

Exploring the Curious Case of Code Prompts

Li Zhang*, Liam Dugan*, Hainiu Xu*, Chris Callison-Burch

University of Pennsylvania

{zharry, ldugan, seacow, ccb}@seas.upenn.edu

Abstract

Recent work has shown that prompting language models with code-like representations of natural language leads to performance improvements on structured reasoning tasks. However, such tasks comprise only a small subset of all natural language tasks. In our work, we seek to answer whether or not code-prompting is the preferred way of interacting with language models *in general*. We compare code and text prompts across three popular GPT models (davinci, code-davinci-002, and text-davinci-002) on a broader selection of tasks (e.g., QA, sentiment, summarization) and find that with few exceptions, code prompts do not consistently outperform text prompts. Furthermore, we show that the style of code prompt has a large effect on performance for some but not all tasks and that fine-tuning on text instructions leads to better relative performance of code prompts.

1 Introduction

Recent work has shown that pre-training language models (LMs) on a mixture of text and program code (e.g., Python or Javascript) makes them more capable of reasoning over natural language (Suzgun et al., 2022). Such program-trained language models (PLMs) significantly outperform text-only LMs on tasks such as math problems and tracking shuffled objects despite such tasks lacking any explicit code formulae (Liang et al., 2022).

Furthermore, prompting such PLMs with *code-like structures* (e.g., Python, JSON, PDDL) instead of text has been shown to lead to performance improvements on structured common sense reasoning (Madaan et al., 2022), event argument extraction (Wang et al., 2022), knowledge graph construction (Bi et al., 2023), story understanding (Dong et al., 2022), and causal reasoning (Zhang et al., 2023).

*Equal contribution.

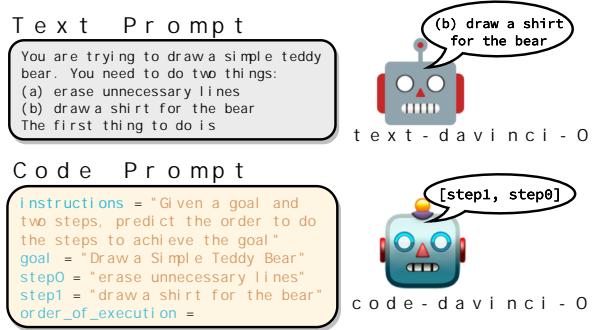


Figure 1: For certain tasks, prompting program-trained language models with code-like representations works better than prompting with text.

Such results naturally lead us to ask whether code-prompting is the preferred way of interacting with PLMs *in general*. While previous work is limited to reasoning tasks, in this work we analyze a broad selection of tasks (e.g., QA, sentiment, summarization) and systematically compare the performance of prompting PLMs with code vs. prompting with text¹. We find that:

- With the exception of some reasoning tasks, code prompts do not outperform text prompts
- The style of code prompt has a large effect on performance for some but not all tasks.
- Fine-tuning on text instructions leads to relative improvements when using code prompts.

2 Experimental Design

Model Selection For our text-based LM we use the original 175 billion parameter davinci model introduced by Brown et al. (2020). For our PLM we use the newer code-davinci-002 model which was explicitly trained on text and code. Neither model underwent any supervised instruction fine-tuning. In addition, we analyze performance on text-davinci-002, which is a vari-

¹The code, prompts, and outputs for our experiments are public at github.com/zharry29/curious_code_prompts

Dataset	Task Category	Num. Eval Examples	Metric	Origin
HellaSwag	Commonsense Reasoning	1000 / 10042	Accuracy	Zellers et al. (2019)
wikiHow Goal-Step	Commonsense Reasoning	1000 / 1073	Accuracy	Zhang et al. (2020)
wikiHow Temporal	Commonsense Reasoning	1000 / 3100	Accuracy	Zhang et al. (2020)
WinoGrande	Commonsense Reasoning	1000 / 1767	Accuracy	Sakaguchi et al. (2021)
OpenPI	Commonsense Reasoning	111 / 111	ROUGE-F1	Tandon et al. (2020)
ANLI	Natural Language Inference	1000 / 3000	Accuracy	Nie et al. (2020)
Yelp	Sentiment Analysis	1000 / 10000	Pearson’s r	Zhang et al. (2015)
IMDb	Sentiment Analysis	1000 / 25000	Accuracy	Maas et al. (2011)
HotpotQA	Question Answering	1000 / 7405	Macro-F1	Yang et al. (2018)
SQuAD	Question Answering	1000 / 11873	Macro-F1	Rajpurkar et al. (2018)
CNN/Daily Mail	Summarization	1000 / 13368	ROUGE-2	Nallapati et al. (2016)
XSUM	Summarization	1000 / 11332	ROUGE-2	Narayan et al. (2018)

