

# The Magic of IF: Investigating Causal Reasoning Abilities in Large Language Models of Code

Xiao Liu<sup>1</sup>, Da Yin<sup>2</sup>, Chen Zhang<sup>1</sup>, Yansong Feng<sup>1,3\*</sup> and Dongyan Zhao<sup>1,4,5</sup>

<sup>1</sup>Wangxuan Institute of Computer Technology, Peking University

<sup>2</sup>Computer Science Department, University of California, Los Angeles

<sup>3</sup>The MOE Key Laboratory of Computational Linguistics, Peking University

<sup>4</sup>Beijing Institute for General Artificial Intelligence

<sup>5</sup>State Key Laboratory of Media Convergence Production Technology and Systems

{lxlisa, zhangch, fengyansong, zhaody}@pku.edu.cn

da.yin@cs.ucla.edu

## Abstract

Causal reasoning, the ability to identify cause-and-effect relationship, is crucial in human thinking. Although large language models (LLMs) succeed in many NLP tasks, it is still challenging for them to conduct complex causal reasoning like abductive reasoning and counterfactual reasoning. Given the fact that programming code may express causal relations more often and explicitly with conditional statements like *if*, we want to explore whether Code-LLMs acquire better causal reasoning abilities. Our experiments show that compared to text-only LLMs, Code-LLMs with code prompts are significantly better in causal reasoning. We further intervene on the prompts from different aspects, and discover that the programming structure is crucial in code prompt design, while Code-LLMs are robust towards format perturbations. Code and data are available at <https://github.com/xxxiao1/magic-if>.

## 1 Introduction

Human beings rely heavily on the capacity for *causal reasoning* (Sloman, 2005; Hagmayer et al., 2007). People understand the observed facts, predict future events, and speculate about what might have happened if things had been different with the help of their causal reasoning skills. For instance, when we go home and find a mess, we probably want to figure out why it happened. If we determine that a bird flew into the house, we might then consider whether the mess could have been avoided if we had closed the window.

Although large language models (LLMs) demonstrate great language understanding and generation abilities, it is still challenging for them to perform complex causal reasoning such as the example above. Powerful LLMs are able to understand single cause-and-effect relations (Brown et al., 2020;

\* Corresponding author.

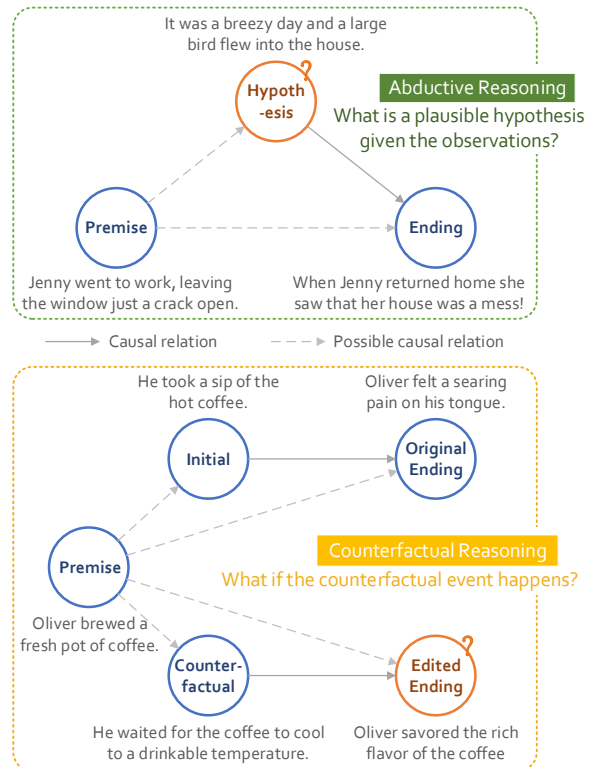


Figure 1: Causal relationships between events in two causal reasoning tasks.

Wang et al., 2021), like *a man losing his balance causes him to fall*. However, when it comes to more complex causal structures involving multiple events and alternative branches (like *close the window* or not), LLMs perform much inferior to humans (Bhagavatula et al., 2019; Qin et al., 2019). In this paper, we consider two challenging causal reasoning tasks: abductive reasoning and counterfactual reasoning. Abductive reasoning requires models to generate a plausible reason for the *ending* while being consistent with the *premise*. Counterfactual reasoning asks what will occur in the *counterfactual branch*. Causal relationships between events in these tasks are shown in Figure 1.

A potential difficulty for LLMs to learn complex

```

# task: generate a plausible explanatory hypothesis given the premise and the ending
def main():
    premise()
    if hypothesis():
        ending()
def premise():
    # The Smiths were having holidays done of the children.
def ending():
    # Ty's face lit up as he ran to the new toy, happily posing for photos.
def hypothesis():
    # Ty was given a new toy.
: model output

# task: generate an ending with three sentences given the premise and the hypothesis
def main():
    premise()
    if hypothesis_1():
        ending_1()
    elif hypothesis_2():
        # minimally revise ending_1
        ending_2()
def premise():
    # Janice was excited to bring cupcakes to her work for her birthday.
def hypothesis_1():
    # She worked all day on making the perfect frosting.
def hypothesis_2():
    # She completely rushed making the frosting.
def ending_1():
    # Each cupcake was truly a work of art.
    # Everyone at her work loved them.
    # Janice was thrilled and happy for the rest of the day.
    # end
def ending_2():
    # The frosting was a complete disaster.
    # Everyone at her work hated them.
    # Janice was sad and embarrassed for the rest of the day.
    # end

```

Figure 2: Example code prompts of abductive reasoning and counterfactual reasoning.

causal structures is that they are rarely expressed explicitly in the text. News articles or narratives may contain multiple events with causal relationships, like an incident and a chain of consequences. However, these events are often written chronologically, and it is hard to extract the causal structure from the text without further annotation. Branches are expressed rarer in text, except for the multi-branching storytelling style (Nisi and Haahr, 2006).

