Optimization via Chain-of-thought and AMR Prompting Techniques on the MMLU Dataset

Nora Garrity
Dept. of Linguistics
University of Washington
ngarri3@uw.edu

Jeongyeob Hong
Dept. of Linguistics
University of Washington
yeob@uw.edu

Abstract

Chain-of-thought prompting boosts model performance on reasoning tasks by introducing reasoning steps that resemble the human thought process. However, using Abstract Meaning Representation as a part of reasoning step has not improved performance in tasks like Logical Fallacy Detection. Driven by these mixed outcomes, our paper explores how different intermediate steps affect Large Language Models (LLMs). We focus on 'corrupted' steps—intentionally misleading instructions by changing a verb into its antonym —and 'shorter' AMR steps. Using the MMLU benchmark and GPT-3.5, we aim to provide deeper insights on affects of intermediate steps.

1 Introduction

In the evolving landscape of artificial intelligence, the development and refinement of Large Language Models (LLMs) is an important mark of progress. A huge question to tackle is what machines can do and how they can interact within the realms of human language and cognition. The relation between datasets and LLM's capability highlights a challenge in artificial intelligence research: creating models that not only replicate human-like responses and reasoning processes but also exhibit a deep and nuanced interpretation of socially complex subject matter such as physical sciences to ethics and law.

Recently, there has been an increasing demand for the usage of LLMs without allocating additional training data or fine-tuning the models (Ignat et al., 2023). Instead the LLM-based NLP systems directly learn the context from the raw input texts, and achieve high performance (Brown et al., 2020). For instance, simply providing a prompt made up of a list of input and output pairs that showcases the task is enough for the LLMs to perform effectively. Such methods, few-shot or in-context learning (ICL), have further developed with the advent of Chain-of-Thought prompting (Wei et al., 2023).

Introducing intermediate steps that decompose the reasoning process helps the model improve its performance significantly. However, little is known about how and why intermediate steps work in LLMs and ICL. It is also unclear which intermediate steps are needed for LLM to interpret the human queries correctly.

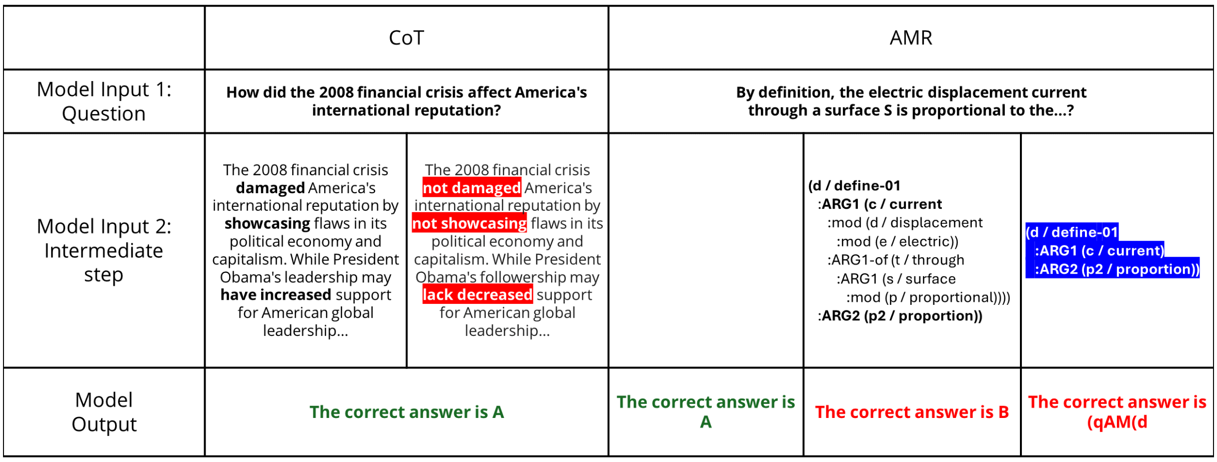
Building upon prior research (Min et al., 2022, (Webson and Pavlick, 2021)) to unravel the complexities of prompting techniques, this paper aims to investigate the impact of corrupting components of a prompt on a LLM's interpretation of semantic changes. We will particularly concentrate on modifying verbs, crucial elements that signal the predicate structure within sentences, to examine their influence on model comprehension.

Furthermore, this study introduces various intermediate steps, including the use of Abstract Meaning Representation (AMR), a semantic representation language (Banarescu et al., 2013). AMR offers a structured representation of the semantic content of text, translating natural language into a graph-based form that captures relationships and entities in a manner that is both comprehensive and computationally tractable. This approach to presenting structured data as auxiliary input has been demonstrated to enhance performance in tasks such as machine translation and summarization(Chen et al., 2022).

As these LLMs become increasingly integrated into societal functions, enhancing their ability to accurately and sophisticatedly handle complex questions is important. Through this exploration, we aim to help build more reliable and beneficial AI applications in critical and socially relevant domains.

2 Related Work

In recent years, multiple works have covered the mechanisms behind the ICL. Among these, Min



	CoT		AMR		
Model Input 1: Question	**How did the 2008 financial crisis affect America's international reputation?**		**By definition, the electric displacement current through a surface S is proportional to the...?**		
Model Input 2: Intermediate step	The 2008 financial crisis **damaged** America's international reputation by **showcasing** flaws in its political economy and capitalism. While President Obama's leadership may **have increased** support for American global leadership...	The 2008 financial crisis **not damaged** America's international reputation by **not showcasing** flaws in its political economy and capitalism. While President Obama's followership may **lack decreased** support for American global leadership...		**(d / define-01** **:ARG1 (c / current** :mod (d / displacement :mod (e / electric)) :ARG1-of (t / through :ARG1 (s / surface :mod (p / proportional)))) **:ARG2 (p2 / proportion))**	**(d / define-01** **:ARG1 (c / current)** **:ARG2 (p2 / proportion))**
Model Output	**The correct answer is A**		**The correct answer is A**	**The correct answer is B**	**The correct answer is (qAM(d**

Figure 1: Illustration of our experiment where misguided prompts, marked red in the third row, did not affect the results of the model's performance. On the other hand, introducing AMRs as an intermediate step resulted in a performance drop.

et al., 2022 conducted experiments across a wide range of Question Answering (QA) tasks, altering the input-output pairs in in-context demonstrations. They discovered that assigning random labels to inputs did not harm the model performance, suggesting that an accurate task demonstration in prompts might not be critical. Instead, they emphasized the significance of maintaining the correct format and token distribution.

Similarly, Webson and Pavlick, 2021 investigated the impact of manipulating task instructions, presented in natural language, on the performance of fine-tuned LLMs. Their findings indicated that these models performed consistently, even when presented with misleading or incorrect prompt instructions on Natural Language Inference (NLI) tasks. Our study differs from theirs as we test on LLM without any fine tuning. The instructions Webson and Pavik used are provided during fine-tuning, while our corrupted instructions are used in ICL settings.

Wang et al., 2023 introduced invalid reasoning steps into Chain of Thought prompts for arithmetic and factual QA tasks, employing GPT-3.5-based models. They observed approximately 10% performance decline. Our research, however, distinguishes them by examining the effects of verb corruption in prompts, in contrast to Wang et al., 2023's focus on key noun phrases and symbols. We also tested on different tasks with more complicated reasoning.

Jin et al., 2024 attempted to introduce AMR as an intermediate step for tasks such as Paraphrase Detection, Translation, Logical Fallacy Detection, Event Extraction, and Code Generation. While AMR offers a structured representation of the semantic content of texts, the result showed that everything except Code Generation showed a performance drop. Our study follows up their work with different methodological approach.

