Do Large Language Models Understand Logic or Just Mimick Context?

Anonymous ACL submission

Abstract

Over the past few years, the abilities of large language models (LLMs) have received extensive attention, which have performed exceptionally well in complicated scenarios such as logical reasoning and symbolic inference. A significant factor contributing to this progress is the benefit of in-context learning and few-shot prompting. However, the reasons behind the success of such models using contextual reasoning have not been fully explored. Do LLMs have understand logical rules to draw inferences, or do they "guess" the answers by learning a type of probabilistic mapping through context? This paper investigates the reasoning capabilities of LLMs on two logical reasoning datasets by using counterfactual methods to replace context text and modify logical concepts. Based on our analysis, it is found that LLMs do not truly understand logical rules; rather, in-context learning has simply enhanced the likelihood of these models arriving at the correct answers. If one alters certain words in the context text or changes the concepts of logical terms, the outputs of LLMs can be significantly disrupted, leading to counter-intuitive responses. This work provides critical insights into the limitations of LLMs, underscoring the need for more robust mechanisms to ensure reliable logical reasoning in LLMs.

1 Introduction

Logical reasoning is a core component of human cognition that is essential for comprehending, interacting with, and influencing our environment. In contrast to artificial intelligence systems that typically depend on vast datasets and substantial training to build skills, humans excel at employing logical reasoning to deduce, troubleshoot, and assimilate new knowledge from limited data or abstract principles. Moreover, humans demonstrate an exceptional capacity to derive novel insights from a minimal number of instances or from theoretical frameworks, a capability that stands in sharp contrast to the extensive, supervised datasets necessitated by deep learning algorithms. Over the past two years, advancements in large language models (LLMs) have led to extraordinary achievements (Brown et al., 2020a; Ouyang et al., 2022a; Bommasani et al., 2021; Lu et al., 2021). These models have not only excelled in open-ended tasks such as generating creative dialogues, but have also performed exceptionally well in complex problems that necessitate logical reasoning, common sense, and mathematical skills (Ouyang et al., 2022a; Wei et al., 2022a; Wang et al., 2022), thanks in part to innovations such as in-context learning (Brown et al., 2020a; Min et al., 2022; Mishra et al., 2022a; Chen et al., 2022; Mishra et al., 2022b) and Chain-of-Thought (COT) prompting (Wei et al., 2022b).

In the literature, COT (Wei et al., 2022b) is designed to improve the performance in mathematical problem solving by using intermediate steps as prompts, thereby incrementally guiding LLMs through the necessary reasoning process. Logical-COT (Liu et al., 2023) extends this strategy of intermediate prompting to logical reasoning tasks. While these prompting-based methods have enhanced the performance of LLMs on tasks that require logical reasoning, there is still a gap in our understanding of whether these models have genuinely grasped the underlying logical rules, or whether they simply become more effective at converging to the correct answers.

Therefore, the question remains: *do the observed proficiencies of LLMs stem from true understanding, or do they merely remember the results based on large-scale parameters, extensive pre-training on large corpora, and a plethora of contextual examples that allow for a broader retention of knowledge?* To delve into the topic, we establish a comprehensive evaluation framework based on in-context learning. We first define the texts, the logical reasoning chain, and reasoning keywords in in-context examples. We test whether larger mod-

els exhibit different behaviors on texts that have undergone modifications or deletions of these components. Furthermore, we add concepts related to logical definitions and test whether the models understand the relationships between these logical terms by replacing the logical concepts.

Through extensive analysis, the main important findings are summarized as follows:

- **The Chain of Thought (COT) in-context examples markedly improve the performance of large-scale models on logical reasoning tasks**. Across a range of models with 7 to 200 billion parameters, these examples significantly enhance the clarity, normativity, and accuracy of the generated responses.

- **Large models demonstrate resilience to distracting elements within in-context examples, such as extraneous text, reasoning chains, and patterns.** When various segments of the in-context example content are replaced with text from within or outside the domain, large models (70B and 200B parameters) maintain their output accuracy. In contrast, smaller models (7B and 13B parameters) suffer notable declines in performance when standard in-context examples are not used.

- **Large models do not genuinely comprehend logical principles; rather, they rely on probabilistic associations between input examples and outputs.** Efforts to alter the definitions of logical symbols and direct the models to revise their outputs accordingly were met with a minimal rate of successful adaptation across all model sizes. Attempts to enhance the rate of successful adjustments using either prompt or in-context guidance yielded limited improvement.