On the other hand, causal relations are exhibited more commonly in code. Conditional statements like `if` direct the computer to execute certain commands, provided a condition is met. This explicitly demonstrates the causal relationship between the *condition block* and the *execution block*. Code can also express branching with `elif` or `switch` statements, and the nesting feature enables code to describe more complex structures<sup>1</sup>.

This motivates us to utilize code models in natural language causal reasoning. Recently, large language models of code (Code-LLMs) are receiving increasing attention (Chen et al., 2021; Xu et al., 2022). They exhibit strong code generation performance, and their structural prediction abilities help complete structural natural language tasks like argument graph generation (Madaan et al., 2022) and event argument extraction (Wang et al., 2022b). Being pre-trained on code with abundant causal expressions, Code-LLMs may also have gained better causal reasoning abilities.

We conduct experiments on the unsupervised abductive reasoning and counterfactual reasoning tasks. To generate task outputs, we design code prompts like Figure 2 to clearly represent the causal structures of the tasks. Results show that Code-

<sup>1</sup>Although causal expressions like *if* are also used in natural languages, representing complex causal structures in text is not as clear and structured as in code.

LLMs with code prompts perform much better than text-only LLMs and previous methods. To better understand why the code prompts are effective, we break down the prompts and analyze the influence of different aspects. We find that Code-LLMs are very sensitive to the *programming structure* (specifically, the conditional statements), while being robust towards format perturbations and programming language changes.

Our main contributions are as follows: 1) We design code prompts to tackle causal reasoning tasks, by leveraging conditional statements in code to represent causal structures. 2) We evaluate Code-LLMs with code prompts on the abductive reasoning and counterfactual reasoning tasks, and exhibit that code models with code prompts are better causal reasoners than text models. 3) We break down the code prompt in detail and find that the programming structure is crucial to the performance.

## 2 Modeling Causal Structure with Code

We convert the input of causal reasoning tasks into the form of code prompt for Code-LLMs to understand better. We expect the prompts to meet two requirements: 1) clearly represent the causal relationships between events, and 2) as most Code-LLMs only support generating at the end, the target output should appear at the end of the prompts. The first requirement is addressed with conditional statements. However, for the second, the target prediction is not always the last part of the conditional statements, e.g., in abductive reasoning we want to predict the hypothesis, which is the condition in the `if` structure. To address this, we uniformly use functions to represent events. As shown in Figure 2, the causal structure is described in the main function. All the event functions are listed afterwards,

	BLEU <sub>4</sub>	ROUGE <sub>L</sub>	CIDEr	BERTScore		BLEU <sub>4</sub>	ROUGE <sub>L</sub>	BERTScore
DELOREAN	1.6	19.1	7.9	41.7	DELOREAN	21.4	40.7	63.4
COLD	1.8	19.5	10.7	42.7	CGMH	41.3	-	73.8
DIFFUSION	7.1	28.3	30.7	-	EDUCAT	44.1	-	74.1
DAVINCI <sub>002</sub>	4.9	27.0	26.6	56.8	DAVINCI <sub>002</sub>	49.0	54.7	73.0
DAVINCI <sub>003</sub>	4.6	25.8	10.7	57.1	DAVINCI <sub>003</sub>	30.6	45.2	69.4
CODEX	<b>13.7</b>	<b>39.6</b>	<b>81.8</b>	<b>64.9</b>	CODEX	<b>66.8</b>	<b>70.0</b>	<b>82.5</b>

(a) Abductive reasoning.

(b) Counterfactual reasoning.

Table 1: Automatic evaluation results on two unsupervised causal reasoning tasks in the zero-shot setting. Numbers are in percentages (%).

	CODEX	Neutral	DAVINCI <sub>002</sub>
<b>Abductive Reasoning</b>			
Coherence with Premise	<b>34%</b>	48.5%	17.5%
Coherence with Ending	<b>32%</b>	42.5%	25.5%
Overall Coherence	<b>40%</b>	38%	22%
<b>Counterfactual Reasoning</b>			
Coherence	<b>36.5%</b>	39.5%	24%
Preservation	<b>47.5%</b>	39.5%	13%

Table 2: Human evaluation of comparing CODEX and DAVINCI<sub>002</sub>.

leaving the target event function at the last.

**Abductive Reasoning.** Abductive reasoning requires models to generate a plausible hypothesis  $H$  given the observations: premise  $P$  and ending  $E$ . The chronological order of these three events is  $P \rightarrow H \rightarrow E$ , and the hypothesis causes the ending to occur.

In Figure 2, we regard the task definition as an instruction and place it as a comment at the beginning of the prompt. The causal structure is represented in the main function like: executing the premise, and if the hypothesis is met, executing the ending<sup>2</sup>. The content of each event is presented as a comment of its function. The hypothesis function is placed at the last, leaving for models to complete. The generation process stops with a line break.