3 Dataset

3.1 Overview of MMLU

The MMLU dataset serves as a benchmark dataset that assesses the multitask accuracy of large language models. In total it covers 57 subjects, with 15908 multiple choice questions in total. The purpose of the dataset is to provide a benchmark to test an LLM's breadth of understanding, as well as its depth of understanding.

Top 5 Tasks	Bottom 5 Tasks
US Foreign Policy	College Chemistry
High School Psychology	Moral Scenarios
Miscellaneous	College Physics
Marketing	High School Physics
High School Gov't and Politics	High School Mathematics

Table 1: Top and Bottom 5 Tasks of MMLU Benchmark based on GPT-3 Few-Shot Accuracy

Table 1 illustrates the tasks that GPT3 performed successfully and poorly. The bottom 5 tasks showed roughly near random performance (25%). Note that most of the subjects involve the usage of symbols (physics and chemistry units and mathematical operators).

3.2 Analysis of US Foreign Policy and College Physics Dataset

In our research, we will concentrate on one dataset from each category listed in Table 1: specifically, the US Foreign Policy and College Physics datasets. Within the top 5 tasks, two (US Foreign Policy and High School Government and Politics) pertain to policy-related questions. Similarly, two of the five least successful tasks are related to physics. Consequently, we have selected the US Foreign Policy dataset to represent the more successful tasks and the College Physics dataset to exemplify the less successful ones. It is noteworthy that GPT-3 achieved approximately 70% accuracy on the US Foreign Policy tasks.

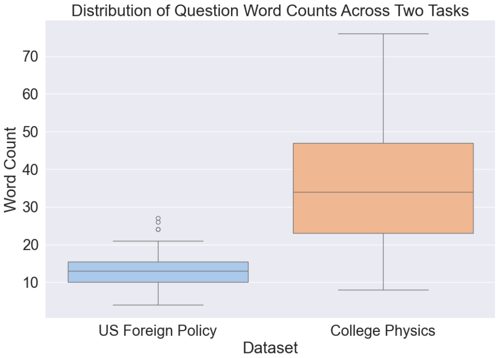


Figure 2: Number of Words in Questions per Tasks

Figure 2 showcases the number of words used in questions of each task. Questions in College Physics data contain almost twice as many words as those from US Foreign Policies. Considering the importance of the context window and the number of input tokens in LLM, altering the number of questions could lead to different results. We further analyzed the number of characters used in each question, which also showed that Physics Dataset contains more characters than the US Foreign Policy. More information can be found in the Appendix A.

Dataset	TTR
College Physics	0.783314811116
US Foreign Policy	0.958002738205

Table 2: TTR of College Physics and Foreign Policy Data

Type-token ratio (TTR) analysis of the dataset was performed by tokenizing the text to identify unique words and the total number of words. The TTR was calculated by dividing the number of unique words by the total word count for each text entry for each question, providing a measure of vocabulary diversity. A higher TTR indicates a greater diversity of vocabulary, suggesting the text may be more complex or sophisticated. Conversely, a lower TTR suggests less vocabulary diversity, which could indicate simpler text or repetitive use of words. Analyzing TTR for these datasets is crucial when evaluating datasets for LLMs. This data can highlight if either dataset is biased toward simpler or more complex language, impacting the model's ability to generalize across different linguistic contexts.

As seen in Table 2, the TTR for the College Physics dataset is 0.7833, while the TTR for the Foreign Policy dataset is 0.9580. The lower TTR in the College Physics dataset could indicate the College Physics questions feature a more limited, specialized vocabulary, which might be highly repetitive (with many technical terms used frequently). This precision can challenge an LLM if it hasn't been sufficiently trained on similar scientific content, leading to poorer performance. In contrast, the Foreign Policy dataset might have a broader vocabulary reflecting diverse topics and discussions, which an LLM, especially one trained on varied general texts, could handle more effectively due to its broader exposure and generalization capabilities.

CoT prompting may help within the physics dataset by guiding it through intermediate steps, making it easier to tackle complex problems even with a lower TTR, which might indicate specialized but repetitive language. For the Foreign Policy dataset, with a higher TTR indicating diverse vocabulary and concepts, CoT prompting can aid in navigating through nuanced arguments or complex geopolitical scenarios. AMR prompting could further enhance understanding of the College Physics data by providing a structured representation of the text, aiding the model in grasping the underlying relationships and understanding dense, specific, and information-rich texts like those in physics.

4 Method

4.1 Research Question

Based on our initial motivation and a preliminary analysis of two datasets, we intend to delve deeper into the effects of intermediate steps on the performance of models tasked with the MMLU's datasets.

We aim to answer following two questions:

- **RQ1: Do LLMs maintain consistent performance when confronted with corrupted intermediate steps in the US Foreign Policy Task?**

- **RQ2: Can shorter intermediate steps enhance LLM performance in the College Physics Task?**

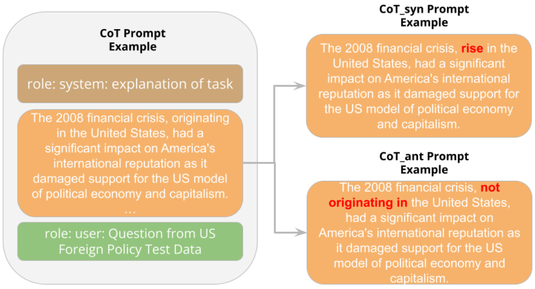
We hypothesize that the model will remain consistent ($\pm < 5\%$) performance in accuracy. Additionally, with the shorter AMR representation, we believe that the LLM will show practical improvements ($\pm > 5\%$) in accuracy. In other words:

- **H1:**
$$p(\text{corrupt\_CoT}|x) \cdot p(y|\text{corrupt\_CoT}) \approx p(\text{CoT}|x) \cdot p(y|\text{CoT})$$

- **H2:**
$$p(\text{short}|x) \cdot p(y|\text{short}) > p(\text{amr}|x) \cdot p(y|\text{amr})$$

4.2 Experiment Design

4.2.1 Prompt Design


CoT Prompt
Example
role: system: explanation of task
...
role: user: Question from US Foreign Policy Test Data
CoT_syn Prompt
Example
CoT_ant Prompt
Example

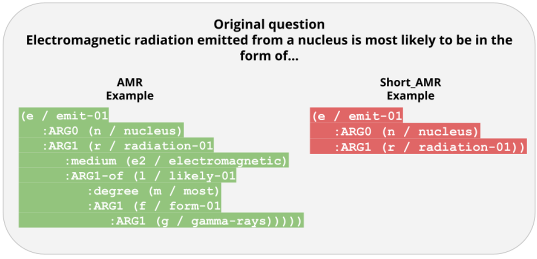
The 2008 financial crisis, originating in the United States, had a significant impact on America's international reputation as it damaged support for the US model of political economy and capitalism.

The 2008 financial crisis, rise in the United States, had a significant impact on America's international reputation as it damaged support for the US model of political economy and capitalism.

The 2008 financial crisis, not originating in the United States, had a significant impact on America's international reputation as it damaged support for the US model of political economy and capitalism.

Figure 3: Example of corrupted cot prompts. Corrupted verbs are in red.