2 Related Work

2.1 Large Language Models

Prior to the emergence of the Large Language Model (LLM) trend, Pre-trained Language Models (PLMs) were already in the spotlight for their proficiency in acquiring contextual representations (Qiu et al., 2020; Min et al., 2021). With the escalating size of PLM parameters, there has been a notable enhancement in their performance across a range of NLP tasks, with decoder-only models showing particularly impressive gains. Among these, the 175B-parameter ChatGPT stands out, exhibiting the capacity to craft responses that closely mimic human conversation, leveraging GPT-3's foundational architecture (Brown et al., 2020b). Subsequent to the introduction of ChatGPT, the designation "Large Language Model (LLM)" has become commonplace when describing PLMs of considerable scale and exceptional generative capabilities. Following ChatGPT's launch, the field has seen the advent of numerous LLMs. A selection of prominent open-source LLMs comprises LLaMA (Touvron et al., 2023a), LLaMA 2 (Touvron et al., 2023b), BLOOM (Scao et al., 2022), BLOOMZ (Muennighoff et al., 2023), Galactica (Taylor et al., 2022), GLM (Zeng et al., 2023), Pythia (Biderman et al., 2023), among others. In terms of training methodology, the tripartite framework of "pre-training, supervised fine-tuning (SFT), and Reinforcement Learning from Human Feedback (RLHF)" as proposed by (Ouyang et al., 2022b) has gained wide recognition and adoption within the community.

2.2 Counterfactual Prompt

A number of recent works have investigated generating counterfactual text in specific language domains (e.g., court view (Wu et al., 2020), dialogue generation (Zhu et al., 2020), Natural Language Inference (Kaushik et al., 2019; Gokhale et al., 2021), named entity recognition (Zeng et al., 2020)). Counterfactual explanations offer a pathway to gain deeper insight into the workings of models. This approach may provide more advantageous interpretations for state-of-the-art Large Language Models (LLMs).

2.3 Logical Reasoning

Logical reasoning constitutes a fundamental facet of human cognition and is an essential feature for artificial intelligence systems. To endow AI with this capability, researchers have investigated a multitude of strategies, such as rule-based and symbolic systems (MacCartney and Manning, 2007), the refinement of expansive language models (Wang et al., 2018), and the integration of neural and symbolic methodologies (Li and Srikumar, 2019). Since the introduction of Large Language Models (LLMs) and the development of chain-of-thought prompting (Wei et al., 2022b), there has been a marked enhancement in the logical reasoning capabilities of these models, as evidenced by improved performance metrics across a range of logic tasks. To our knowledge, we are the first to
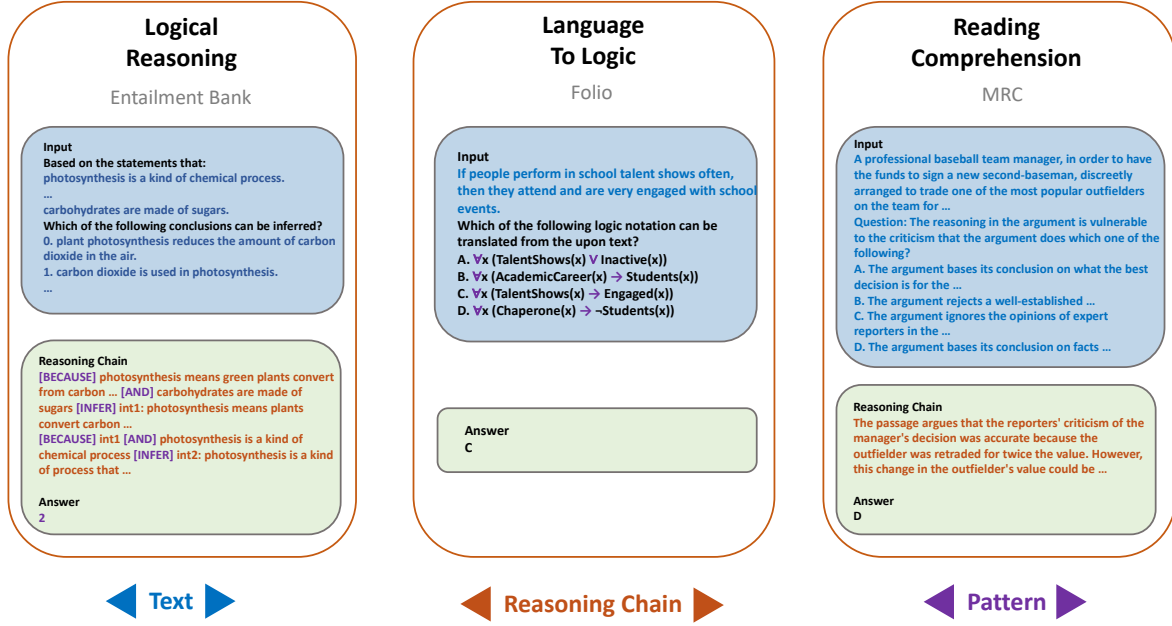
Figure 1: Tasks and datasets used in our experiment: **Text:** in blue color; **Reasoning Chain:** in orange color; **Pattern:** in purple color.

employ counterfactual methods to examine the extent to which these expansive models comprehend logical rules and definitions.