**Counterfactual Reasoning.** Counterfactual reasoning aims to rewrite a story under a counterfactual condition. As in Figure 1, the input consists of four parts: the premise  $P$ , the initial context  $C_1$ , the original ending  $E_1$ , and the counterfactual context  $C_2$ . Models are asked to generate the counterfactual ending  $E_2$  that *minimally* modifies the original ending  $E_1$  and is coherent with the counterfactual context  $C_2$ .

The causal relationships are represented with the `if-elif` structure. The premise  $P$  is executed first, and then if the initial context  $C_1$  is met, the original ending  $E_1$  is executed; otherwise, if the counterfac-

tual context  $C_2$  is met, the counterfactual ending  $E_2$  will be executed. For ease of exposition, we call the context hypothesis as well, being consistent with the former task. The event contents are also written as comments for event functions. We use `#` end to mark the finish of the ending.

### 3 Evaluation

**Datasets.** We experiment on the *ART* dataset (Bhagavatula et al., 2019) for the evaluation of abductive reasoning, and the TimeTravel dataset (Qin et al., 2019) for counterfactual reasoning, with 3,561 and 1,871 test instances, respectively.

**Models.** We experiment with CODEX (Chen et al., 2021), trained on a blend of code and text, as the Code-LLM. The specific version is `code-davinci-002`. We compare with two LLMs: the latest versions of GPT-3 (Brown et al., 2020) `text-davinci-002` and `text-davinci-003` (referred to as DAVINCI<sub>002</sub> and DAVINCI<sub>003</sub>). Both of them originate from CODEX and are tuned with instructions. We follow OpenAI’s default settings in CODEX and DAVINCI decoding, and the text prompts for DAVINCI are in Figure A.1.

We also compare with previous unsupervised methods on these tasks, including DELOREAN (Qin et al., 2020), COLD (Qin et al., 2022), DIFFUSION (Li et al., 2022), CGMH (Miao et al., 2019), and EDUCAT (Chen et al., 2022a)<sup>3</sup>. Appendix A.3

<sup>2</sup>Although not entirely accurate, this approximates the actual underlying causal relationships.

<sup>3</sup>All these methods except DIFFUSION use GPT-2 (Radford et al., 2019) as the base model, and the model size ranges from medium to XL.

	BLEU <sub>4</sub>	ROUGE <sub>L</sub>	CIDEr	BERTScore		BLEU <sub>4</sub>	ROUGE <sub>L</sub>	BERTScore
CODEX <sub>text</sub>	11.7	37.5	78.5	62.5	CODEX <sub>text</sub>	55.1	61.3	77.8
CODEX <sub>code</sub>	13.7	39.6	81.8	64.9	CODEX <sub>code</sub>	66.8	70.0	82.5
CODEX <sub>code</sub> *	<b>16.5</b>	<b>42.0</b>	<b>91.6</b>	<b>66.3</b>	CODEX <sub>code</sub> *	<b>73.3</b>	<b>74.7</b>	<b>85.3</b>
DAVINCI <sub>text</sub>	4.9	27.0	26.6	56.8	DAVINCI <sub>text</sub>	<b>49.0</b>	<b>54.7</b>	<b>73.0</b>
DAVINCI <sub>code</sub>	6.7	31.1	46.2	59.9	DAVINCI <sub>code</sub>	40.4	48.5	70.5
DAVINCI <sub>code</sub> *	<b>9.0</b>	<b>35.0</b>	<b>64.0</b>	<b>62.2</b>	DAVINCI <sub>code</sub> *	43.7	52.0	72.8

(a) Abductive reasoning.

(b) Counterfactual reasoning.

Table 3: Effect of exchanging prompts for CODEX and DAVINCI<sub>002</sub> (%). \* indicates the best code prompt experimented in Section 4.

provides a brief introduction of these methods.

**Automatic Evaluation.** We use the following automatic evaluation metrics: BLEU<sub>4</sub> (Papineni et al., 2002), ROUGE<sub>L</sub> (Lin, 2004), CIDEr (Vedantam et al., 2015) and BERTScore (Zhang et al., 2019) based on BERT-base for abductive reasoning; BLEU<sub>4</sub>, ROUGE<sub>L</sub> and BERTScore for counterfactual reasoning.

Table 1 reports the automatic evaluation results in the zero-shot setting. CODEX significantly outperforms previous methods and DAVINCI on both tasks (with significance level  $\alpha = 0.01$ ), exhibiting strong causal reasoning ability. Although the two DAVINCI models are based on CODEX, their causal reasoning abilities may be weakened during instruction tuning, and this phenomenon is called *alignment tax* (Ouyang et al., 2022). DAVINCI<sub>003</sub> underperforms DAVINCI<sub>002</sub> on most metrics, probably because it tends to generate longer and more discursive outputs, which do not comply with the tasks.

**Human Evaluation.** We conduct pairwise comparison between CODEX and DAVINCI<sub>002</sub> on 100 test examples. Annotators are asked to choose the better output given the task requirements. For abductive reasoning, the outputs are rated from three aspects: coherence with the premise, coherence with the ending, and the overall coherence. For counterfactual reasoning, the outputs are rated from coherence with the context and the extent of preserving the original ending. Each example is rated by at least two annotators, and the average inter-rater reliability is 0.64.

The results are shown in Table 2. CODEX outperforms DAVINCI<sub>002</sub> in all aspects. It better considers the context in generation, and is able to preserve the original content in counterfactual reasoning.