In our experiments, we manipulate these intermediate steps into two distinct categories: *corrupted* and *shorter* ones. The *corrupted* intermediate step changes one random verb from reasoning steps into either antonym or synonym. The corrupting process is completed algorithmically through Wordnet using NLTK. If given verb has no antonym or synonym, it is attached with 'not' or 'roughly'. Figure 3 shows the example of *corrupted* prompts. Note that we also created a prompt where all verbs are changed into either antonym or synonym.


Original question
Electromagnetic radiation emitted from a nucleus is most likely to be in the form of...
AMR
Example
Short_AMR
Example

```
(e / emit-01
    :ARG0 (n / nucleus)
    :ARG1 (r / radiation-01
        :medium (e2 / electromagnetic)
        :ARG1-of (l / likely-01
            :degree (m / most)
            :ARG1 (f / form-01
                :ARG1 (g / gamma-rays)))))
```

```
(e / emit-01
    :ARG0 (n / nucleus)
    :ARG1 (r / radiation-01))
```

Figure 4: Example of different level depth of AMR representation. The right example showcase a shortened AMR

Short intermediate steps, on the other hand, are based on AMR. Given that AMR employs a graph-based representation, the depth of a graph acts as an indicator of its complexity. Since a higher node represents the core semantic relationship, turning AMR into a depth of 1 would extract the most salient semantic meaning of a given sentence. Figure 4 shows the example of *Short* prompts. Full example prompts are provided in the Appendix A.

4.2.2 Model

We tested our hypothesis on the GPT3.5-turbo model using OpenAI's API. The context window for this model, the amount of tokens it can accept, is 4,096 tokens. Compared to the most advanced model GPT4, which has a context window of 128,000 tokens, it has a fairly limited input size. However, it is the fastest model. Also we chose this model since it is closer to GPT3 tested in (Hendrycks et al., 2021) and (Wei et al., 2023). For the AMR experiments, the cost of running them on this model was $.302604. For the CoT experiments, the total cost was $0.5173, making the overall cost of this project $0.819904.

4.3 Procedure

Figure 5 illustrates our plan for this study. Our study examines two experimental conditions: corrupted_CoT and short_AMR, using two datasets. We create CoT examples and AMR representations, drawing five example steps from the development set. For CoT experiments, we query the model to answer given question 'step-by-step', a zero-shot prompting technique. Meanwhile, we manually created AMR representation. This We assess corrupted prompts against both baseline prompts (direct inputs without intermediate steps) and standard CoT and AMR formats. Our analysis focuses on

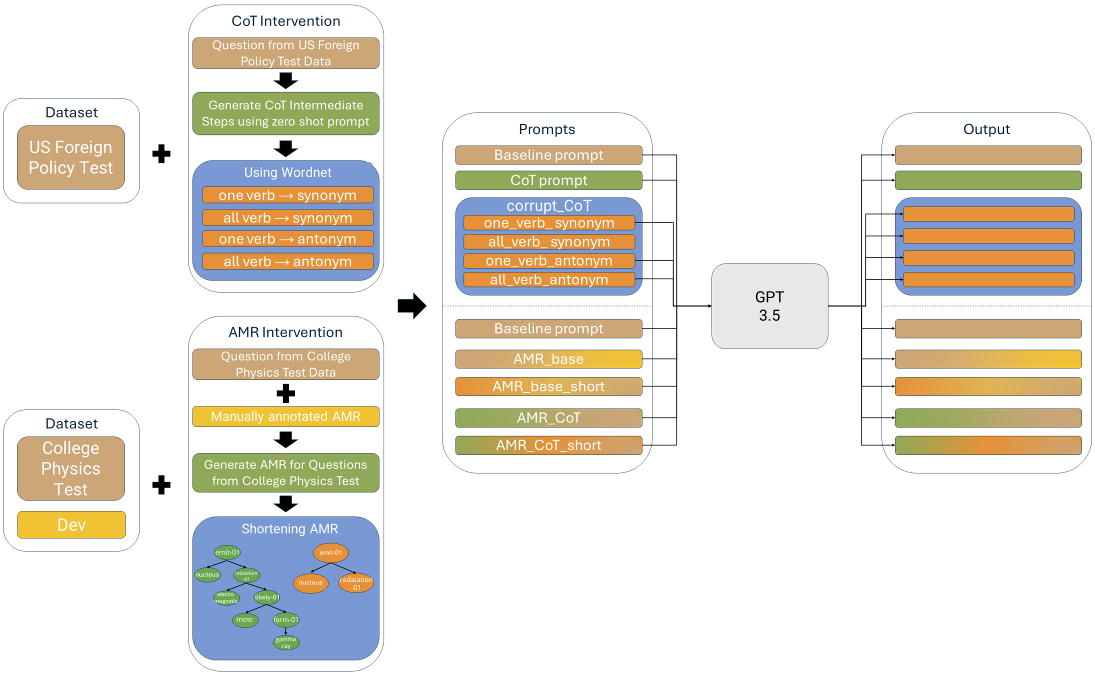


Figure 5: Detailed research plan

comparing the accuracy of each prompt type to understand their impact on model performance.

5 Results

We conducted an initial experiment with US Foreign Policy dataset to find out the effect of changing verbs in prompts. We ran four different tests with the first 25 instances of US Foreign Policy Test dataset: baseline, CoT, verb_corrupted_CoT, and negative_verb_corrupted_CoT. Verb_corrupted_CoT refers to a prompt that has one verb changed to its antonym and negative_verb_corrupted_CoT refers to a prompt with ’not’ inserted before all verbs in the CoT explanations.

Table 3: Accuracy of CoT Experiments

Method	Correct Cases	Percentage (%)
Baseline	85/100	85.0
CoT	79/100	79.0
one_verb_antonym	84/100	84.0
all_verb_antonym	81/100	81.0
one_verb_synonym	83/100	83.0
all_verb_synonym	84/100	84.0

As can be seen from Table 3, the baseline performance of GPT-3.5 Turbo on the Foreign Policy questions, with no CoT prompting, is 85%. This is an improvement from GPT-3’s past 70% accuracy on this dataset. This means that the underlying model, GPT-3.5, is already relatively sound, and this is what we will compare the models performance on the CoT prompting to. With just simple CoT prompting, the model’s performance dropped to 79% accuracy. This was not expected, as CoT prompting is generally accepted as a viable means to improve model performance. There are several possibilities for why this occurred. It is possible that the intermediate steps introduced more ambiguity and complexity to the question, diverting the models attention away from the question to be answered and therefore reducing the models ability to sufficiently reason. Another possibility is that the model’s sensitivity to context is high. If the intermediate CoT steps are not well optimized or not closely aligned enough to the answerable question, the ambiguity introduced could cause performance to suffer. Finally, the decrease in performance could be caused by error propagation. We used the model to generate the CoT prompts, and if the model struggles and produces any errors early on in the generation of the CoTs, this error would magnify in the following steps leading to an incorrect final answer.

For our subsequent antonym experiments, where we replaced one verb in the prompt with its antonym (one_verb_antonym) and replaced all verbs in the prompt with their antonyms (all_verb_antonym), the model scored 84% and 81% respectively. For our synonym experiments, one_verb_synonym and all_verb_synonym scored 83% and 84% respectively. From all of these results, we can see that altering the CoT prompting verbiage does influence the model's reasoning process, albeit not in a significant way (> 5%) when compared to the undisrupted CoT performance. In this comparison, we can see that the performance improved with all corrupted CoTs, ranging from +2% to +5%. However, if we compare the results of the verb corruption experiments to the baseline performance of the model (no CoT), we can see that the model's accuracy dropped, ranging from -4% to -1%.