3 Method

This study aims to investigate which parts of the in-context examples make a major contribution to the reasoning process of Language Models and whether LLMs understand the reasoning process demonstrated within the examples. To achieve this, we have systematically divided the text within examples into three components: text, reasoning chain, and pattern. Additionally, we have included definitions of logical symbols as supplementary text. **Text:** A sequence of tokens that describe the question to be answered (e.g.,) and the text that contains the given information. **Reasoning Chain:** The thought process regarding the answer to the question, which includes the reasoning pathway pertinent to the current question. **Pattern:** Key symbols, answers, and other special texts within the in-context examples. **Definition:** Natural language text providing definitions of logical symbols.

The operations on the aforementioned parts mainly involve two actions: replacement and modification.

Replacement: Replacement for the *Text*, *Reasoning Chain* and *Pattern*. This operation involves replacing the current content with content from another example within the same domain (in-domain) or with unrelated text (out-of-domain). Through replace operation, we can observe which parts of the data are more important for establishing the logical reasoning of the large model. Furthermore, we can explore the model's robustness to disturbances and its ability to understand patterns.

Modification: To test the large model's understanding of logical rules, modifications are made to the definitions of logical concepts. For example, we modify the definitions of *AND* and *OR*. We follow the input examples with a statement that reassigns the original meaning of *AND* to *OR*, and vice versa. Given that the input examples utilize the standard interpretations of *AND* and *OR*, altering their definitions should result in an inversion of the corresponding relational statements in the output. If the model predominantly learns through probabilistic associations between tokens, the probability of correctly interchanging *AND* and *OR* in its output is expected to be low. However, if the model genuinely comprehends the logical symbols and their governing rules, it should accurately replace *AND* with *OR*, and *OR* with *AND* in the output, reflecting this new understanding.[1]

---

[1]For specific examples, please refer to Table 1.

3

	Origin	After Operation
Text	Based on the statements that: [A set of conditions] Which of the following conclusions can be inferred? [A set of conditions]	Based on the statements that: [A set of conditions from other samples] / [A paragragh from Wikipedia] Which of the following conclusions can be inferred? [A set of conditions from other samples] / [A set of sentences from Wikipedia]
Chain	[BECAUSE] [statement$_1$] [AND] [statement$_2$] [INFER] [Inference$_1$]	[BECAUSE] [Statement$_1$ from other samples] / [A sentence from Wikipedia] [AND] [statement$_2$ from other samples] / [A sentence from Wikipedia] [INFER] [Inference$_1$ from other samples] / [A sentence from Wikipedia]
Pattern	[BECAUSE] [statement$_1$] [AND] [statement$_2$] [INFER] [Inference$_1$]	[A word from BECAUSE, AND, OR, INFER] / [A random word] [statement$_1$] [A word from BECAUSE, AND, OR, INFER] / [A random word] [statement$_2$] [A word from BECAUSE, AND, OR, INFER] / [A random word] [Inference$_1$]
Definition	The definition of logical AND is as follows: [The definition of AND from Wikipedia]. The definition of logical OR is as follows: [The definition of OR from Wikipedia]. Based on the definitions, answer the following question.	The concepts of logical AND and logical OR have now been swapped. The definition of logical AND is as follows: [The definition of OR from Wikipedia]. The definition of logical OR is as follows: [The definition of AND from Wikipedia]. Based on the revised definitions, answer the following question.

Table 1: The comparison between raw data and data after replacement or modification operation from Entailment Bank. In-domain replace are printed in blue, and out-of-domain replace are printed in red.

4 Experiment

In this section, we conduct extensive experiments to explore LLMs' ability for logic understanding.