Table 1: The 12 evaluation tasks. Macro F1 is based on Rajpurkar et al. (2016). For each task, we randomly sample a fixed set of 1000 examples from its validation or test set for evaluation. For OpenPI we are limited to 111 examples.

ant of `code-davinci-002` trained explicitly on human demonstrations using supervised fine-tuning². We include this model to help us determine whether or not fine-tuning PLMs on text instructions affects their ability to interpret code prompts. All three models were queried through the OpenAI API³ and our experiments cost approximately \$2700 in total (see Appendix F for the full cost breakdown).

Task Selection Following the methodology of Sanh et al. (2022) we select tasks in a top-down fashion by first choosing the categories of interest (e.g. Question Answering, Sentiment Analysis, Summarization) and then selecting datasets from within those categories. We pay special attention to common sense and causal reasoning tasks as PLMs prompted with code have been shown to perform well on such tasks. The resulting 12 tasks are listed in Table 1 and include Commonsense Reasoning, Natural Language Inference, Sentiment Analysis, Question Answering, and Summarization. More details on each task can be found in Appendix A.

Prompt Formulation We collect text prompts for each task using the PromptSource dataset (Bach et al., 2022), a publicly available collection of crowd-sourced prompt templates. For tasks with many prompts, we randomly select one from those provided in the dataset. For a few tasks absent on PromptSource, we write the prompts ourselves.

For our code prompts, we manually write four custom code prompts per task. The code prompt types are as follows, from least to most Pythonic.

(i). **Vanilla (Vanilla):** instructions and inputs are given as variables with generic names;

²<https://platform.openai.com/docs/model-index-for-researchers>

³<https://openai.com/blog/openai-api>

- (ii). **Var Identifier (VI):** instructions and inputs are given as variables with meaningful names;
- (iii). **Var Identifier + Comments (VIC):** instructions and inputs are given as variables with meaningful names along with comments explaining their purpose;
- (iv). **Class + Var Identifier + Comments (CVIC):** instructions and inputs are given as a task-specific class. Functionality is “implemented” as member functions.

Figure 2 shows an example of the different styles of code prompts for the wikiHow temporal ordering task. Note that we attempt to write our code prompts such that we match the wording of the text-based PromptSource prompt as closely as possible.

At inference time, for each test example, we randomly sample in-context examples from the training set and add them to the context window until the maximum context length is reached. This process circumvents the bias caused by static in-context examples. We conduct an ablation study where we vary the random seed and show that this process produces consistent results (see Appendix D).

3 Results

What is the best type of code prompt? We compare performance across the four code prompt types from Section 2 on all 12 tasks using `code-davinci-002` and report our results in Figure 3. We find that no single type of code prompt performs significantly better than the others across all tasks and that the relative difference in performance between code prompts also varies significantly across tasks. For example, on IMDb and SQuAD all code prompts have roughly even performance while for tasks such as wikiHow-Temporal and WinoGrande we see a near 14% accuracy dif-

Text Prompt

```
You are trying to {goal}. You
need to do two things:
(a) {step0}
(b) {step1}
The first thing to do
is {first}
```

Code Prompt (a)

```
input0 = "Given a goal and two steps,
predict the correct order to do the
steps to achieve the goal"
input1 = "{goal}"
step0 = "{step0}"
step1 = "{step1}"
label = [{first}, {second}]
```

Code Prompt var identifier

```
instructions = "Given a goal and two
steps, predict the correct order to do
the steps to achieve the goal"
goal = "{goal}"
step0 = "{step0}"
step1 = "{step1}"
order_of_exec = [{first}, {second}]
```