**Contributions of the Model and the Prompt.** We exchange the prompts of code and text models, to measure the contributions of the model and the

prompt. The results are in Table 3. We find that CODEX performs better with the code prompt, as the code prompt clearly describes the causal relation between events. Code prompts benefit the text model DAVINCI<sub>002</sub> on abductive reasoning, but have negative impacts on counterfactual reasoning. A possible reason is that the causal structure in counterfactual reasoning is more complicated, leading to a more complex code which is harder for text models to understand.

## 4 What are Crucial in Code Prompts?

To paint a better picture of the key points in the code prompts, we intervene on the prompts from four aspects and measure the influences of the interventions. The four aspects we select are *information*, *structure*, *format*, and *language*. The former two, the prior information provided and the programming structure of functions, are content-related; the latter two, the code format and programming languages, are form-related. An ideal model should rely on the content and be insensitive to form perturbations. The interventions are described below, with examples in Figure A.2.

**Information.** We study two types of prior information: task instructions and function names. In *No Instruction*, we remove the task instruction from the prompts. In *Function Name Perturbation*, we replace original function names with anonymous functionX. For example, we replace premise() and hypothesis() in Figure 2 with functionA() and functionB(), respectively. It eliminates the information in function names and only allows models to learn the event relations from programming structures.

**Structure.** The first way to intervene in the programming structure is to convert the conditional structures into sequential structures, referred to as *Sequential Structure*. The events are executed sequentially, like premise(), hypothesis(),

		BLEU <sub>4</sub>	ROUGE <sub>L</sub>	CIDEr	BERTScore
	CODEX	13.7	39.6	81.8	64.9
<b>Information</b>	No Instruction	12.1	37.4	73.8	62.9
	Function Name Perturbation	15.1	39.1	77.8	64.6
<b>Structure</b>	Sequential Structure	9.6	36.8	72.0	63.5
	Disruption	7.9	30.3	49.8	58.5
<b>Format</b>	Class	16.0	41.0	87.4	65.8
	Print	13.8	39.4	82.0	65.0
	Return	13.0	40.3	83.4	65.5
<b>Language</b>	Java	<b>16.5</b>	<b>42.0</b>	<b>91.6</b>	<b>66.3</b>
	C	15.5	41.0	88.0	65.6

Table 4: Intervention results on abductive reasoning (%).

ending() in abductive reasoning. In the second way called *Disruption*, we randomly disrupt the positions of the functions in the conditional structure. For instance, if hypothesis(): ending() can be disrupted into if ending(): hypothesis(). We also apply the function name perturbation in disruption to eliminate the impact of function names.

**Format.** We test three formats besides the original one: *Class*, *Print* and *Return*. The first one converts the original code into a class. We define the programming structure in the `__init__` method, and move the event functions into the class. In *Print*, we represent the content of events as a string and print it in the function body, like `def premise(): print("The Smiths ...")`. And in *Return*, the string is the return value of event functions.

**Language.** We also convert the original Python programs into two other languages, *Java* and *C*, to evaluate the influence of programming languages.

**Intervention Results.** We evaluate the influence of interventions on abductive reasoning in Table 4, and the results on counterfactual reasoning are in Table A.2. The absence of prior information causes a small decrease in results. Even if the instruction or function names are not provided, CODEX is able to perform causal reasoning based on conditional statements. Changes in the programming structure have a larger negative impact. Comparing *Function Name Perturbation* and *Disruption*, the alteration of two characters (swap B and C in `functionB` and `functionC`) results in a major drop, showing that the conditional structure that reasonably depicts the relations between events is crucial in CODEX reasoning.

CODEX is quite robust towards format and language changes. Settings like *Class* and *Java* are even better than the original one, revealing that the performance can be further improved with delicate prompt engineering.

## 5 Conclusion

We investigate the causal reasoning ability of Code-LLMs. With code prompts of conditional statements, Code-LLMs achieve great performance in abductive and counterfactual reasoning, outperforming text-only LLMs significantly. Our study on different aspects of code prompts shows that providing a reasonable causal structure in code can help generate plausible outputs, and Code-LLMs are robust towards format perturbations.

## Limitations

**Language** Our experiments are conducted on English, as all Code-LLMs we know are pre-trained on English programming languages. Fundamentally, most popular programming languages are English-based, but international programming languages (which work in multiple languages) like Scratch, or non-English-based programming languages like Qalb also emerge. We look forward to the appearance of Code-LLMs on these programming languages.

**Prompt Engineering** We manually design the prompts without prompt engineering techniques such as prompt search. The searched prompts may outperform the ones we used, but our experiments on interventions show that CODEX is fairly robust towards format perturbations.

**Model** LLMs update quickly. From the time we submitted the paper until now, several new LLMs have been released. We try to compare their performance with ours. We select three new LLMs: CHATGPT, GPT-4 (OpenAI, 2023), and BARD<sup>4</sup>, and feed the text prompts to them. Because we do not have access to some of their APIs, we only experiment on a subset of 100 instances and report

<sup>4</sup>Experiments are done with models updated to May 10, 2023.

	BLEU <sub>4</sub>	ROUGE <sub>L</sub>	CIDEr	BERTScore		BLEU <sub>4</sub>	ROUGE <sub>L</sub>	BERTScore
CODEX	<b>15.0</b>	<b>39.8</b>	<b>82.2</b>	<b>67.8</b>	CODEX	<b>68.4</b>	<b>70.3</b>	<b>84.7</b>
CHATGPT	5.1	26.9	17.5	62.6	CHATGPT	15.3	34.7	70.0
GPT-4	6.3	29.2	27.8	65.1	GPT-4	38.5	55.5	78.6
BARD	5.7	31.5	14.8	66.0	BARD	12.1	22.0	62.1

(a) Abductive reasoning.