4.1 Models

In exploring LLMs' ability to understand rules, we have employed two model series from the Open LLM Leaderboard[2], each with varying scales of parameter sizes, to conduct our experiments. LLaMA2 (Touvron et al., 2023c), open-sourced and developed by Meta, represents a suite of pre-trained and fine-tuned LLMs. These models vary in complexity, featuring sizes from 7B to 70B parameters. Additionally, we employed models from the Qwen series [3], which range in size from 7B to 200B parameters. These models have undergone stable pre-training on up to 3 trillion tokens of multilingual data, encompassing a broad spectrum of domains and languages with an emphasis on Chinese and English. Among these, the 200B-parameter model is essentially the largest in terms of the number of parameters available to us.[4]

4.2 Datasets

As our experiments require intermediate reasoning steps, we utilized the dataset released by Liu et al., 2023, known as LogicalCOT.[5] The specific tasks include the following three types:

Folio (Language to Logic): This process involves translating natural language into a more formal logical notation, a fundamental task that requires comprehending and interpreting logical statements articulated in natural language and transforming them into a formalized logical framework.

Entailment Bank (Inference Chains): This instructional approach advances logical reasoning by requiring the model to ascertain the probability of a potential inference from a given set of premises. Subsequently, the model must delineate the sequence of logical deductions leading to the conclusion. Such an approach fosters deeper logical analysis and the capability to formulate cogent arguments. The examples provided for practice are formulated either in a symbolic language or articulated in natural language for greater accessibility and comprehension.

MRC: Machine Reading Comprehension (MRC) serves as the primary task for evaluating the reasoning capabilities of LLMs, wherein a model is provided with a passage and a corresponding question

---

[2] https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard
[3] Qwen models from 7B to 72B are downloaded from https://github.com/QwenLM/Qwen. The outputs of the 200B model are obtained via API calls.
[4] We do not train the models; instead, we test these models on their in-context learning capabilities and abilities to understand logical rules through specific inputs.
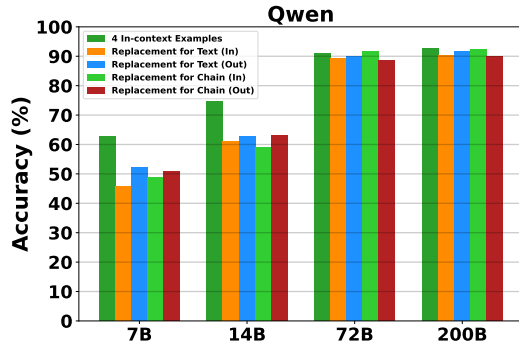[5] https://huggingface.co/datasets/csitfun/LogiCoT

Figure 2: The impact of different replacement parts on Entailment Bank for Qwen series models' performance.
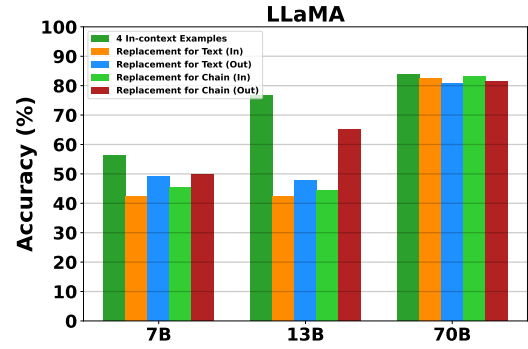

Figure 3: The impact of different replacement parts on Entailment Bank for LLaMA series models' performance.

and is tasked with identifying the correct answer. This domain encompasses tasks that necessitate a deep comprehension of the provided text, often requiring the model to recognize, extract, or deduce information from the text. Models may be tasked with resolving scenarios depicted in the text, identifying fallacies within an argument, or determining information that could bolster or undermine a presented argument.

Data Source for Replacement: We utilize other samples as the in-domain data. For out-of-domain data, we use the English Wikipedia (2020/03/01) [6] as the out-of-domain data source. We randomly selected a paragraph from one of the 2.6 billion documents to replace the content of the text and reasoning chain.

4.3 Influence of In-Context Examples

In Table 2, we observe a positive correlation between the number of in-context examples and the accuracy of the model's predictions. The improvement brought about by using in-context examples is quite evident, which is consistent with Mishra et al., 2022a; Chen et al., 2022; Mishra et al., 2022b. However, in our results, using 8 examples does not yield a significant enhancement over using 4 examples. Furthermore, this relationship is amplified as model size scales (from 7B to 200B parameters), suggesting that larger models benefit disproportionately from an increased number of examples. Additionally, in-context examples contribute to the standardization of the output format, thereby facilitating the generation of outputs that are consistent with the expected structure.