Code Prompt- var identifier + comments

```
"""Given a goal and two steps, predict the correct
order to do the steps to achieve the goal"""

# The goal that someone is trying to achieve
goal = "{goal}"

# One of the steps that needs to be taken
step0 = "{step0}"

# Another one of the steps that need be taken
step1 = "{step1}"

# The list of correct order of those two steps
order_of_exec = [{first}, {second}]
```

Code Prompt - class + var identifier + comments

```
import order_steps
class Event:
    """Given a goal and two steps, predict the correct
    order to do the steps to achieve the goal"""
    def __init__(self, goal, step0, step1):
        self.goal = goal # The goal someone is trying to accomplish
        self.step0 = step0 # One of the steps that need be taken
        self.step1 = step1 # Another step that need be taken
    def get_order_of_steps(self):
        # Output a list of correct order of the two steps to be taken
        return order_steps(self.goal, self.step0, self.step1)

event = Event(goal="{goal}", step0="{step0}", step1="{step1}")
assert event.get_order_of_steps == [{first}, {second}]
```

Figure 2: An example of the four styles of manually written code prompts used in our analysis (Vanilla, VI, VIC, and CVIC) for the wikiHow temporal ordering task. At test time, variables in braces are replaced with information from the dataset item (as shown in Figure 1). For this task, {goal}, {step0}, {step1} refer to the article title and the steps to order while {first} and {second} refer to the true ordering of the steps.

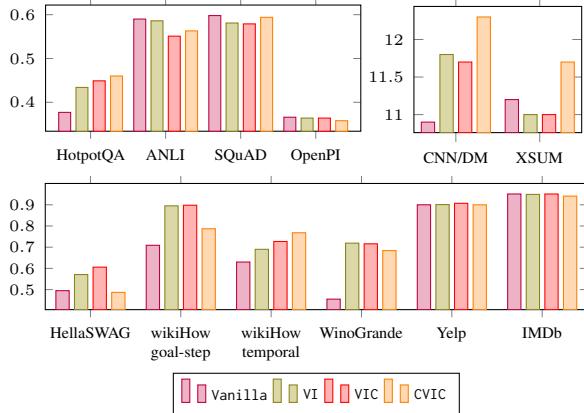


Figure 3: Comparison of code-davinci-002 across the four types of code prompts. Figures are split to allow for different y-axis scales. We see that different prompts do better on different tasks and while some tasks have high variance over prompt types, others do not.

ference between the worst and best prompt.

In Appendix C, we calculate the average rank of each code prompt type relative to each other and find that the “Var Identifier + Comments” (VIC) prompt is the best across all tasks on average (2.25 avg. rank). We thus use this prompt type for our comparison in all future sections.

How many in-context examples should we include in our code prompt? We would like to also investigate how the number of in-context examples in the prompt affects models’ ability to perform the task. We therefore conducted an experiment where we filled the context window of

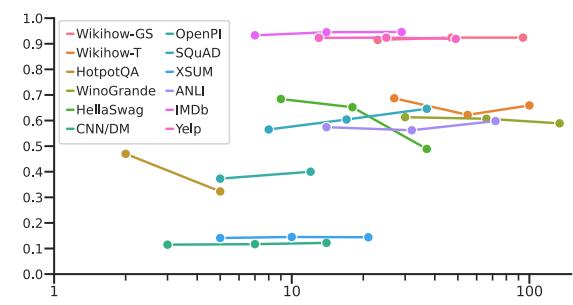


Figure 4: Performance score (y-axis) vs number of in-context examples (x-axis, in log scale) using code prompts (VIC) with code-davinci-002. We see that increasing number of examples does not always increase performance and in some cases makes it worse.

code-davinci-002 with in-context examples up to 2000 tokens, 4000 tokens, and 8000 tokens and plotted the validation accuracy of the model with respect to the number of examples in Figure 4.

Contrary to expectations, we find that the number of in-context examples has little effect on model performance for most tasks and actually has a *negative* effect on some tasks. This is especially interesting given that previous work on in-context learning with text prompts finds roughly monotonic improvement from adding more in-context examples (Liu et al., 2021). While further research is necessary, it seems that code prompts may have different scaling behavior than text prompts when used in in-context learning.