(b) Counterfactual reasoning.

Table 5: Automatic evaluation results on a subset of 100 instances in the zero-shot setting. Numbers are in percentages (%).

the results in Table 5. CODEX outperforms all these models in the automatic evaluation, but part of the reason is that these models provide more detailed outputs than the reference. We provide a case study in Appendix A.5.

Since CODEX is no longer available to the public, we provide CODEX generation results in our GitHub repository. We also looked for alternatives and tried two open source Code-LLMs CODEGEN (Nijkamp et al., 2022) (version CodeGen-16B-Mono) and STARCODER (Li et al., 2023) with our code prompts. However, as shown in the case study, their performance is not comparable to CODEX, probably because they are more than ten times smaller in size.

## Ethics Statement

Our work is based on off-the-shelf LLMs. As the results may inherit the underlying bias of LLMs, they cannot be used individually without human supervision. The Codex API was free when the experiments were conducted, and the Davinci APIs cost \$0.02 per thousand tokens. We conduct all the experiments with less than \$100. We recruit annotators for human evaluation from friends and colleagues of authors. All annotators are fairly paid with more than \$10 per hour.

## Acknowledgments

This work is supported in part by NSFC (62161160339). We would like to thank the anonymous reviewers for the helpful discussions and suggestions. For any correspondence, please contact Yansong Feng.

## References

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2019. Abductive commonsense reasoning. In *International Conference on Learning Representations*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Du-Seong Chang and Key-Sun Choi. 2005. Causal relation extraction using cue phrase and lexical pair probabilities. In *International Conference on Natural Language Processing*, pages 61–70. Springer.

Jiangjie Chen, Chun Gan, Sijie Cheng, Hao Zhou, Yanghua Xiao, and Lei Li. 2022a. Unsupervised editing for counterfactual stories. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10473–10481.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. 2022b. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *arXiv preprint arXiv:2211.12588*.

Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. 2021. Excar: Event graph knowledge enhanced explainable causal reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2354–2363.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2022. Pal: Program-aided language models. *arXiv preprint arXiv:2211.10435*.

Andrew S Gordon, Cosmin A Bejan, and Kenji Sagae. 2011. Commonsense causal reasoning using millions of personal stories. In *Twenty-Fifth AAAI Conference on Artificial Intelligence*.

York Haggmayer, Steven A Sloman, David A Lagnado, and Michael R Waldmann. 2007. Causal reasoning through intervention. *Causal learning: Psychology, philosophy, and computation*, pages 86–100.

- Yushi Hu, Chia-Hsuan Lee, Tianbao Xie, Tao Yu, Noah A. Smith, and Mari Ostendorf. 2022. In-context learning for few-shot dialogue state tracking. *ArXiv*, abs/2203.08568.
- Pengfei Li and Kezhi Mao. 2019. Knowledge-oriented convolutional neural network for causal relation extraction from natural language texts. *Expert Systems with Applications*, 115:512–523.
- Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, et al. 2023. Starcoder: may the source be with you! *arXiv preprint arXiv:2305.06161*.
- Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B Hashimoto. 2022. Diffusion-lm improves controllable text generation. *arXiv preprint arXiv:2205.14217*.
- Zhongyang Li, Tongfei Chen, and Benjamin Van Durme. 2019. Learning to rank for plausible plausibility. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4818–4823.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Xiao Liu, Da Yin, Yansong Feng, Yuting Wu, and Dongyan Zhao. 2021. Everything has a cause: Leveraging causal inference in legal text analysis. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1928–1941.
- Aman Madaan, Shuyan Zhou, Uri Alon, Yiming Yang, and Graham Neubig. 2022. Language models of code are few-shot commonsense learners. *arXiv preprint arXiv:2210.07128*.
- Ning Miao, Hao Zhou, Lili Mou, Rui Yan, and Lei Li. 2019. Cgmh: Constrained sentence generation by metropolis-hastings sampling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6834–6842.
- Erik Nijkamp, Bo Pang, Hiroaki Hayashi, Lifu Tu, Huan Wang, Yingbo Zhou, Silvio Savarese, and Caiming Xiong. 2022. Codegen: An open large language model for code with multi-turn program synthesis. *arXiv preprint arXiv:2203.13474*.
- Valentina Nisi and Mads Haahr. 2006. Weird view: interactive multilinear narratives and real-life community stories. *Crossings*, 2:27.
- OpenAI. 2023. Gpt-4 technical report. *arXiv*.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. 2019. Counterfactual story reasoning and generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5043–5053.
- Lianhui Qin, Vered Shwartz, Peter West, Chandra Bhagavatula, Jena D Hwang, Ronan Le Bras, Antoine Bosselut, and Yejin Choi. 2020. Back to the future: Unsupervised backprop-based decoding for counterfactual and abductive commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 794–805.
- Lianhui Qin, Sean Welleck, Daniel Khashabi, and Yejin Choi. 2022. Cold decoding: Energy-based constrained text generation with langevin dynamics. *arXiv preprint arXiv:2202.11705*.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Bryan Rink, Cosmin Adrian Bejan, and Sanda Harabagiu. 2010. Learning textual graph patterns to detect causal event relations. In *Twenty-Third International FLAIRS Conference*.
- Steven Sloman. 2005. *Causal models: How people think about the world and its alternatives*. Oxford University Press.
- Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.
- Jiashuo Wang, Yi Cheng, and Wenjie Li. 2022a. Care: Causality reasoning for empathetic responses by conditional graph generation. *arXiv preprint arXiv:2211.00255*.
- Xingyao Wang, Sha Li, and Heng Ji. 2022b. Code4struct: Code generation for few-shot structured prediction from natural language. *arXiv preprint arXiv:2210.12810*.
- Zirui Wang, Adams Wei Yu, Orhan Firat, and Yuan Cao. 2021. Towards zero-label language learning. *arXiv preprint arXiv:2109.09193*.

