
	Choice 2	I was laughing.	
	Rationale	The answer is 1 because: The woman was crying.	
	Prediction	w/o CoT: **2**	w/ CoT: **1**
88	Premise	The woman contacted the real estate agent.	
	Question	cause	
	Choice 1	The woman planned to buy a condo.	
	Choice 2	The woman needed to clean her house.	
	Rationale	The answer is 1 because: the woman's conversation did not show any signs of change and the agent was not willing to deal with the real estate agent.	
	Prediction	w/o CoT: **2**	w/ CoT: **1**

Table 3: Examples that were incorrectly predicted by the basic model but correctly predicted by the Chain-of-Thought model. Only the CoT model will use the "Rationale". "Val idx" means the index of the example in the validation set.

(1) Flawed rationale. The first category of error arises from **flawed reasoning**: the model produces a fluent rationale with good coverage over the choices and premise, yet fails to align coherently with the premise. In QA pair 43, given the premise "The student was in a rush to get to school on time.", the two choices are "*He left his assignment at home.*" and "*He brought his lunch to school.*" The rationale generated by the model to determine the effect is "The student brought his lunch to school.", which is contradictory to the premise. In QA pair 32, given the premise "I coughed.", the two choices are "*I inhaled smoke.*" and "*I lowered my voice.*" The rationale to predict the cause generated by the model is "My friend's advice to my friend was not to cough, thus raising the tension, which caused my breathing to worsen." The rationale seems to be coherent to the premise and the choice "*I lowered my voice.*", but it lacks coherence with the premise and fails to concentrate on the premise's causative factor.

(2) Irrelevant rationale. Another significant type of error involves the production of **irrelevant rationales**: the model randomly creates rationales that bear no relevance to either the premise or the choices presented. This highlights the model is incapable of processing and comprehending the questions. For instance, in QA pair 25, given the premise "The driver got a flat tire.", the two choices are "*He went over the speed limit.*" and "*He ran over a nail.*" The rationale to determine the cause is "In my home, my car was hit. I lost my brakes.", which is irrelevant to both of the choices. In QA pair 3, given the premise "I wanted to conserve energy.", the two choices are "*I swept the floor in the unoccupied room.*" and "*I shut off the light in the unoccupied room.*" The rationale to determine the effect is "The light in the unoccupied room led to my loss of a breathing space.". The rationale seems to have the term "unoccupied room" that is present in both choices, yet the overall meaning of the sentence is unrelated to the premise and both choices.

(3) Mismatched rationale. The last type of error pertains to **mismatched rationale and predictions**. The model provides a reasonable rationale but fails to predict the label correctly. Take QA pair 73 as an example, with the premise "The seasons changed from summer to autumn." and choices "*People evacuated their homes.*","*Leaves fell from the trees.*". Even though the model generates a rationale "The leaves fell from the trees." for the effect that is consistent with the correct choice, the model still picks the wrong answer.

An example that was incorrectly predicted by the basic model but correctly predicted by the CoT model is QA pair 88. To predict the cause of the premise "The woman contacted the real estate agent.", two choices are "*The woman planned to buy a condo.*" and "*The woman needed to clean her house.*" By generating a rationale of "The real estate agent suggested that she sell the condo for $1 million.", the model can select the correct choice.

It is important to note that despite the improved accuracy CoT brings to the model, the reasoning behind a handful of examples—incorrectly predicted by the basic model but accurately by the CoT model—remains flawed. The same three categories of errors persist within these instances. In QA pair 5, to find the effect of the premise "I doubted the salesman's pitch.", the rationale generated for the two choices "*I turned his offer down.*","*He persuaded me to buy the product.*" is "The salesman offered me a lower price.". This example highlights issues of both flawed reasoning and mismatched rationale and predictions. The reasoning provided contradicts the premise, yet, the model still manages to choose the correct answer despite this opposition. Moreover, in QA pair 59, to predict the

cause of the premise "The man lost the competition.", while the two choices provided are "*The competition was sabotaged.*" and "*He intimidated his competitors.*", the generated rationale "As I stood in a parking lot, the man stopped walking, and his voice was lower than normal." is completely irrelevant to the context.

5 Step 3: External Knowledge

In this Section, we introduce the method implementation in § 5.1, path selection in § 5.2, result analysis in § 5.3, result analysis in § 5.4, and error analysis in § 5.5 at Step 3.

5.1 Implementation

We solve the problem with the help of external knowledge bases (KBs), in particular, ConceptNet (Speer et al., 2017) [6]. The concept path from keywords in the question and those in the options is supposed to explicitly show the commonsense reasoning process and thus assist the decision-making.

5.2 Path Selection

In this assignment, we follow the idea in Assignment 1 [7] to extract keyword pairs from questions and choices and then perform path searching in ConceptNet.