Dataset	Metric	davinci			code-002			text-002		
		+Text	+Code	Δ	+Text	+Code	Δ	+Text	+Code	Δ
Hellaswag	Accuracy	0.321	0.307	-0.014	0.652	0.606	-0.046	0.717	0.773	+0.046
wikiHow goal-step	Accuracy	0.347	0.302	-0.045	0.924	0.898	-0.026	0.919	0.915	-0.004
wikiHow temporal	Accuracy	0.495	0.532	+0.037	0.622	0.727	+0.105	0.688	0.761	+0.073
Yelp	Pearson ρ	0.913	0.896	-0.017	0.924	0.907	-0.017	0.919	0.904	-0.015
IMDb	Accuracy	0.872	0.935	+0.063	0.945	0.951	+0.006	0.940	0.952	+0.012
WinoGrande	Accuracy	0.513	0.500	-0.013	0.607	0.716	+0.109	0.628	0.726	+0.098
ANLI	Accuracy	0.333	0.360	+0.027	0.562	0.551	-0.011	0.504	0.557	+0.053
HotpotQA	Macro-F1	-	-	-	0.470	0.449	-0.021	0.490	0.350	-0.140
SQuAD	Macro-F1	0.482	0.466	-0.016	0.604	0.579	-0.025	0.670	0.656	-0.014
OpenPI	ROUGE-F1	-	-	-	37.33	36.36	-0.970	35.60	31.30	-4.300
CNN/Daily Mail	ROUGE-2	9.28	9.13	-0.150	11.74	11.67	-0.070	13.63	13.55	-0.080
XSUM	ROUGE-2	9.38	6.83	-2.550	14.51	11.03	-3.580	14.48	13.26	-1.220

Table 2: Performance of the three LMs when using code prompts (+Code) vs. using text prompts (+Text). Blank cells indicate tasks for which single test examples could not fit in the context window. Color indicates whether or not code prompts are **better**, **slightly better**, **slightly worse**, or **worse** than text prompts. We see that while code prompts outperform text prompts for certain tasks (such as wikiHow temporal and WinoGrande) text prompts are better on average. We also find that instruction fine-tuning (text-002) allows for better code prompt utilization.

Which is better: code or text prompts? In our main experiment we compare the performance of the three GPT models on code prompts (VIC style) and text prompts across the 12 datasets. Given the results from Figure 4, we fill the context window of all models with in-context examples up to 4000 tokens to serve as a middle ground for comparing code and text prompts. We report the results of our main experiment in Table 2 and see several surprising trends.

First, we find that prompting PLMs with code leads to substantial increases in performance for certain few reasoning tasks but that this trend does not hold across all tasks—or even all reasoning tasks. For example, when using code prompts with code-davinci-002, we see a 10.5% accuracy increase on wikiHow temporal ordering but a 2.6% accuracy decrease on wikiHow goal-step inference despite both being commonsense reasoning tasks and having identical source material.

Second, we find that supervised instruction fine-tuning on natural language demonstrations does not hurt model performance on code. Rather, we observe that code prompts outperform text prompts on *more* tasks when using text-davinci-002 than when using code-davinci-002 despite the fact that text-davinci-002 received no additional fine-tuning on code instructions.

Finally, we find that LMs not explicitly trained on code can also benefit from code prompting on certain reasoning tasks. In particular, code prompts outperform text prompts on davinci for 3 out of our 12 tasks—the same proportion as code-davinci-002. The tasks that benefit from

code prompts also seem to be largely consistent across the three types of models tested, suggesting some underlying trend as to which tasks systematically benefit from structured input.

4 Conclusion

In this work we investigate whether or not there exists a systematic performance difference between prompting PLMs with code or with text. We confirm that there are indeed tasks for which code prompting is significantly more effective than text prompting and that this finding holds across different types of models. However, for most tasks, we find that text prompting is still the best method for eliciting few-shot generalization from PLMs.

Given this result it seems reasonable to attempt to predict which tasks will benefit from code prompts and which tasks will not. However, we show that making such predictions based on simple heuristics such as domain and task category is difficult and that the larger trends remain unclear. Future work should seek to investigate the core mechanism behind what makes code prompting effective for certain tasks.

Finally, concurrent to our work, a new line of research has emerged wherein models generate code and *execute* that code to produce valid output (Chen et al., 2022; Mishra et al., 2022; Gao et al., 2022; Lyu et al., 2023). Future work should consider whether or not the tasks that benefit from executable code prompts and non-executable code prompts have any overlap.