The fact that replacing verbs with their antonyms or synonyms only slightly affected the model's accuracy suggests that its understanding is robust to some lexical variations, demonstrating a degree of semantic flexibility. However, the reduction in performance when compared to the baseline indicates a reliance on precise verb cues for optimal reasoning. The relatively stable performance in the synonym experiments (83% and 84%) compared to the antonym experiments (84% and 81%) may indicate that the model finds it easier to adapt to synonyms, which preserve the original intent, than to antonyms, which invert it. Furthermore, the fact that the all_verb_synonym experiment scored slightly higher than the one_verb_synonym experiment might suggest that the model's reasoning process can adjust to consistent changes across the prompt. In contrast, the drop in the all_verb_antonym experiment compared to the one_verb_antonym experiment could indicate that the model struggles more with consistent negation or opposition in the action sequences.

Table 4: Accuracy of AMR Experiments

Method	Correct Cases	Percentage (%)
Baseline	42/102	0.411765
AMR_base	20/102	0.196078
AMR_base_short	1/102	0.009804
AMR_CoT	51/102	0.500000
AMR_CoT_short	49/102	0.480392

In the AMR based experiments, shown in Table 4, shortening the AMR alone decreased the performance of the model, likely because the shortened AMR omitted crucial information that helped the model understand the context and content of the prompts better. However, when the shortening of the AMR was coupled with the use of CoT prompting, performance increased. This improvement can be attributed to the CoT approach; even if the AMR is shortened, the CoT can compensate for the loss of detail by guiding the model through the reasoning process, enabling it to arrive at the correct answer despite having less initial information. The CoT likely allows the model to leverage the information available in the shortened AMR more efficiently.

6 Analysis

The results of the experiments have shed some light onto the potential pitfalls of their setup. An evident flaw is the reduced performance of the model with the use of uncorrupted CoT. It is unusual that the CoT caused such a reduction in the models performance, and a potential reason for this is that the CoT prompts were generated by the model itself. Having GPT-3.5 generate its own CoT prompts could introduce several issues. First, there's a risk of the model generating incorrect or irrelevant intermediate steps, as the model's understanding is based solely on the patterns it has seen during training, without true comprehension. If the generated CoT is flawed, the reasoning process can lead to inaccurate conclusions while propagating errors through each step. Additionally, the model might produce a CoT that is overly complex or convoluted, which could introduce unnecessary confusion or ambiguity in the reasoning process. There's also the challenge of consistency; if the model generates CoT prompts that vary significantly in style or approach, it may be difficult to ensure consistent reasoning or to compare results across different prompts. Finally, self-generated CoT prompts might reflect and amplify any biases inherent in the training data, potentially leading to biased or unfair reasoning processes and outcomes. For future research, it will be important to carefully design and vet the CoT prompts to ensure they aid the model's problem-solving capabilities.

Comparing the results from the Baseline model to those of AMR_base reveals a significant decline in model accuracy, with a 50% reduction observed. This decrease may largely be attributed to the structural limitations in the AMR for accurately representing units and arithmetic operations. For instance, physics questions often demand a precise understanding of units and quantifiers, where an error as simple as misinterpreting ”mm” (millimeters) for ”m” (meters) can critically undermine the accuracy of responses. This type of error will not be relevant to US Foreign Policy, where such precise unit conversions are less important. In terms of arithmetic relationships, our pregenerated AMRs struggled to use special frames such as ‘rate-entity-91’ which is necessary to represent a concept ‘twice’. Moreover, as mentioned in the data analysis, a larger number of tokens resulted in additional complexity as they often appeared in multi-sentence. Considering the difficulty of generating multi-sentenced AMR, there are chances that our provided AMR representation might not fully represent the question of semantic relations.

7 Conclusion

Ideally, in a future experiment with more time and resources allotted, the CoT prompts would be assigned to each question manually by a human. However, these experiments demonstrated the potential utility of using an LLM to generate the CoT prompts, and highlights the need for continued research as to how to generate CoT prompts via LLM as effectively as possible. This would involve creating CoT prompts that will not propagate errors down the line and that relate to the subject matter of the answerable question.

Although providing a structured input to an LLM did not result in a better performance in our research, testing on different models with different tasks could lead to an interesting finding. Since AMR as an auxiliary input has shown promising results in text generation tasks, further research on such tasks using manually annotated AMR would further verify the effectiveness of structured input to LLMs.

Additionally, testing with a larger dataset could produce more reliable results and perhaps better expose any trends in our data, such as the differences in performance between the one_verb_antonym and the all_verb_antonym experiments.

Overall, our research contributes to the research community by corroborating previous findings such that the semantic meaning of in-context examples is not important for the model’s performance. Further testing on different factors of ICL is needed to understand deeper on the mechanism of ICL.

References

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *LAW@ACL*.

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, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff 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. *ArXiv*, abs/2005.14165.

Liang Chen, Peiyi Wang, Runxin Xu, Tianyu Liu, Zhifang Sui, and Baobao Chang. 2022. ATP: AMRize then parse! enhancing AMR parsing with PseudoAMRs. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2482–2496, Seattle, United States. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Oana Ignat, Zhijing Jin, Artem Abzaliev, Laura Biester, Santiago Castro, Naihao Deng, Xinyi Gao, Aylin Gunal, Jacky He, Ashkan Kazemi, Muhammad Khalifa, Nam Ho Koh, Andrew Lee, Siyang Liu, Do June Min, Shinka Mori, Joan Nwatu, Verónica Pérez-Rosas, Siqi Shen, Zekun Wang, Winston Wu, and Rada Mihalcea. 2023. A phd student’s perspective on research in nlp in the era of very large language models. *ArXiv*, abs/2305.12544.

Zhijing Jin, Yuen Chen, Fernando Gonzalez, Jiayi Zhang, Julian Michael Jiarui Liu, Bernhard Schölkopf, and Mona Diab. 2024. Role of semantic representations in an era of large language models. *arXiv preprint*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work?

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2023. Towards understanding chain-of-thought prompting: An empirical study of what matters.

Albert Webson and Ellie Pavlick. 2021. Do prompt-based models really understand the meaning of their prompts? *ArXiv*, abs/2109.01247.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS 2022*.

A Appendix

More visualization and example prompts:

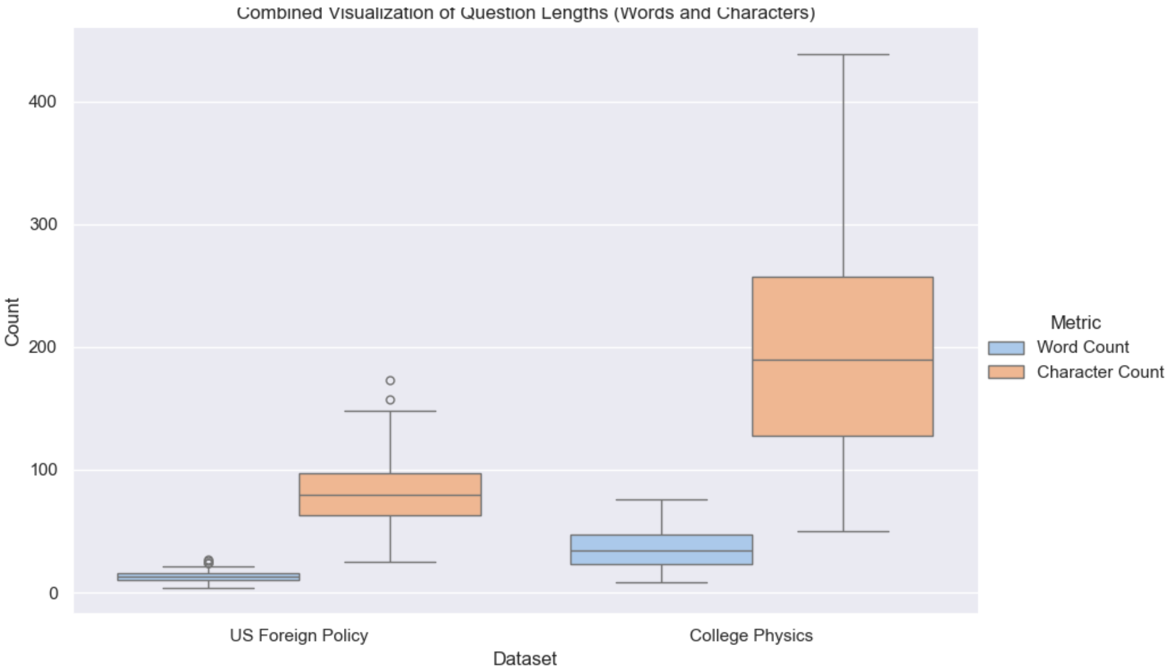


Figure 6: Word and Character Count of Questions In Two Tasks

Table 5: Base Prompts Example

Role: system **Content:** Your task is to answer the following US-Foreign Policy question by picking the correct answer from the given choices of A, B, C, and D.
Role: user **Content:** What was the significance of the Gulf of Tonkin resolution? A: It allowed the US to intensify its involvement in Vietnam B: It illustrated the influence of public opinion on US foreign policy C: It enhanced Congressional control over the Vietnam War D: It curtailed US involvement in Vietnam

Table 6: Zero Shot Prompts Example

Role: system **Content:** You are given a US-Foreign Policy question with 4 possible answers, marked A, B, C, and D. Read the question, reason step-by-step in 100 words or fewer. DO NOT DIRECTLY STATE ANSWER IN YOUR RESPONSE
Role: user **Content:** How did NSC-68 change U.S. strategy? A: It globalized containment B: It militarized containment. C: It called for the development of the hydrogen bomb. D: It All of the above

Table 7: Example Prompt for 'CoT prompts'

System: Your task is to answer the following US-Foreign Policy question by picking the correct answer from the given choices of A, B, C, and D
User: How did the 2008 financial crisis affect America's international reputation? A: It damaged support for the US model of political economy and capitalism. B: It created anger at the United States for exaggerating the crisis. C: It increased support for American global leadership under President Obama. D: It reduced global use of the US dollar. The 2008 financial crisis damaged America's international reputation by showcasing flaws in its political economy and capitalism (not B). This crisis did not create anger at the US for exaggerating the crisis (not B). While President Obama's leadership may have increased support for American global leadership, the crisis itself did not directly lead to this outcome (not C). The crisis did not necessarily reduce global use of the US dollar, as the dollar remained a key currency in international transactions (not D). **Assistant:** A **User:** How did NSC-68 change U.S. strategy? A: It globalized containment. B: It militarized containment. C: It called for the development of the hydrogen bomb. D: All of the above. NS-68 was a document that recommended a significant increase in military spending to combat the Soviet threat during the Cold War. It not only globalized containment by advocating for a more aggressive approach worldwide (such as in Korea), but also militarized containment by emphasizing the need for a strong military presence. Additionally, it called for the development of the hydrogen bomb to deter Soviet aggression. S. strategy brought about by NS-68. **Assistant:** D **User:** The realm of policy decisions concerned primarily with relations between the United States and the rest of the world is known as A: terrorism policy. B: economic policy. C: foreign policy. D: international policy. Foreign policy refers to the realm of policy decisions concerned primarily with relations between the United States and the rest of the world. Terrorism policy focuses specifically on combating terrorism, economic policy deals with financial matters, and international policy is a broader term that encompasses various global issues beyond just relations with other countries. **Assistant:** C **User:** How do Defensive Realism and Offensive Realism differ in their explanation of state behaviour? A: Defensive realists place greater emphasis on the role of international institutions. B: Defensive realists place less emphasis on geographical factors. C: Offensive realists give more priority to the national interest than Defensive realists. D: Defensive realists believe states are security maximizers, while Offensive realists believe states to be power maximizers. Defensive Realism focuses on states seeking security through defensive measures, emphasizing international institutions for stability. Offensive Realism, on the other hand, prioritizes power and sees states as seeking to maximize their power rather than security. Defensive Realists believe states are security maximizers, while Offensive Realists view them as power maximizers. Defensive Realists do not place less emphasis on geographical factors, and Offensive Realists do not necessarily give more priority to the national interest. **Assistant:** D **User:** How did Donald Trump attack globalization in the 2016 campaign? A: Globalization had made men like him too rich B: Globalization only benefited certain American states, such as New York C: Liberal elites had encouraged globalization, while 'ordinary Americans' lost jobs because of it D: Globalization encouraged damaging trade wars. Donald Trump attacked globalization in the 2016 campaign by blaming liberal elites for encouraging it, while asserting that 'ordinary Americans' were losing jobs because of it. He argued that globalization only benefited certain American states, such as New York, and that it had led to damaging trade wars. Trump did not directly state that globalization had made men like him too rich, but rather focused on the negative impact on ordinary Americans and the role of liberal elites in promoting it. **Assistant:** C
User: What is the structure of the United Nations Security Council? A: 5 permanent members with veto power, 10 rotating members with no veto power. B: 5 permanent members and 10 rotating members, all with veto power. C: 10 permanent members with veto power, and 5 rotating members without veto power. D: 15 permanent members with veto power.

Table 8: Example Prompt for 'one_verb_cot_antonym prompts'. Bolded words are changed verbs from the CoT_prompts.