- Yuhuai Wu, Albert Q Jiang, Wenda Li, Markus N Rabe, Charles Staats, Mateja Jamnik, and Christian Szegedy. 2022. Autoformalization with large language models. *arXiv preprint arXiv:2205.12615*.
- Frank F Xu, Uri Alon, Graham Neubig, and Vincent Josua Hellendoorn. 2022. A systematic evaluation of large language models of code. In *Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming*, pages 1–10.
- Jiayao Zhang, Hongming Zhang, Weijie Su, and Dan Roth. 2022. Rock: Causal inference principles for reasoning about commonsense causality. In *International Conference on Machine Learning*, pages 26750–26771. PMLR.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.



## A Appendix

### A.1 Related Work

**Causal Reasoning** There is a growing interest in the NLP community to equip models with causal reasoning abilities. Chang and Choi (2005); Gordon et al. (2011) measure causality between words and phrases with statistical methods, Rink et al. (2010); Li and Mao (2019) use explicit semantic cues, and Liu et al. (2021); Zhang et al. (2022) discover causal relations with causal inference methods like propensity score matching. Li et al. (2019) finetune LLMs on causal event corpus, and Du et al. (2021); Wang et al. (2022a) augment LLMs with causal knowledge graphs. Contrast to them, we explore the causal reasoning abilities acquired by Code-LLMs in pre-training.

**Applying Code-LLMs to Natural Language Tasks** With the recent development of Code-LLMs, several works attempt to solve natural language tasks with code models. They mainly focus on two areas: numerical reasoning and structural prediction. Gao et al. (2022); Chen et al. (2022b); Wu et al. (2022) apply Code-LLMs to numerical reasoning. They generate programs with Code-LLMs and feed the programs into an external interpreter to derive the answer. Madaan et al. (2022); Wang et al. (2022b) leverage the text-to-structure translation ability of Code-LLMs to perform structural prediction tasks. They ask models to generate structures in the form of code, and convert the generated code into the task output format. In addition, Hu et al. (2022) takes advantages of Code-LLMs on text-to-SQL generation. Different from them, we leverage the causal reasoning ability of Code-LLMs, and ask them to generate natural language events given the causal structure.

### A.2 Prompts

Figure A.1 demonstrates the prompts of probing DAVINCI. Specifically, the language conversion is made automatically by CODEX with the instruction # python to java/c. Figure A.2 shows the interventions on code prompts for abductive reasoning.

### A.3 Models for Comparison

We compare with previous unsupervised methods on the two tasks, including DELOREAN (Qin et al., 2020), COLD (Qin et al., 2022), and DIFFUSION (Li et al., 2022) on abductive reasoning; and CGMH (Miao et al., 2019), EDUCAT (Chen et al.,

2022a), DELOREAN, and COLD on counterfactual reasoning. Among them, DELOREAN and COLD are constraint-based models. They regard the task requirements as constraints (for example, the generated text should be consistent with the premise, and coherent with the ending in the abductive reasoning task), and iteratively update text representation to meet the constraints. CGMH and EDUCAT are editing-based models targeted for counterfactual reasoning. They start from the original ending and edit it to meet the counterfactual context. DIFFUSION builds a controllable LM based on continuous diffusions to perform control tasks including abductive reasoning.

### A.4 Additional Results

	Min-Edit	BERTScore
DELOREAN	52.9	73.7
COLD	56.8	73.5
CODEX	<b>58.0</b>	<b>79.5</b>

Table A.1: Counterfactual reasoning results in the first-sentence setting (%).

**First-Sentence Setting of Counterfactual Reasoning** Endings in the original counterfactual reasoning data TimeTravel are of three sentences. Due to the computation constraint of COLD (Qin et al., 2022), it is evaluated in a first-sentence setting: only the first sentence of the original ending is used, and models are asked to generate a one-sentence counterfactual ending. We conduct experiments in the first-sentence setting with the metrics used in Qin et al. (2022). As shown in Table A.1, CODEX outperforms previous methods in this setting.

**Intervention on Counterfactual Reasoning** Table A.2 demonstrates the intervention results on counterfactual reasoning. The observations are similar to those in the abductive reasoning task: changes in the programming structure affect CODEX’s performance largely, changes in the information affect less, and CODEX is robust towards format and language changes.

**One-shot Setting** We also conduct experiments in the one-shot setting. Models are shown with one demonstration example in the in-context learning manner, and the example is identical among the models. As shown in Table A.3, both DAVINCI<sub>1002</sub> and CODEX are better than in the

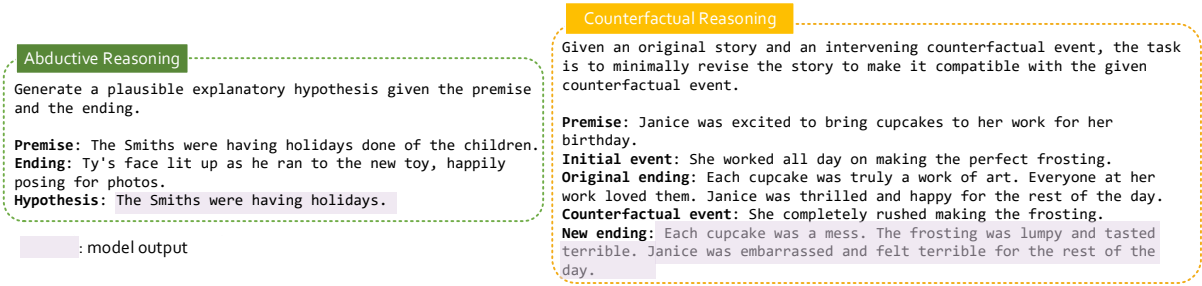


Figure A.1: Example text prompts of abductive reasoning and counterfactual reasoning.