Limitations

One significant limitation to our study is that, as of March 23rd 2023, OpenAI has deprecated access to code-davinci-002⁴, thus rendering our results non-replicable for any team not granted special access to these models by OpenAI. We did not anticipate this deprecation while conducting this work and we believe this raises serious questions about the usage of API-based language models in scholarly work.

Another limitation is that the 12 tasks we selected may not be representative of the broader population of natural language tasks. Had we conducted our experiments on a larger selection of tasks there may have been larger-scale trends that we would have been able to uncover.

The largest and most pressing limitation with our work is that the models we are testing on have closed-source pre-training datasets. Thus, we are unable to verify the extent to which our task datasets have been included in the training or instruction fine-tuning data. Given that the training data for most of the models tested in this work cuts off in late 2021, this is a very strong possibility. Our results should be viewed with this limitation strongly in mind.

Finally, while we experimented with different code prompts, the search space of possible prompts is very large. Thus, it is very likely that there exists some prompt that outperforms our chosen prompts for each task. Drawing conclusions based on a limited sampling of prompts is tenuous and while methods exist for searching the space of all prompts, such techniques lack interpretability and erase any distinction between code and text prompt (Li and Liang, 2021).

Acknowledgements

The paper is dedicated to the late Prof. Dragomir Radev, the first mentor in NLP of the author Li Zhang, for igniting his passion for research and inspiring him to be a better human being.

We thank Shuyan Zhou, Aman Madaan, and Niket Tandon for valuable discussions about this work and we thank Alyssa Hwang for her contributions to the structure, presentation, and narrative of the final paper.

This research is based upon work supported in part by the DARPA KAIROS Program (con-

tract FA8750-19-2-1004), the DARPA LwLL Program (contract FA8750-19-2-0201), the Office of the Director of National Intelligence (ODNI) via the IARPA HIATUS Program (contract 2022-22072200005), the NSF (Award 1928631), and gifts from Roblox and Salesforce. Approved for Public Release, Distribution Unlimited. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DARPA, ODNI, IARPA, NSF, the U.S. Government, or of Roblox or Salesforce. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Stephen H. Bach, Victor Sanh, Zheng-Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesh Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-David, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Alan Fries, Maged S. Al-shaibani, Shanya Sharma, Urmiish Thakker, Khalid Almubarak, Xiangru Tang, Dragomir Radev, Mike Tian-Jian Jiang, and Alexander M. Rush. 2022. *Promptsource: An integrated development environment and repository for natural language prompts*.

Zhen Bi, Jing Chen, Yinuo Jiang, Feiyu Xiong, Wei Guo, Huajun Chen, and Ningyu Zhang. 2023. *Codekgc: Code language model for generative knowledge graph construction*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. *Language models are few-shot learners*. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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

Yijiang River Dong, Lara J. Martin, and Chris Callison-Burch. 2022. *Corpus: Detecting story inconsistency*.

⁴<https://platform.openai.com/docs/model-index-for-researchers>

cies via codex-bootstrapped neurosymbolic reasoning.

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*.

Karl Moritz Hermann, Tomas Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. [Teaching machines to read and comprehend](#). In *NIPS*, pages 1693–1701.

Xiang Lisa Li and Percy Liang. 2021. [Prefix-tuning: Optimizing continuous prompts for generation](#).

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. [What makes good in-context examples for gpt-3?](#)

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. [Faithful chain-of-thought reasoning](#).

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. [Learning word vectors for sentiment analysis](#). In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

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*.

Swaroop Mishra, Matthew Finlayson, Pan Lu, Leonard Tang, Sean Welleck, Chitta Baral, Tanmay Rajpurohit, Oyvind Tafjord, Ashish Sabharwal, Peter Clark, et al. 2022. Lila: A unified benchmark for mathematical reasoning. *arXiv preprint arXiv:2210.17517*.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gülcehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don’t give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial nli: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901.

Long Ouyang, Jeff 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 Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#).

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don’t know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

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

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, et al. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Maarten Sap, Vered Shwartz, Antoine Bosselut, Yejin Choi, and Dan Roth. 2020. Introductory tutorial: Commonsense reasoning for natural language processing. *ACL 2020*, page 27.