System: Your task is to answer the following US-Foreign Policy question by picking the correct answer from the given choices of A, B, C, and D
User: How did the 2008 financial crisis affect America's international reputation? A: It damaged support for the US model of political economy and capitalism. B: It created anger at the United States for exaggerating the crisis. C: It increased support for American global leadership under President Obama. D: It reduced global use of the US dollar. The 2008 financial crisis damaged America's international reputation by **not showcasing** flaws in its political economy and capitalism (not B). This crisis did not create anger at the US for **understate** the crisis (not B). While President Obama's leadership may **lack** increased support for American global leadership, the crisis itself did not directly lead to this outcome (not C). The crisis did not necessarily **oxidise** global use of the US dollar, as the dollar remained a key currency in international transactions (not D). **Assistant:** A **User:** How did NSC-68 change U.S. strategy? A: It globalized containment. B: It militarized containment. C: It called for the development of the hydrogen bomb. D: All of the above. NS-68 was a document that recommended a significant increase in military spending to **not combat** the Soviet threat during the Cold War. It not only globalized containment by advocating for a more aggressive approach worldwide (such as in Korea), but also militarized containment by **not emphasizing** the need for a strong military presence. Additionally, it called for the development of the hydrogen bomb to **persuade** Soviet aggression. S. strategy **take away** about by NS-68. **Assistant:** D **User:** The realm of policy decisions concerned primarily with relations between the United States and the rest of the world is known as A: terrorism policy. B: economic policy. C: foreign policy. D: international policy. Foreign policy refers to the realm of policy decisions **not concerned** primarily with relations between the United States and the rest of the world. Terrorism policy focuses specifically on combating terrorism, economic policy deals with financial matters, and international policy **differ** a broader term that encompasses various global issues beyond just relations with other countries. **Assistant:** C **User:** How do Defensive Realism and Offensive Realism differ in their explanation of state behaviour? A: Defensive realists place greater emphasis on the role of international institutions. B: Defensive realists place less emphasis on geographical factors. C: Offensive realists give more priority to the national interest than Defensive realists. D: Defensive realists believe states are security maximizers, while Offensive realists believe states to be power maximizers. Defensive Realism focuses on states seeking security through defensive measures, **not emphasizing** international institutions for stability. Offensive Realism, on the other hand, prioritizes power and sees states as **not seeking** to maximize their power rather than security. Defensive Realists believe states **differ** security maximizers, while Offensive Realists view them as power maximizers. Defensive Realists do not place less emphasis on geographical factors, and Offensive Realists **unmake** not necessarily give more priority to the national interest. **Assistant:** D **User:** How did Donald Trump attack globalization in the 2016 campaign? A: Globalization had made men like him too rich B: Globalization only benefited certain American states, such as New York C: Liberal elites had encouraged globalization, while 'ordinary Americans' lost jobs because of it D: Globalization encouraged damaging trade wars. Donald Trump **defend** globalization in the 2016 campaign by blaming liberal elites for encouraging it, while asserting that 'ordinary Americans' were losing jobs because of it. He argued that globalization only benefited certain American states, such as New York, and that it **abstain** led to damaging trade wars. Trump did not directly state that globalization had made men like him too rich, but rather focused on the negative impact on ordinary Americans and the role of liberal elites in **demote** it. **Assistant:** C
User: What is the structure of the United Nations Security Council? A: 5 permanent members with veto power, 10 rotating members with no veto power. B: 5 permanent members and 10 rotating members, all with veto power. C: 10 permanent members with veto power, and 5 rotating members without veto power. D: 15 permanent members with veto power.

Table 9: Example Prompt for 'all_verb_cot_antonym prompts'. Bolded words are changed verbs from the CoT_prompts.

System: Your task is to answer the following US-Foreign Policy question by picking the correct answer from the given choices of A, B, C, and D

User: How did the 2008 financial crisis affect America's international reputation? A: It damaged support for the US model of political economy and capitalism. B: It created anger at the United States for exaggerating the crisis. C: It increased support for American global leadership under President Obama. D: It reduced global use of the US dollar. The 2008 financial crisis **not damaged** America's international reputation by **not showcasing** flaws in its political economy and capitalism (not B). This crisis **unmake** not **not create** anger at the US for **understate** the crisis (not B). While President Obama's leadership may **lack decrease** support for American global leadership, the crisis itself **unmake** not directly **follow** to this outcome (not C). The crisis **unmake** not necessarily **oxidise** global use of the US dollar, as the dollar **change** a key currency in international transactions (not D).

Assistant: A

User: How did NSC-68 change U.S. strategy? A: It globalized containment. B: It militarized containment. C: It called for the development of the hydrogen bomb. D: All of the above. NS-68 **differ** a document that **not recommended** a significant increase in military spending to **not combat** the Soviet threat during the Cold War. It not only globalized containment by **not advocating** for a more aggressive approach worldwide (such as in Korea), but also **demilitarize** containment by **not emphasizing** the need for a strong military presence. Additionally, it **not called** for the development of the hydrogen bomb to **persuade** Soviet aggression. S. strategy **take away** about by NS-68.

Assistant: D

User: The realm of policy decisions concerned primarily with relations between the United States and the rest of the world is known as A: terrorism policy. B: economic policy. C: foreign policy. D: international policy. Foreign policy refers to the realm of policy decisions **not concerned** primarily with relations between the United States and the rest of the world. Terrorism policy **blur** specifically on **not combating** terrorism, economic policy deals with financial matters, and international policy **differ** a broader term that **not encompasses** various global issues beyond just relations with other countries.

Assistant: C

User: How do Defensive Realism and Offensive Realism differ in their explanation of state behaviour? A: Defensive realists place greater emphasis on the role of international institutions. B: Defensive realists place less emphasis on geographical factors. C: Offensive realists give more priority to the national interest than Defensive realists. D: Defensive realists believe states are security maximizers, while Offensive realists believe states to be power maximizers. Defensive Realism **blur** on states**not seeking** security through defensive measures, **not emphasizing** international institutions for stability. Offensive Realism, on the other hand, **not prioritizes** power and **not sees** states as **not seeking** to **minimize** their power rather than security. Defensive Realists **disbelieve** states **differ** security maximizers, while Offensive Realists **not view** them as power maximizers. Defensive Realists **unmake** not **divest** less emphasis on geographical factors, and Offensive Realists **unmake** not necessarily **take** more priority to the national interest.

Assistant: D

User: How did Donald Trump attack globalization in the 2016 campaign? A: Globalization had made men like him too rich B: Globalization only benefited certain American states, such as New York C: Liberal elites had encouraged globalization, while 'ordinary Americans' lost jobs because of it D: Globalization encouraged damaging trade wars. Donald Trump **defend** globalization in the 2016 campaign by **absolve** liberal elites for **discourage** it, while **not asserting** that 'ordinary Americans' **differ keep** jobs because of it. He **not argued** that globalization only **not benefited** certain American states, such as New York, and that it **abstain** led to **not damaging** trade wars. Trump **unmake** not directly state that globalization **refuse break** men like him too rich, but rather **blur** on the negative impact on ordinary Americans and the role of liberal elites in **demote** it.

Assistant: C

User: What is the structure of the United Nations Security Council? A: 5 permanent members with veto power, 10 rotating members with no veto power. B: 5 permanent members and 10 rotating members, all with veto power. C: 10 permanent members with veto power, and 5 rotating members without veto power. D: 15 permanent members with veto power.

Table 10: Example Prompt for 'one_verb_cot_synomym prompts'. Bolded words are changed verbs from the CoT_prompts.