		BLEU <sub>4</sub>	ROUGE <sub>L</sub>	BERTScore
	CODEX	66.8	70.0	82.5
<b>Information</b>	No Instruction	55.4	60.1	77.0
	Function Name Perturbation	65.4	69.0	82.2
<b>Structure</b>	Sequential Structure	43.4	50.2	68.2
	Disruption	16.0	23.5	55.2
<b>Format</b>	Class	63.6	67.4	81.1
	Print	<b>73.3</b>	<b>74.7</b>	<b>85.3</b>
	Return	69.4	70.5	83.0
<b>Language</b>	Java	71.1	73.5	84.5
	C	71.9	74.2	85.0

Table A.2: Intervention results on counterfactual reasoning (%).

zero-shot setting, while CODEX still largely outperforms DAVINCI<sub>002</sub>, showing that the advantage of CODEX is robust across different settings.

### A.5 Case Study

We randomly select some generation examples and demonstrate them in Table A.4. Comparing CODEX and DAVINCI, CODEX generations are more coherent with the context, while DAVINCI sometimes cannot take into account the premise. CODEX also understands the task instruction well and better preserves the original ending in counterfactual reasoning. Generations of more powerful LLMs like CHATGPT and GPT-4 are coherent with the context, but they add much detail and barely keep the original ending. Although open source Code-LLMs like CODEGEN and STARCODER can follow the code prompts and generate sentences in the required format, most of their outputs are inconsistent with the premise and the ending.

	BLEU <sub>4</sub>	ROUGE <sub>L</sub>	CIDEr	BERTScore		BLEU <sub>4</sub>	ROUGE <sub>L</sub>	BERTScore
DAVINCI <sub>002</sub>	8.2	33.5	55.9	61.7	DAVINCI <sub>002</sub>	53.5	58.8	76.0
CODEX	<b>17.9</b>	<b>42.3</b>	<b>91.7</b>	<b>67.1</b>	CODEX	<b>74.3</b>	<b>76.2</b>	<b>86.1</b>

(a) Abductive reasoning.

(b) Counterfactual reasoning.

Table A.3: Evaluation results in the one-shot setting (%).

Abductive Reasoning
<b>Premise:</b> Angie went to a cocktail party hosted by her best friend.
<b>Ending:</b> Angie decided to be quiet about what she overheard.
<b>Hypothesis [CODEX]:</b> Angie overheard her best friend talking about her behind her back.
<b>Hypothesis [DAVINCI<sub>002</sub>]:</b> Angie decided to be quiet about what she overheard because she didn't want to start any drama with her best friend.
<b>Hypothesis [DAVINCI<sub>003</sub>]:</b> Angie overheard a conversation between two people at the party that made her uncomfortable, so she decided to stay quiet and not tell anyone what she heard.
<b>Hypothesis [CHATGPT]:</b> Angie overheard something that she felt was better left unsaid.
<b>Hypothesis [GPT-4]:</b> Angie overheard a sensitive conversation that she felt was not her place to share.
<b>Hypothesis [BARD]:</b> Angie overheard something at the party that she did not want to share with her best friend.
<b>Hypothesis [CODEGEN]:</b> Angie's best friend was a bartender.
<b>Hypothesis [STARCODER]:</b> If judge is cheating on Annie, then judge was correct in trolling an English major.
Counterfactual Reasoning
<b>Premise:</b> My grandpa took my brothers and I fishing in the summer.
<b>Initial Event:</b> We went to a small pond close to his house.
<b>Original Ending:</b> He set up our fishing poles and we began to fish. It was so boring when nobody caught any fish. We were disappointed to go home empty handed that day.
<b>Counterfactual Event:</b> We caught four huge trouts within the first hour.
<b>Counterfactual Ending [CODEX]:</b> He set up our fishing poles and we began to fish. It was so exciting when we caught four huge trouts within the first hour. We were happy to go home with four trouts that day.
<b>Counterfactual Ending [DAVINCI<sub>002</sub>]:</b> We were so excited to have caught four trouts within the first hour! We were proud to show our grandpa our catches.
<b>Counterfactual Ending [DAVINCI<sub>003</sub>]:</b> He set up our fishing poles and we began to fish. To our pleasant surprise, we caught four huge trouts within the first hour. We were delighted to take home our catches that day.
<b>Counterfactual Ending [CHATGPT]:</b> As soon as we cast our lines, we felt a tug and reeled in four huge trouts within the first hour. We couldn't believe our luck and spent the rest of the day catching even more fish. Grandpa was proud of us and we were thrilled to bring home a big haul for dinner. It was a memorable day of fishing that we would never forget.
<b>Counterfactual Ending [GPT-4]:</b> He set up our fishing poles and we began to fish. To our surprise, we caught four huge trouts within the first hour. We were thrilled with our success and couldn't wait to show off our catch when we got home.
<b>Counterfactual Ending [BARD]:</b> We caught four huge trouts within the first hour! We were so excited that we forgot all about being bored. We had a great time fishing with grandpa that day.
<b>Counterfactual Ending [CODEGEN]:</b> He set up our fishing poles and we began to fish. It was so boring when nobody caught any fish. We were disappointed to go home empty handed that day. We caught four huge trouts within the first hour.
<b>Counterfactual Ending [STARCODER]:</b> We were finally happy to catch all of the trouts. And while he is a very kind man, you will never see him again. We will always love our old family in China better than ever before.