Mirac Suzgun, Nathan Scales, Nathanael Schärlí, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.

Niket Tandon, Keisuke Sakaguchi, Bhavana Dalvi, Dheeraj Rajagopal, Peter Clark, Michal Guerquin, Kyle Richardson, and Eduard Hovy. 2020. A dataset for tracking entities in open domain procedural text. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6408–6417.

Xingyao Wang, Sha Li, and Heng Ji. 2022. Code4struct: Code generation for few-shot structured prediction from natural language. *arXiv preprint arXiv:2210.12810*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800.

Li Zhang, Qing Lyu, and Chris Callison-Burch. 2020. Reasoning about goals, steps, and temporal ordering with wikihow. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4630–4639.

Li Zhang, Hainiu Xu, Yue Yang, Shuyan Zhou, Weiqiu You, Manni Arora, and Chris Callison-Burch. 2023. **Causal reasoning of entities and events in procedural texts**. In *Findings of the Association for Computational Linguistics: EACL 2023*, Dubrovnik, Croatia. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. **Character-level convolutional networks for text classification**. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

A Detailed Task Description

Summarization is the task of composing a concise description of a lengthy text. Given a long narrative, the model is tasked with composing a short summary that contains the salient events in the original text.

For our study, we select the *CNN/Daily Mail* (Hermann et al., 2015; Nallapati et al., 2016) and *XSUM* (Narayan et al., 2018) datasets as both are variants on the challenging abstractive summarization task. *XSUM* tasks models with generating extremely concise 1 to 2 sentence summaries of news articles and *CNN/Daily Mail* tasks models with generating reasonably concise but longer abstractive summaries. For both *CNN/Daily Mail* and *XSUM* datasets, we use ROUGE-2 score for evaluation.

Question Answering (QA) is the task of composing answers given a question and an optional context passage. When this context passage is provided the task is referred to as “open-book” QA and when it is not it is referred to as “closed-book” QA. Open-book QA tasks examine language models’ ability to understand and extract information from their context while Closed-book QA tasks evaluate

the amount of knowledge encapsulated in language models during pre-training.

For our study we pick two open-book QA datasets, *SQuADv2* (Rajpurkar et al., 2018) and *HotpotQA* (Yang et al., 2018), which allow us to focus our evaluation on how structured prompts affect models’ ability to comprehend long text input.

For both *SQuADv2* and *HotpotQA*, we evaluate model performance based on the macro-averaged F1 score as proposed in Rajpurkar et al. (2016). This metric measures the average overlap between the prediction and ground truth answer. It is calculated by treating the prediction and ground truth as bags of tokens, and first computing their F1. Then, the maximum F1 score is taken over all of the ground truth answers for a given question, and that score is averaged over all of the questions to get the final result.

Commonsense Reasoning is a machine reasoning task that demands the use of commonsense knowledge which is oftentimes implicitly present in the text (Sap et al., 2020). The customary formulation of commonsense reasoning tasks are *Classification*, where the input is a context, optionally with candidate answers as choices, and the output is a label from a pre-defined label space, and *Question Answering* (QA), where the input is a context followed by a reasoning question and the output is in free-form language.

In this study, we selected four Classification style commonsense reasoning tasks: *wikiHow Temporal* and *wikiHow Goal-Step* (Zhang et al., 2020), *ANLI* (Nie et al., 2020), and *HellaSwag* (Zellers et al., 2019). We also included one Question Answering style task with *OpenPI* (Tandon et al., 2020). In addition, we evaluate our models on *WinoGrande* a comprehensive reasoning benchmark dataset (Sakaguchi et al., 2021).

For *wikiHow Goal-Step*, *wikiHow Temporal*, *HellaSwag*, *WinoGrande*, and *ANLI*, we use classification accuracy as the evaluation metric. To evaluate *OpenPI*, we use F1 score based on the ROUGE metric as described in the original paper (Tandon et al., 2020).

Sentiment Analysis is a task that is concerned with judging emotion and its degree in text. Given a passage, a language model is tasked with classifying the sentiment (positive, negative, neutral) and/or its degree (strongly, weakly, moderately).