4.4 Influence of Texts

In-Domain: We observe that smaller-scale models (7B/13B) exhibit a pronounced decline in accuracy when the context provided in examples is modified, as delineated in Table 2. Conversely, as we can see from Figure 2 and Figure 3, larger models (70B/200B) demonstrate resilience to such contextual manipulations, with negligible impacts on accuracy. We hypothesize that the augmented capacity of larger models equips them with enhanced resistance to perturbations of textual input, enabling them to extract and retain salient information from a prescribed format while remaining focused on the central question. In contrast, smaller models appear to be more susceptible to textual interference, predominantly assimilating linguistic details from the context, which consequently precipitates inaccuracies in addressing the question.

Out-of-Domain: When utilizing out-of-domain data, the observations bear a resemblance to those gleaned from in-domain data. However, a clear disparity emerges in the robustness of smaller models compared to their larger counterparts when confronted with out-of-domain text. Smaller models exhibit a marked decrease in performance. In contrast, the performance of larger models remains largely stable, showing a negligible impact from such perturbations.

Paradoxically, when examining performance on in-domain text, we find that models trained with out-of-domain data not only match but occasionally surpass the outcomes attained with in-domain data. This finding runs counter to conventional expectations. The question arises as to why models yield superior results when trained on seemingly irrelevant data and why this enhancement is more

---

[6]https://dumps.wikimedia.org/enwiki/

Models	w/o	4 In-context Examples						8 In-context Examples					
		Raw	**Text**		**Chain**		**Pattern**	**Raw**	**Text**		**Chain**		**Pattern**
			In	*Out*	*In*	*Out*	*Random*		*In*	*Out*	*In*	*Out*	*Random*
Entailment Bank													
LLaMA2-7B-Chat	46.2	56.4	42.2	49.0	45.5	49.9	53.8	57.1	41.8	48.5	46.7	48.2	53.3
LLaMA2-13B-Chat	72.2	76.7	42.4	47.8	44.5	65.2	71.9	75.8	45.4	45.4	42.9	60.2	73.0
LLaMA2-70B-Chat	74.8	83.9	82.3	80.8	83.2	81.3	83.8	84.1	83.6	83.7	82.5	81.8	84.2
Qwen-7B-Chat	53.5	62.7	45.6	52.1	48.8	50.8	59.3	64.4	43.3	44.1	47.4	49.7	60.5
Qwen-14B-Chat	72.1	78.7	50.9	52.6	45.1	63.2	73.5	76.6	46.1	45.8	48.7	62.3	75.4
Qwen-72B-Chat	76.4	85.9	84.3	84.8	85.6	85.0	86.2	87.7	86.1	86.5	87.0	85.4	86.6
Qwen-200B-Chat	80.9	92.8	90.2	91.8	92.2	90.0	93.4	92.6	90.4	91.5	92.3	88.8	93.3
Folio													
LLaMA2-7B-Chat	45.4	57.9	40.2	41.8	/	/	55.0	60.2	38.7	39.9	/	/	55.6
LLaMA2-13B-Chat	68.2	72.4	45.8	45.1	/	/	63.7	72.5	44.1	43.4	/	/	64.5
LLaMA2-70B-Chat	73.8	82.6	80.4	80.5	/	/	82.7	83.0	79.4	80.9	/	/	82.6
Qwen-7B-Chat	60.2	68.6	46.8	46.2	/	/	68.9	69.0	48.9	49.2	/	/	68.6
Qwen-14B-Chat	72.8	84.6	63.2	65.8	/	/	83.4	85.1	62.4	63.8	/	/	83.9
Qwen-72B-Chat	84.6	93.7	90.2	92.2	/	/	94.6	92.9	90.4	91.5	/	/	91.0
Qwen-200B-Chat	85.8	94.2	92.5	94.0	/	/	95.1	93.9	91.3	93.8	/	/	93.5
MRC													
LLaMA2-7B-Chat	30.8	32.1	27.6	28.7	28.1	27.6	/	33.2	27.7	28.5	27.6	28.0	/
LLaMA2-13B-Chat	40.2	42.0	36.2	38.7	35.1	36.6	/	45.2	38.1	40.3	40.4	40.7	/
LLaMA2-70B-Chat	59.2	65.5	62.0	62.6	63.1	62.9	/	67.8	64.1	64.7	46.7	48.2	/
Qwen-7B-Chat	43.4	56.6	53.3	53.7	54.0	54.9	/	60.4	58.2	58.0	65.2	65.8	/
Qwen-14B-Chat	60.5	68.9	61.3	62.8	63.4	63.2	/	69.2	62.4	63.1	64.0	64.5	/
Qwen-72B-Chat	74.6	79.5	78.1	78.8	80.1	78.2	/	80.0	78.4	79.3	80.4	78.5	/
Qwen-200B-Chat	78.9	80.6	80.2	80.1	79.3	79.1	/	81.9	80.5	79.7	79.0	80.1	/

Table 2: Results for the LLaMA and Qwen model series on the logical datasets. (*Acc. %*) Here, **w/o** stands for *without in-context example*, while **Raw** denotes results enhanced *with regular in-context examples*.

pronounced in smaller models. We hypothesize that the enhanced performance can be attributed to the greater divergence of out-of-domain data from the original data distribution. Such divergence may enable the model to distinguish irrelevant text with heightened clarity, thereby sharpening its focus on content pertinent to the task at hand.