The core idea of our solution is to select the most relevant choice to the question, where the relevance is calculated based on the reasoning paths in ConceptNet between keywords in the question and those in each choice.

As shown in the left part of Figure 1, for each example, we first extract keywords from the question and choices. For each choice in the example, we pair up each keyword of the choice with each question keyword. As the right part of Figure 1 demonstrates, the keyword pairs are denoted as dotted lines between the question keywords ("Q_kw") and choice keywords ("C_kw"). We ignore the red lines because the two words on the ends are the same.

For each keyword pair {source_word, target_word}, we search for valid paths from the source word to the target word via the bidirectional weighted breadth-first search and Dijkstra's algorithm (Dijkstra, 1959) in the knowledge base. After searching, we obtain valid concept paths. As illustrated in Figure 2, the green pair in Figure 1 <"card", "debt"> has a valid path "card" → "credit" → "charge" → "tax" → "money" → "debt", where the relations between each two words are all "RelatedTo".

We visualize the best path of every <C_kw, Q_kw> pair using Graphviz[8]. As shown in Figure 3, the path from "card" to "debt" is labelled with word names, relation types ("RelatedTo"), weights, and directions.

To enhance the efficiency of search in the extensive and dense knowledge graph, we implement a set of heuristic optimizations. Following the processes of path searching and node matching, we illustrate the identified paths through figures and transform the edge information into natural language sentences.

At last, we make the option selection based on the attributes of the identified paths and calculate the accuracy score for evaluation. Specifically, **the path with the largest average edge weights is considered the best for a <question, choice> pair**. The best path is transformed into natural language sentences (using official templates) and serves as the rationale for this choice, as shown in Table 4.

```
{ICL_Examples}

Q: {Premise} {Question}
1) {Choice1}
2) {Choice2}
Rationale: {Rationale}
```

Table 4: Prompt template for LLM generation with rationales generated according to the path searching results in the ConceptNet. The in-context learning examples `{ICL_Examples}` are the same as that in Table 2.

5.3 Result Analysis

Path not found. Since path searching in the ConceptNet is costly, we use the K keyword pair to generate the rationale. Besides, we set the bidirectional search's max depth to limit each path search's run time to approximately one minute. If no path is found, the rationale will be set as "None".

When we set K as 1, the process of rationale generation for 100 validation examples took about 90 minutes. However, there were 39 examples that have a "None" rationale. As a result, the accuracy was 40%.

[6] https://conceptnet.io/

[7] https://github.com/YuweiYin/UBC_CPSC_532V/tree/master/Assignment_1

[8] https://www.graphviz.org/

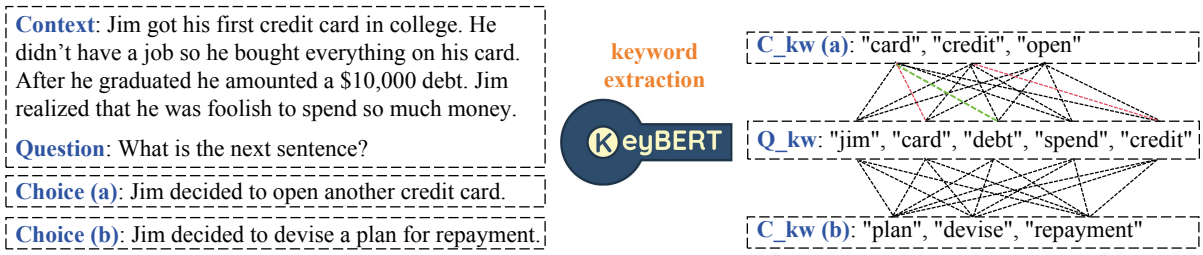

Figure 1: The overview of keyword extraction.

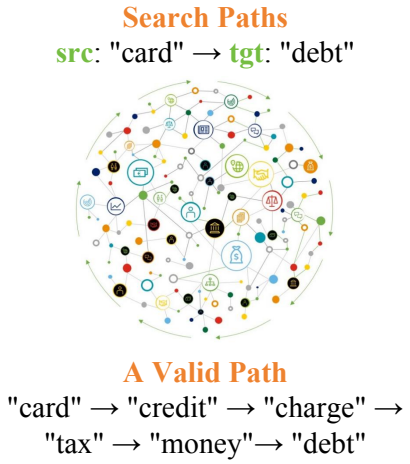

Figure 2: The illustration of path search in a knowledge base.