The selected datasets, namely *IMDb* (Maas et al.,

	Vanilla	VI	VIC	CVIC
HellaSwag	3	2	1	4
wikiHow Goal-Step	4	2	1	3
wikiHow Temporal	4	3	2	1
Yelp	4	2	1	4
IMDb	1	3	1	4
WinoGrande	4	1	2	3
HotpotQA	4	3	2	1
ANLI	1	2	4	3
OpenPI	1	2	3	4
SQuAD	1	3	4	2
CNN/Daily Mail	4	2	3	1
XSUM	2	4	3	1
<i>Mean</i>	2.75	2.42	2.25	2.58
<i>Standard Deviation</i>	1.36	0.76	1.09	1.26

Table 3: Relative performance rank of the four code prompt types from Section 2 across the 12 tasks. Ranks are calculated based on the results reported in Figure 3. We see that the “Variable Identifier + Comments” (VIC) style prompt performs the best out of all code prompt types on average.

2011) and *Yelp* (Zhang et al., 2015), are both constructed using customer reviews. The IMDb dataset proposes a binary classification problem where the input is a movie review and the label space is $\{negative, positive\}$. Yelp proposes a five-way classification problem where the input is a restaurant review and the label space is the number of stars (out of 5) the customers assigned to the restaurant.

For *IMDb*, we use accuracy as the evaluation metric and for *Yelp*, we use Pearson Correlation between the predicted rating and the ground truth rating as the evaluation metric.

B Hyperparameters

For all our experiments regarding GPT-based models, we use a max token of the maximum of possible output tokens in the ground-truth development set. We use a top p of 1, and no frequency and presence penalty. We use a temperature of 0 for classification and multiple-choice tasks and a temperature of 0.7 for generation tasks.

C Ranking of Code Prompt Styles

In Table 3 we report the rank-based statistics of the four code prompt types from Section 2 on our 12 tasks. Ranks are calculated based on the results reported in Figure 3 of the main paper. The numbers in a row reflect the relative standing of each code prompt on the corresponding task. While we note that all code prompts perform within ± 0.5 ranks of

Dataset	Performance	σ
Hellaswag	0.65, 0.67, 0.69, 0.67, 0.67	± 0.01
wikiHow-GS	0.51, 0.51, 0.51, 0.50, 0.51	± 0.00
wikiHow-T	0.62, 0.65, 0.63, 0.63, 0.62	± 0.01
Yelp	0.92, 0.92, 0.92, 0.92, 0.92	± 0.00
IMDb	0.94, 0.94, 0.94, 0.94, 0.94	± 0.00
WinoGrande	0.62, 0.64, 0.61, 0.62, 0.62	± 0.01
HotpotQA	0.35, 0.33, 0.35, 0.35, 0.35	± 0.01
ANLI	0.59, 0.58, 0.57, 0.60, 0.61	± 0.01
OpenPI	36.3, 38.1, 38.3, 37.7, 39.9	± 1.16
SQuAD	0.60, 0.62, 0.61, 0.60, 0.63	± 0.01
CNN/DM	11.7, 12.0, 12.4, 12.3, 12.0	± 0.25
XSUM	14.5, 14.9, 15.5, 15.2, 15.4	± 0.36

Table 4: Comparison across 5 repeated runs of the *code-davinci-002* model with text prompts using different random seeds for sampling in-context examples. We see minimal standard deviation (σ) between the runs.

each other on average, we see that on average the VIC prompt performs the best across all tasks and the Vanilla prompt performs the worst. Looking to the standard deviation section, we see that the VI prompt performs the most consistently across all tasks and that once again the Vanilla prompt performs the least consistently.

D Ablation Study

To see whether the findings in our Results section could be attributed to variance in the random sampling of in-context training examples per test example, we conduct five repeated runs using *code-davinci-002* with different random seeds each time and calculated the standard deviation across the five runs. We report our results in Table 4 and find that the choice of in-context examples accounts for very little of the observed variance across prompt type and context length. This finding is surprising as previous work has shown that the selection and ordering of in-context examples has a very large effect on the performance of models (Liu et al., 2021). However, it seems that our approach of random sampling in-context examples per test item helps to lessen this inherent variance.