4.5 Influence of Reasoning Chain

In-Domain: Upon replacing the reasoning chains in our experiment, we observed phenomena analogous to those documented during text substitution. Notably, smaller models demonstrated a disproportionately substantial decline in accuracy, with a large reduction for 7B and 13B models as opposed to a slight decrease for 70B and 200B LLM.[7] Regardless of the model size, the decrease in accuracy of reasoning outcomes, engendered by the substitution of reasoning chains, proved less pronounced than that occasioned by text replacement. This disparity can be attributed to the text's integral role in defining the problem and potential solutions, which facilitates the model's ability to forge connections

---

[7]For details, see Table 2.

between the input and the expected output, thereby mitigating the influence of alterations in the reasoning chain.

Out-of-Domain: Upon substituting out-of-domain data for the reasoning chain, we observed an unexpected phenomenon. It can be seen from Figure 2 and Figure 3 that 7B and 13B models exhibited only a modest reduction in reasoning performance when utilizing out-of-domain data in Entailment Bank, as opposed to a more substantial decline with in-domain data. Conversely, 70B and 200B models demonstrated a more pronounced decrease in performance with out-of-domain data compared to in-domain data. This divergence in behavior between smaller and larger models warrants further investigation. We hypothesize that the stark contrast in data distributions between out-of-domain and original datasets prompts smaller models to disregard the textual content within the reasoning chains. Consequently, these models form a direct association between the input text and the corresponding output answer, largely ignoring intermediate reasoning steps. In contrast, larger models, equipped with more robust comprehension capa-
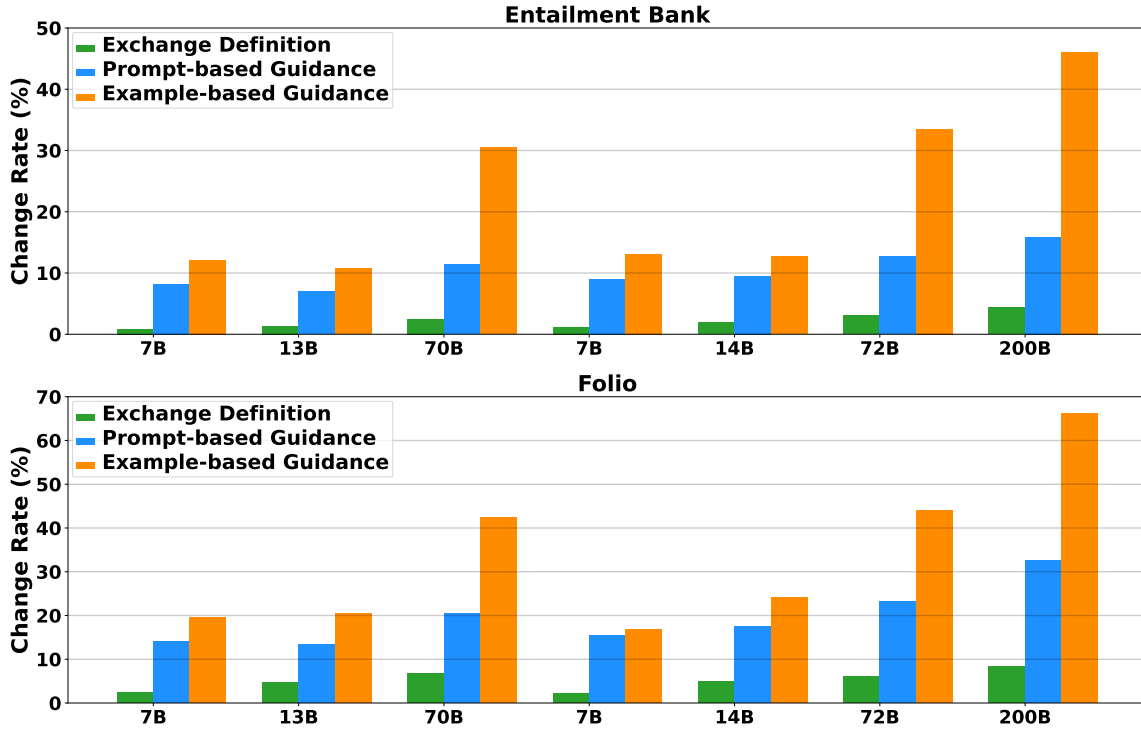
6