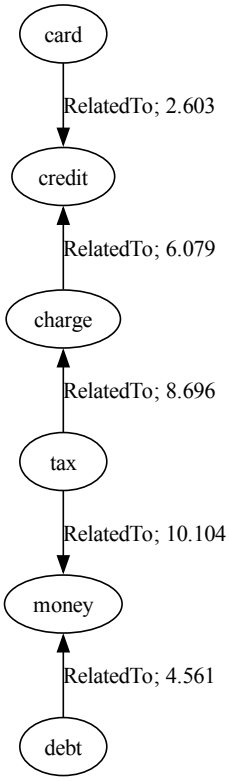

Figure 3: An example of path visualization.

Therefore, we set K as 2. The rationale generation process takes about 180 minutes. Now, there are only 19 examples with a "None" rationale, resulting an accuracy of **47**%, which sounds more reasonable (at least nearly the random guess.)

Method comparison. There are 100 examples in the COPA validation set, where half belong to label 0 (Choice 1) and half to label 1 (Choice 2). It is a balanced binary classification task, meaning a random guess should have an accuracy of 50%. For example, a model that outputs label 1 regardless of the input will have 50 examples correct.

Table 5 shows the experimental results of the three methods on the COPA validation set. The results demonstrate that the **basic model** (51%) and the **KB method** (47%) are nearly random guesses (50%). We believe the main reason the KB method's accuracy is less than 50% is that there are still 19 examples without rationale (a valid path in ConceptNet). Therefore, the accuracy is compromised due to the trade-off between time consumption and performance. The **CoT method** (58%) performs the best, showcasing its advantages. However, when we change the handwritten rationales or random seeds, the results can degrade under 50% easily.

5.4 Case Study

Table 6 presents two instances where the basic or CoT model failed to make accurate predictions, whereas the KB-augmented model identified the correct answers. The success of the KB-augmented model in these cases can be attributed to its ability to discern and utilize relationships between keywords.

In the first example, the model linked "drank" from the premise with "headache" among the choices. Similarly, in the second example, it effectively connected the term "bark" from the premise to "knock" in the choices.

	Basic	CoT	KB
Accuracy	51%	**58%**	47%

Table 5: Experimental results of the three methods. "Basic" refers to the basic LLM model with in-context learning prompts. "CoT" refers to LLM with Chain-of-Thought prompting. "KB" refers to LLM with external knowledge from ConceptNet.

5.5 Error Analysis

For the sampled 20 error examples in the KB method, we categorize them into three main types: **prediction not aligned with rationale** (10 examples), **path not found** (8 examples), and **improper keyword** (2 examples).

(1) Prediction not aligned with rationale. A prevalent challenge observed with the KB-augmented model is the **disconnection between its predictions and the underlying rationales**. Specifically, although the rationale might correctly identify and incorporate the most relevant keywords linking the premise to one of the choices, the model often selects the other choice that does not align with these identified keywords. This discrepancy indicates a gap in the model's ability to correctly apply the logical connections or relevance identified within its rationale to make the most accurate choice.

(2) Path not found. The path not found is described in § 5.3. We alleviate the problem by easing the path searching restrictions, at the cost of running time. Specifically, the path-missing amount was reduced from 39 to 19, and consequently, the running time increased from about 90 minutes to roughly 180 minutes.

(3) Improper keywords. An additional issue encountered with the model involves the **ineffectiveness of selected keywords** in addressing the given problem. For example, consider example 99, which presents a premise "The woman was in a bad mood." accompanied by two choices "*She engaged in small talk with her friend.*" and "*She told her friend to leave her alone.*". The model extracts "friend" and "women" as keywords to predict the effect, yet these terms fall short of capturing the crucial sentimental information. This shows a limitation in the model's keyword extraction process, where the selected keywords do not contain the information required for deducing the correct choice.

6 Step 4: Assignment QA Report

The questions in Step 4 are answered in Table 7.

7 Conclusion

In this study, we set out to evaluate the commonsense reasoning capabilities of LLMs, specifically GPT-NEO, utilizing the COPA dataset across three distinct methodologies within a few-shot learning framework: basic in-context learning, CoT prompting, and KB-augmented approach, the latter incorporating a keyword extraction and sentence generation approach previously established in our Assignment 1. Our findings reveal a notable variance in performance across these methodologies. Specifically, the CoT prompting method emerged as the most effective, significantly outperforming the other approaches. On the other hand, basic in-context learning showed a performance level similar to making random guesses, highlighting its limitations in leveraging the provided context to enhance reasoning capabilities. Surprisingly, the model's accuracy dipped below that of random guessing when applying the KB-augmented approach. This unexpected outcome underscores the complexity of integrating external knowledge bases effectively and suggests areas for further investigation and improvement in model design and knowledge integration strategies. Additionally, our observations show a noticeable disconnect between the rationales generated by the model and its subsequent predictions. Such discrepancies point to a broader challenge of inconsistency in the content generated by smaller LLMs like GPT-NEO, suggesting that while these models can simulate logical processes, aligning these processes with their final decisions remains a challenge. This inconsistency emphasizes the complexity of achieving coherence in model outputs and the need for further refinement in how smaller LLMs integrate and apply reasoning processes to ensure alignment between their explanations and conclusions.