E Evaluation on *text-davinci-003*

While conducting our research into the differences between code and text prompts, OpenAI released the *text-davinci-003* model. This model differs from *text-davinci-002* in that it is trained using Reinforcement Learning with Human Feedback (RLHF) instead of supervised instruction fine-tuning (Ouyang et al., 2022). Out

Task	code-002 (base)	text-002 (+IFT)	text-003 (+RLHF)
HellaSwag	0.652	0.717	0.714
wikiHow GS	0.924	0.919	0.510
wikiHow T	0.622	0.688	0.815
Yelp	0.924	0.919	0.903
IMDb	0.945	0.940	0.938
WinoGrande	0.607	0.628	0.735
ANLI	0.562	0.504	0.549
HotpotQA	0.470	0.490	0.378
SQuAD	0.604	0.670	0.663
OpenPI	37.33	35.60	39.06
CNN/DM	11.74	13.63	12.64
XSUM	14.51	14.48	13.36

Table 5: Performance of the three GPT-3.5 models across our 12 datasets with **text prompts**. (+IFT) indicates the addition of supervised instruction fine-tuning and (+RLHF) indicates the addition of training using Reinforcement Learning from Human Feedback (Ouyang et al., 2022). We see that RLHF does not always improve performance and that for some tasks (HotpotQA and wikiHow Goal-Step) it causes large degradations in performance.

of curiosity, to see the effect of this new training paradigm, we conducted experiments comparing this new `text-davinci-003` model to the other GPT-3.5 models (`text-davinci-002` and `code-davinci-002`). We report the results of our comparison across the 12 evaluation tasks in Table 5.

We see that while `text-davinci-003` outperforms all previous models on *wikiHow Temporal*, *WinoGrande*, and *OpenPI*, it does significantly worse than previous models on *wikiHow Goal-Step* and *HotpotQA*. Such large reductions in performance are to be somewhat expected when using RLHF given the costly nature of collecting human demonstrations. However, the magnitude of the decreases (-50.1% for *wikiHow* and -11.2% for *HotpotQA*) is nonetheless surprising and such results raise important questions about exactly what is being learned when conducting instruction fine-tuning and whether or not this learned information can generalize to tasks not seen during fine-tuning.

F Evaluation Cost

In this section we report the approximate cost of conducting our experiments. In our study we use four OpenAI models, namely `davinci`, `code-davinci-002`, `text-davinci-002` and `text-davinci-003`. While `code-davinci-002` is free to use at the time of this study, we report the approximate cost of running the experiments

Dataset	Num. Examples	Est. Cost
HellaSwag	1000 / 10042	\$240.48
wikiHow Goal-Step	1000 / 1073	\$240.48
wikiHow Temporal	1000 / 3100	\$240.48
WinoGrande	1000 / 1767	\$240.48
OpenPI	111 / 111	\$28.08
ANLI	1000 / 3000	\$240.48
Yelp	1000 / 10000	\$240.48
IMDb	1000 / 25000	\$240.48
HotpotQA	1000 / 7405	\$241.20
SQuAD	1000 / 11873	\$241.08
CNN/Daily Mail	1000 / 13368	\$257.91
XSUM	1000 / 11332	\$246.66
Total Cost		\$2698.29

Table 6: The total estimated cost of running `davinci`, `text-davinci-002` and `text-davinci-003` for 1000 data samples from each dataset (except for OpenPI).

on the other three models⁵ in Table 6. To estimate the cost of an experiment, we calculate the approximate number of tokens necessary for computing one dataset example and then multiplied that by the number of examples in the dataset. For classification tasks, since we fill up the context window to roughly 4000 tokens for every test example, we estimate the number of tokens to be 4000 (3999 tokens for the prompt and 1 token for the label). To estimate cost for generative tasks (OpenPI, HotpotQA, SQuAD, CNN/Daily Mail, and XSUM), we compute the average generation length from our generated samples and assume the in-context examples take up 3500 tokens. While this calculation results in a fairly loose upper bound, we believe this to be a good estimate of the total cost incurred by the project as such overestimates help offset the cost of other miscellaneous API queries done over the course of the project.

⁵The cost of querying `davinci`, `text-davinci-002` and `text-davinci-003` is \$0.02/1,000 tokens at the time of study. See <https://openai.com/pricing> for more details.