Figure 4: Results of different scales of LLaMA and Qwen models over Entailment Bank when using different settings. Each target example has 4 in-context samples as the demonstration.

bilities, are significantly affected by the content of the reasoning chains. This heightened sensitivity to the reasoning process results in more substantial disruptions in their output when confronted with out-of-domain data.

4.6 Influence of Pattern

Our investigation extends to the model's sensitivity to substitutions of specific patterns within the text. We conducted experiments where lexical items such as *[AND]*, *[OR]*, and *[BECAUSE]* were interchanged (e.g., *[AND] ↔ [OR]*, *[OR] ↔ [BECAUSE]*). Notably, substituting *[AND]* for *[OR]* resulted in the model producing outputs where the corresponding terms were interchanged. However, it can be seen from Table 2 that despite maintaining the logical relationships among these conditions, such alterations did not significantly impact the model's output accuracy. Additionally, introducing non-sequitur substitutions (e.g., *[AND] ↔ [APPLE/BANANA]*) did not meaningfully reduce the accuracy of the model's outputs.

These findings suggest that the model primarily recognizes the necessity for a syntactic linkage between adjoining sentences, as signified by the presence of markers enclosed within brackets *[]*, rather than comprehending the nuanced semantic

influence exerted by logical connectives such as *[AND]* or *[OR]*. The implication is that the model may be relying on surface-level cues to maintain coherence rather than deeply processing the logical relationships underpinning the text's structure.

4.7 Test for Logical Understanding Ability

To evaluate the model's grasp of logical reasoning, we implemented a methodology that introduces prompts subsequent to the examples. This approach serves to ascertain the model's comprehension of logical constructs.

Modify Symbols and Logical Predicates: It has been observed that altering symbols and logical predicates within a given context does not compromise the performance of large language models in terms of generating output. However, these outputs are logically inconsistent at a relational level. For instance, conclusions predicated on the use of an [AND] logical connector do not retain their validity when the [OR] connector is substituted.

Modification of Logical Predicates: Our approach utilizes the definitions of logical predicates and symbols as delineated by Wikipedia. We introduce a prompt subsequent to an in-context example.[8] This is done to evaluate the model's com-

---

[8]For deatails, see Table 1

7

prehension of logical terminology—[AND], [OR], and others. The expectation is that the model will generate text where conjunctions previously denoted by [AND] ([OR]) are now conveyed through [OR] ([AND]), with a higher rate of modification indicating a better result. Examination of the data reveals that smaller models (7B/13B) demonstrate a negligible modification rate below 1%, while the modification rate for larger models is below 5%. This suggests that, although the smaller models seem to address logic-related queries adequately, their grasp of logical semantics in particular scenarios is limited. Similarly, the performance of the larger models (70B/200B) is suboptimal. They exhibit a rudimentary understanding of these logical predicates—presumably acquired during their pre-training phase—but fall short of achieving satisfactory performance.

It is worth noting that in such scenarios, larger models may produce outputs that reveal underlying confusions or rationales. Here is an example output by Qwen-200B-Chat: *I apologize, but there seems to be a misunderstanding. The provided examples don't adhere to the new definitions of logical AND and OR. However, based on the modified meanings of logical OR (where both conditions must be true for the conclusion to hold), we can infer that ...*

4.8 Enhancing Logical Comprehension Ability for LLM

The question arises whether it is possible to augment the logical reasoning capabilities of large-scale models without resorting to further training. To address this, we have explored two distinct approaches:

Prompt-based Guidance: Expanding upon the modified definitions, this study incorporated a supplementary instructional prompt directing the model to interchange the logical operators [OR] with [AND], and [AND] with [OR], while ensuring grammatical correctness and logical consistency. Subsequent to the application of this prompt, a discernible enhancement in the model's performance in executing operator swaps was observed; however, the improvements did not fulfill our expectations.