Assignment Fulfillment

Here, we claim that we have fulfilled the assignment requirements according to the grading rubric. Table 8 shows the criteria-to-section mapping. The code is released on GitHub, and all experimental results are reproducible. [9]

[9] https://github.com/YuweiYin/UBC_CPSC_532V/tree/master/Assignment_2

Val idx	Fields	Basic CoT KB
9	Premise	The man drank heavily at the party.
	Question	effect
	Choice 1	He had a headache the next day.
	Choice 2	He had a runny nose the next day.
	CoT Rationale	The answer is 2 because: The man had a runny nose.
	KB Rationale	"day" is conceptually related to "man" because: Morning is like day; Morning is like time; Minute is like time; Minute is like hour; Hour is like clock; You are likely to find clock in house; You are likely to find family in house; Daughter is like family; Daughter is like girl; Girl is like young; Child is like young; Boy is like child; Boy is like man. "day" is conceptually related to "party" because: Birthday is like day; Birthday is like party.
	Prediction	Basic: **2** CoT: **2** KB: **1**
67	Premise	The dog barked.
	Question	cause
	Choice 1	The cat lounged on the couch.
	Choice 2	A knock sounded at the door.
	CoT Rationale	The answer is 1 because: The cat was knocking on the door, trying to enter the room.
	KB Rationale	"door" is conceptually related to "barked" because: Door is like opening; Window is like opening; Window is like wall; Paint is like wall; Paint is like color; Color is like shade; Shade is like tree; Bark is part of tree; Barked is a form of the word bark. "door" is conceptually related to "dog" because: Door is like house; Home is like house; Nest is like home; Nest is like baby; Baby is like small; Squirrel is like small; Squirrel is like tree; Bark is part of tree; Dog can bark.
	Prediction	Basic: **1** CoT: **1** KB: **2**

Table 6: Examples that were incorrectly predicted by the basic/CoT model but correctly predicted by the ConceptNet model. "Val idx" means the index of the example in the validation set.

References

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. Gpt-neox-20b: An open-source autoregressive language model. *arXiv preprint arXiv:2204.06745*, abs/2204.06745.

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. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, volume 33, pages 1877–1901, Virtual Event. NeurIPS.

Linzheng Chai, Jian Yang, Tao Sun, Hongcheng Guo, Jiaheng Liu, Bing Wang, Xinnian Liang, Jiaqi Bai, Tongliang Li, Qiyao Peng, and Zhoujun Li. 2024. xcot: Cross-lingual instruction tuning for cross-lingual chain-of-thought reasoning. *arXiv preprint arXiv:2401.07037*, abs/2401.07037.

Edsger W. Dijkstra. 1959. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1:269–271.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*, abs/2301.00234.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, abs/2312.10997.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Jia Li, Ge Li, Yongmin Li, and Zhi Jin. 2023. Structured chain-of-thought prompting for code generation. *arXiv preprint arXiv:2305.06599*.

OpenAI. 2022. Chatgpt. *OpenAI Research*.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Question	Answer
1. Basic model	
1.1. Include the error analysis.	See § 3.2
2. CoT	
2.1. Did you manage to improve the performance upon the basic model?	See § 4.2
2.2. If applicable, showing a few examples that were incorrectly predicted by the basic model but correctly predicted by the CoT model.	See § 4.3
2.3. Include the error analysis.	See § 4.4
3. External knowledge	
3.1. How did you choose which path to include?	See § 5.2
3.2. Did you manage to improve the performance upon the basic model?	See § 5.3
3.3. How well does this model do compared to CoT?	See § 5.3
3.4. If applicable, showing a few examples that were incorrectly predicted by the basic/CoT model but correctly predicted by the ConceptNet model.	See § 5.4
3.5. Include the error analysis.	See § 5.5

Table 7: Answers to the questions in Step 4.

Criteria	Fulfillment Statement
Neural model error analysis	See § 3.2, § 4.4, and § 5.5
Neuro-symbolic model error analysis	See § 3.2, § 4.4, and § 5.5
Discussion of performance difference	See § 4.2 and § 5.3
Implementation: Neural Model	See § 2, § 3.1, § 4.1, and § 5.1
Implementation: Neuro-symbolic model	See § 2, § 3.1, § 4.1, and § 5.1

Table 8: Statement of assignment fulfillment according to the marking rubric.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *Logical Formalizations of Commonsense Reasoning, Papers from the 2011 AAAI Spring Symposium, Technical Report SS-11-06, Stanford, California, USA, March 21-23, 2011*. AAAI.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3261–3275.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.