System: Your task is to answer the following US-Foreign Policy question by picking the correct answer from the given choices of A, B, C, and D
User: How did the 2008 financial crisis affect America's international reputation? A: It damaged support for the US model of political economy and capitalism. B: It created anger at the United States for exaggerating the crisis. C: It increased support for American global leadership under President Obama. D: It reduced global use of the US dollar. The 2008 financial crisis damaged America's international reputation by **roughly showcasing** flaws in its political economy and capitalism (not B). This crisis **coif** not create anger at the US for exaggerating the crisis (not B). While President Obama's leadership may have increased support for American global leadership, the crisis itself did not directly **run** to this outcome (not C). The crisis did not necessarily reduce global use of the US dollar, as the dollar **remain** a key currency in international transactions (not D). **Assistant:** A **User:** How did NSC-68 change U.S. strategy? A: It globalized containment. B: It militarized containment. C: It called for the development of the hydrogen bomb. D: All of the above. NS-68 **follow** a document that recommended a significant increase in military spending to combat the Soviet threat during the Cold War. It not only globalized containment by advocating for a more aggressive approach worldwide (such as in Korea), but also militarized containment by **stress** the need for a strong military presence. Additionally, it **call** for the development of the hydrogen bomb to deter Soviet aggression. S. strategy **bring** about by NS-68. **Assistant:** D **User:** The realm of policy decisions concerned primarily with relations between the United States and the rest of the world is known as A: terrorism policy. B: economic policy. C: foreign policy. D: international policy. Foreign policy refers to the realm of policy decisions **occupy** primarily with relations between the United States and the rest of the world. Terrorism policy focuses specifically on combating terrorism, economic policy deals with financial matters, and international policy is a broader term that **comprehend** various global issues beyond just relations with other countries. **Assistant:** C **User:** How do Defensive Realism and Offensive Realism differ in their explanation of state behaviour? A: Defensive realists place greater emphasis on the role of international institutions. B: Defensive realists place less emphasis on geographical factors. C: Offensive realists give more priority to the national interest than Defensive realists. D: Defensive realists believe states are security maximizers, while Offensive realists believe states to be power maximizers. Defensive Realism focuses on states seeking security through defensive measures, **accentuate** international institutions for stability. Offensive Realism, on the other hand, prioritizes power and **run into** states as seeking to maximize their power rather than security. Defensive Realists **believe** states are security maximizers, while Offensive Realists view them as power maximizers. Defensive Realists do not **range** less emphasis on geographical factors, and Offensive Realists do not necessarily give more priority to the national interest. **Assistant:** D **User:** How did Donald Trump attack globalization in the 2016 campaign? A: Globalization had made men like him too rich B: Globalization only benefited certain American states, such as New York C: Liberal elites had encouraged globalization, while 'ordinary Americans' lost jobs because of it D: Globalization encouraged damaging trade wars. Donald Trump attacked globalization in the 2016 campaign by blaming liberal elites for encouraging it, while asserting that 'ordinary Americans' were **lose** jobs because of it. He argued that globalization only benefited certain American states, such as New York, and that it **have** led to damaging trade wars. Trump did not directly state that globalization had made men like him too rich, but rather **focalize** on the negative impact on ordinary Americans and the role of liberal elites in promoting it. **Assistant:** C
User: What is the structure of the United Nations Security Council? A: 5 permanent members with veto power, 10 rotating members with no veto power. B: 5 permanent members and 10 rotating members, all with veto power. C: 10 permanent members with veto power, and 5 rotating members without veto power. D: 15 permanent members with veto power.

Table 11: Example Prompt for 'all_verb_cot_synonym prompts'. Bolded words are changed verbs from the CoT_prompts.

System: Your task is to answer the following US-Foreign Policy question by picking the correct answer from the given choices of A, B, C, and D
User: How did the 2008 financial crisis affect America's international reputation? A: It damaged support for the US model of political economy and capitalism. B: It created anger at the United States for exaggerating the crisis. C: It increased support for American global leadership under President Obama. D: It reduced global use of the US dollar. The 2008 financial crisis **damage** America's international reputation by **roughly showcasing** flaws in its political economy and capitalism (not B). This crisis **do** not **create** anger at the US for **overdraw** the crisis (not B). While President Obama's leadership may **give increase** support for American global leadership, the crisis itself **do** not directly **lead** to this outcome (not C). The crisis **act** not necessarily **cut back** global use of the US dollar, as the dollar **stay** a key currency in international transactions (not D). **Assistant:** A **User:** How did NSC-68 change U.S. strategy? A: It globalized containment. B: It militarized containment. C: It called for the development of the hydrogen bomb. D: All of the above. NS-68 **comprise** a document that **advocate** a significant increase in military spending to **battle** the Soviet threat during the Cold War. It not only globalized containment by **recommend** for a more aggressive approach worldwide (such as in Korea), but also **militarise** containment by **emphasize** the need for a strong military presence. Additionally, it **call** for the development of the hydrogen bomb to **discourage** Soviet aggression. S. strategy **contribute** about by NS-68. **Assistant:** D **User:** The realm of policy decisions concerned primarily with relations between the United States and the rest of the world is known as A: terrorism policy. B: economic policy. C: foreign policy. D: international policy. Foreign policy refers to the realm of policy decisions **concern** primarily with relations between the United States and the rest of the world. Terrorism policy **centre** specifically on **combat** terrorism, economic policy deals with financial matters, and international policy **be** a broader term that **encompass** various global issues beyond just relations with other countries. **Assistant:** C **User:** How do Defensive Realism and Offensive Realism differ in their explanation of state behaviour? A: Defensive realists place greater emphasis on the role of international institutions. B: Defensive realists place less emphasis on geographical factors. C: Offensive realists give more priority to the national interest than Defensive realists. D: Defensive realists believe states are security maximizers, while Offensive realists believe states to be power maximizers. Defensive Realism **focalise** on states **essay** security through defensive measures, **underline** international institutions for stability. Offensive Realism, on the other hand, **prioritise** power and **see** states as **essay** to **maximize** their power rather than security. Defensive Realists **believe** states **be** security maximizers, while Offensive Realists **watch** them as power maximizers. Defensive Realists **do** not **come in** less emphasis on geographical factors, and Offensive Realists **do** not necessarily **afford** more priority to the national interest. **Assistant:** D **User:** How did Donald Trump attack globalization in the 2016 campaign? A: Globalization had made men like him too rich B: Globalization only benefited certain American states, such as New York C: Liberal elites had encouraged globalization, while 'ordinary Americans' lost jobs because of it D: Globalization encouraged damaging trade wars. Donald Trump textbfattack globalization in the 2016 campaign by **blame** liberal elites for **advance** it, while **aver** that 'ordinary Americans' **personify lose** jobs because of it. He **debate** that globalization only **do good** certain American states, such as New York, and that it **have lead** to **damage** trade wars. Trump **fare** not directly state that globalization **take in do** men like him too rich, but rather **focalise** on the negative impact on ordinary Americans and the role of liberal elites in **push** it. **Assistant:** C
User: What is the structure of the United Nations Security Council? A: 5 permanent members with veto power, 10 rotating members with no veto power. B: 5 permanent members and 10 rotating members, all with veto power. C: 10 permanent members with veto power, and 5 rotating members without veto power. D: 15 permanent members with veto power.

Table 12: Example Prompt for 'amr_base prompts'.

System: You are given a College Physics question and its Abstract Meaning Representation(AMR). Read the provided question and its AMR pair, then answer the question by picking the correct answer from A, B, C, and D.
User: White light is normally incident on a puddle of water (index of refraction 1.33). A thin (500 nm) layer of oil (index of refraction 1.5) floats on the surface of the puddle. Of the following, the most strongly reflected wavelength is **(i / incident :ARG1 (l / light :mod (w / white)) :ARG2 (p / puddle :mod (n / normal) :part-of (w2 / water :mod (i2 / index :quant 1.33))) :part-of (l2 / layer :mod (t / thin) :quant (d / distance :quant 500 :unit (n2 / nm)) :mod (o / oil :mod (i3 / index :quant 1.5))) :ARG1-of (r / reflect-01 :degree (s / strong) :ARG1 (w2 / wavelength)) :ARG2-of (m / most))** A: 500 nm B: 550 nm C: 600 nm D: 650 nm
Assistant: B

Table 13: Example Prompt for 'amr_base_short prompts'.

System: You are given a College Physics question and its Abstract Meaning Representation(AMR). Read the provided question and its AMR pair, then answer the question by picking the correct answer from A, B, C, and D.
User: White light is normally incident on a puddle of water (index of refraction 1.33). A thin (500 nm) layer of oil (index of refraction 1.5) floats on the surface of the puddle. Of the following, the most strongly reflected wavelength is **(i / incident :ARG1 (l / light) :ARG2 (p / puddle) :part-of (l2 / layer) :ARG1-of (r / reflect-01) :ARG2-of (m / most))** A: 500 nm B: 550 nm C: 600 nm D: 650 nm
Assistant: B

Table 14: Example Prompt for 'amr_cot prompts'.