Table A.4: Examples of model generations.

### 1. Information: No instruction

```
def main():
    premise()
    if hypothesis():
        ending()

def premise():
    # The Smiths were having holidays done of the children.

def ending():
    # Ty's face lit up as he ran to the new toy, happily
    posing for photos.

def hypothesis():
    #
```

### 2. Information: Function Name Perturbation

```
# task: generate a plausible explanatory hypothesis given the
premise and the ending

def main():
    functionA()
    if functionB():
        functionC()

def functionA():
    # The Smiths were having holidays done of the children.

def functionC():
    # Ty's face lit up as he ran to the new toy, happily
    posing for photos.

def functionB():
    #
```

### 3. Structure: Sequential Structure

```
# task: generate a plausible explanatory hypothesis given the
premise and the ending

def main():
    premise()
    hypothesis()
    ending()

def premise():
    # The Smiths were having holidays done of the children.

def ending():
    # Ty's face lit up as he ran to the new toy, happily
    posing for photos.

def hypothesis():
    #
```

### 4. Structure: Disruption

```
# task: generate a plausible explanatory hypothesis given the
premise and the ending

def main():
    functionA()
    if functionB():
        functionC()

def functionA():
    # The Smiths were having holidays done of the children.

def functionB():
    # Ty's face lit up as he ran to the new toy, happily
    posing for photos.

def functionC():
    #
```

### 5. Format: Class

```
# task: generate a plausible explanatory hypothesis given the
premise and the ending

class Story:
    def __init__(self):
        self.premise()
        if self.hypothesis():
            self.ending()

    def premise(self):
        # The Smiths were having holidays done of the
        children.

    def ending(self):
        # Ty's face lit up as he ran to the new toy, happily
        posing for photos.

    def hypothesis(self):
        #
```

### 6. Format: Print

```
# task: generate a plausible explanatory hypothesis given the
premise and the ending

def main():
    premise()
    if hypothesis():
        ending()

def premise():
    print("The Smiths were having holidays done of the
    children.")

def ending():
    print("Ty's face lit up as he ran to the new toy, happily
    posing for photos.")

def hypothesis():
    print(")
```

### 7. Format: Return

```
# task: generate a plausible explanatory hypothesis given the
premise and the ending

def main():
    premise()
    if hypothesis():
        ending()

def premise():
    return("The Smiths were having holidays done of the
    children.")

def ending():
    return("Ty's face lit up as he ran to the new toy, happily
    posing for photos.")

def hypothesis():
    return(")
```

### 8. Language: Java

```
// task: generate a plausible explanatory hypothesis given the
premise and the ending

public class Story {
    public static void main(String[] args) {
        premise();
        if (hypothesis()) {
            ending();
        }
    }

    public static void premise() {
        // The Smiths were having holidays done of the
        children.
    }

    public static void ending() {
        // Ty's face lit up as he ran to the new toy, happily
        posing for photos.
    }

    public static boolean hypothesis() {
        //
```

### 9. Language: C

```
// task: generate a plausible explanatory hypothesis given the
premise and the ending

int main() {
    premise();
    if (hypothesis()) {
        ending();
    }
}

void premise() {
    // The Smiths were having holidays done of the children.
}

void ending() {
    // Ty's face lit up as he ran to the new toy, happily
    posing for photos.
}

int hypothesis() {
    //
```

Figure A.2: Examples of code prompt interventions in abductive reasoning.

## ACL 2023 Responsible NLP Checklist

---

### A For every submission:

- A1. Did you describe the limitations of your work?  
*Limitation Section*
- A2. Did you discuss any potential risks of your work?  
*Ethics Statement*
- A3. Do the abstract and introduction summarize the paper's main claims?  
*1. Introduction*
- A4. Have you used AI writing assistants when working on this paper?  
*Left blank.*

### B Did you use or create scientific artifacts?

*3. Evaluation*

- B1. Did you cite the creators of artifacts you used?  
*3. Evaluation*
- B2. Did you discuss the license or terms for use and / or distribution of any artifacts?  
*Not applicable. In the supplementary data*
- B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?  
*3. Evaluation*
- B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?  
*Not applicable. The data we use is created and checked by previous work.*
- B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?  
*Limitation Section*
- B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.  
*3. Evaluation*

### C Did you run computational experiments?

*3 & 4*

- C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?  
*Not applicable. The parameters and computational budget are not public available.*

---

*The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.*

- C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?  
*Limitation Section*
- C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?  
*3. Evaluation*
- C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?  
*In the supplementary code*
- D**  **Did you use human annotators (e.g., crowdworkers) or research with human participants?**  
*Appendix A.4*
- D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?  
*The instructions are briefly introduced in Appendix A.4*
- D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?  
*Ethics Statement*
- D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?  
*Appendix A.4*
- D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?  
*Not applicable. Ethics review is not required.*
- D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?  
*Not applicable. Ethics Statement*