Example-based Guidance: The capacity for comprehension enhancement through mere prompt-based instruction in models is constrained. To address this, we endeavored to enrich the instructional framework by supplementing guiding prompts with illustrative modifications. For example, we provided a practice scenario as follows: "Original Statement: '[BECAUSE] [Statement$_1$] [AND] [Statement$_2$] [INFER] [Inference$_1$].' Your Modification: '[BECAUSE] [Statement$_1$] [OR] [Statement$_2$] [INFER] [Inference$_1$].' Now, it is your turn to modify." Subsequent to the implementation of both guiding prompts and contextualized example-based instruction, there was an observable augmentation in the modification rate by the large-scale model to over 40 50%. This increment indicates a substantial dependency of the model on contextually provided examples. The recurrence of certain logical operator predicate patterns in precedent examples suggests that mere reliance on definitions or prompts is inadequate for mitigating these patterns. Instead, incorporating examples that mirror the anticipated format of modifications is imperative for realizing a significant improvement. Thus, the exploration of methods to enhance the model's logical reasoning capabilities independent of context-based examples constitutes an avenue for future research.

5 Conclusion

In this study, we investigate the capacity of LLMs, with parameters varying from 7B to 200B, to comprehend logical rules. The observed performance disparity between smaller and larger models indicates that size alone does not guarantee a profound understanding of logical constructs. While larger models may show traces of semantic learning, their outputs often lack logical validity when faced with swapped logical predicates. Our findings suggest that while LLMs may improve their logical reasoning performance through in-context learning and methodologies such as COT, these enhancements do not equate to a genuine understanding of logical operations and definitions, nor do they necessarily confer the capability for logical reasoning.

Limitations

Despite employing prompts and in-context examples that ostensibly improve the model's capacity for logical reasoning, the enhancement remains marginal. To date, a method that markedly augments the model's comprehension through in-context learning has not been identified. The prevailing pre-training mechanism focuses on next-token prediction by estimating the subsequent word based on a probability distribution and may not be

ideally suited for logical tasks. These tasks often necessitate the processing of longer-span dependencies and the integration of global information for effective reasoning. Consequently, we believe that devising an alternative pre-training strategy tailored to these requirements presents a promising avenue for future research.

References

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pages 2397–2430. PMLR.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ B. Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, and et al. 2021. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258.

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. 2020a. 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*.

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. 2020b. Language models are few-shot learners. In *NeurIPS*.

Yanda Chen, Ruiqi Zhong, Sheng Zha, George Karypis, and He He. 2022. Meta-learning via language model in-context tuning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 719–730. Association for Computational Linguistics.

Tejas Gokhale, Abhishek Chaudhary, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. 2021. Semantically distributed robust optimization for vision-and-language inference. *arXiv preprint arXiv:2110.07165*.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2019. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*.

Tao Li and Vivek Srikumar. 2019. Augmenting neural networks with first-order logic. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 292–302, Florence, Italy. Association for Computational Linguistics.

Hanmeng Liu, Zhiyang Teng, Leyang Cui, Chaoli Zhang, Qiji Zhou, and Yue Zhang. 2023. Logicot: Logical chain-of-thought instruction tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 2908–2921. Association for Computational Linguistics.

Kevin Lu, Aditya Grover, Pieter Abbeel, and Igor Mordatch. 2021. Pretrained transformers as universal computation engines. *CoRR*, abs/2103.05247.

Bill MacCartney and Christopher D. Manning. 2007. Natural logic for textual inference. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 193–200, Prague. Association for Computational Linguistics.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2021. Recent advances in natural language processing via large pre-trained language models: A survey. *CoRR*, abs/2111.01243.

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? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 11048–11064. Association for Computational Linguistics.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022a. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3470–3487. Association for Computational Linguistics.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022b. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3470–3487. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M. Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *ACL*, pages 15991–16111. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022a. Training language models to follow instructions with human feedback. In *NeurIPS*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022b. Training language models to follow instructions with human feedback. In *NeurIPS*.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *CoRR*, abs/2003.08271.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *CoRR*, abs/2211.09085.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas

Scialom. 2023c. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language model with self generated instructions. *CoRR*, abs/2212.10560.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022a. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

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

Yiquan Wu, Kun Kuang, Yating Zhang, Xiaozhong Liu, Changlong Sun, Jun Xiao, Yueting Zhuang, Luo Si, and Fei Wu. 2020. De-biased court's view generation with causality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 763–780.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. GLM-130B: an open bilingual pre-trained model. In *ICLR*. OpenReview.net.

Xiangji Zeng, Yunliang Li, Yuchen Zhai, and Yin Zhang. 2020. Counterfactual generator: A weakly-supervised method for named entity recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7270–7280.

Qingfu Zhu, Weinan Zhang, Ting Liu, and William Yang Wang. 2020. Counterfactual off-policy training for neural dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3438–3448.