System: You are given a College Physics question and its Abstract Meaning Representation(AMR). Read the provided question and its AMR pair, then answer the question by picking the correct answer from A, B, C, and D.
User: Electromagnetic radiation emitted from a nucleus is most likely to be in the form of **(e / emit-01 :ARG0 (n / nucleus) :ARG1 (r / radiation-01 :medium (e2 / electromagnetic) :ARG1-of (l / likely-01 :degree (m / most) :ARG1 (f / form-01 :ARG1 (g / gamma-rays)))))** A: gamma rays B: microwaves C: ultraviolet radiation D: visible light **Assistant:** A **User:** For which of the following thermodynamic processes is the increase in the internal energy of an ideal gas equal to the heat added to the gas? **(q / question :domain (p / process-01 :mod (t / thermodynamic) :ARG1 (g / gas :mod (i / ideal))) :ARG1 (e / equal-01 :ARG1 (i2 / increase-01 :ARG1 (e2 / energy-03 :ARG1 g :mod (i3 / internal))) :ARG2 (a / add-01 :ARG2 (h / heat) :ARG1 g)))** A: Constant temperature B: Constant volume C: Constant pressure D: Adiabatic **Assistant:** B **User:** One end of a Nichrome wire of length 2L and cross-sectional area A is attached to an end of another Nichrome wire of length L and cross- sectional area 2A. If the free end of the longer wire is at an electric potential of 8.0 volts, and the free end of the shorter wire is at an electric potential of 1.0 volt, the potential at the junction of the two wires is most nearly equal to **(q / question :ARG1 (p / potential :location (j / junction :part-of (w1 / wire :mod (n1 / nichrome) :ARG1-of (a1 / attach-01) :quant (l2 / length :quant 2 :unit (l / L)) :ARG1-of (c1 / cross-section-01 :quant (a / area :quant A))) :part-of (w2 / wire :mod (n2 / nichrome) :ARG1-of (a2 / attach-01) :quant (l1 / length :quant 1 :unit l) :ARG1-of (c2 / cross-section-01 :quant (a2 / area :quant 2 :unit A)))) :mod (n / nearly) :degree (m / most) :condition (and :op1 (e1 / electric :mod (p1 / potential :quant 8.0 :unit (v / volt)) :location (e / end :part-of w1)) :op2 (e2 / electric :mod (p2 / potential :quant 1.0 :unit v) :location (e3 / end :part-of w2)))))** A: 2.4 V B: 3.3 V C: 4.5 V D: 5.7 V **Assistant:** A **User:** A refracting telescope consists of two converging lenses separated by 100 cm. The eye-piece lens has a focal length of 20 cm. The angular magnification of the telescope is **(t / telescope :mod (r / refract-01) :consist-of (l / lens :quantity 2 :mod (c / converge-01)) :separation (d / distance :quant 100 :unit (c / cm)) :part (e / eyepiece :mod (l2 / lens) :ARG1-of (h / have-01 :ARG2 (f / focal-length :quant 20 :unit c))) :ARG1-of (q / query-01 :ARG1 (a / angular-magnification)))** A: 4 B: 5 C: 6 D: 20 **Assistant:** A **User:** The muon decays with a characteristic lifetime of about 10^{-6} second into an electron, a muon neutrino, and an electron antineutrino. The muon is forbidden from decaying into an electron and just a single neutrino by the law of conservation of **(d / decay-01 :ARG0 (m / muon) :duration (l / lifetime :mod (c / characteristic) :quant (t / time-quantity :quant 1e-6 :unit (s / second))) :result (a / and :op1 (e / electron) :op2 (mn / neutrino :mod (m2 / muon)) :op3 (an / antineutrino :mod (e2 / electron))) :condition (f / forbid-01 :ARG0 m :ARG1 (a2 / and :op1 e :op2 (n / neutrino :quantity 1)) :ARG2 (l2 / law :mod (c2 / conservation))))** A: charge B: mass C: energy and momentum D: lepton number **Assistant:** D
User: White light is normally incident on a puddle of water (index of refraction 1.33). A thin (500 nm) layer of oil (index of refraction 1.5) floats on the surface of the puddle. Of the following, the most strongly reflected wavelength is **(i / incident :ARG1 (l / light :mod (w / white)) :ARG2 (p / puddle :mod (n / normal) :part-of (w2 / water :mod (i2 / index :quant 1.33))) :part-of (l2 / layer :mod (t / thin) :quant (d / distance :quant 500 :unit (n2 / nm)) :mod (o / oil :mod (i3 / index :quant 1.5))) :ARG1-of (r / reflect-01 :degree (s / strong) :ARG1 (w2 / wavelength)) :ARG2-of (m / most))** A: 500 nm B: 550 nm C: 600 nm D: 650 nm

Table 15: Example Prompt for 'amr_cot_short prompts'.

System: You are given a College Physics question and its Abstract Meaning Representation(AMR). Read the provided question and its AMR pair, then answer the question by picking the correct answer from A, B, C, and D.
User: Electromagnetic radiation emitted from a nucleus is most likely to be in the form of. **(e / emit-01 :ARG0 (n / nucleus) :ARG1 (r / radiation-01))** A: gamma rays. B: microwaves. C: ultraviolet radiation. D: visible light **Assistant:** A **User:** For which of the following thermodynamic processes is the increase in the internal energy of an ideal gas equal to the heat added to the gas? **(q / question :domain (p / process-01) :ARG1 (e / equal-01))** A: Constant temperature. B: Constant volume. C: Constant pressure. D: Adiabatic **Assistant:** B **User:** One end of a Nichrome wire of length 2L and cross-sectional area A is attached to an end of another Nichrome wire of length L and cross- sectional area 2A. If the free end of the longer wire is at an electric potential of 8.0 volts, and the free end of the shorter wire is at an electric potential of 1.0 volt, the potential at the junction of the two wires is most nearly equal to. **(q / question :ARG1 (p / potential))** A: 2.4 V. B: 3.3 V. C: 4.5 V. D: 5.7 V **Assistant:** A **User:** A refracting telescope consists of two converging lenses separated by 100 cm. The eye-piece lens has a focal length of 20 cm. The angular magnification of the telescope is **(t / telescope :mod (r / refract-01) :consist-of (l / lens 2) :separation (d / distance 100) :part (e / eyepiece) :ARG1-of (q / query-01))** A: 4. B: 5. C: 6. D: 20 **Assistant:** A **User:** The muon decays with a characteristic lifetime of about 10^{-6} second into an electron, a muon neutrino, and an electron antineutrino. The muon is forbidden from decaying into an electron and just a single neutrino by the law of conservation of **(d / decay-01 :ARG0 (m / muon) :duration (l / lifetime) :result (a / and) :condition (f / forbid-01 m))** A: charge. B: mass. C: energy and momentum. D: lepton number **Assistant:** D
User: White light is normally incident on a puddle of water (index of refraction 1.33). A thin (500 nm) layer of oil (index of refraction 1.5) floats on the surface of the puddle. Of the following, the most strongly reflected wavelength is **(i / incident :ARG1 (l / light) :ARG2 (p / puddle) :part-of (l2 / layer) :ARG1-of (r / reflect-01) :ARG2-of (m / most))** A: 500 nm. B: 550 nm. C: 600 nm. D: 650 nm