# Instruction-following Evaluation through Verbalizer Manipulation

**Shiyang Li**[1] **Jun Yan**[2*] **Hai Wang**[1] **Zheng Tang**[1] **Xiang Ren**[2]
**Vijay Srinivasan**[1] **Hongxia Jin**[1]
[1]Samsung Research America [2]University of Southern California
{shiyang.li, h.wang2, zheng.tang, v.srinivasan, hongxia.jin}@samsung.com
{yanjun,xiangren}@usc.edu

## Abstract

While instruction-tuned models have shown remarkable success in various natural language processing tasks, accurately evaluating their ability to follow instructions remains challenging. Existing benchmarks primarily focus on common instructions that align well with what the model learned during training. However, proficiency in responding to these instructions does not necessarily imply strong ability in instruction following. In this paper, we propose a novel instruction-following evaluation protocol called verbalizer manipulation. It instructs the model to verbalize the task label with words aligning with model priors to different extents, adopting verbalizers from highly aligned (e.g., outputting "positive" for positive sentiment), to minimally aligned (e.g., outputting "negative" for positive sentiment). Verbalizer manipulation can be seamlessly integrated with any classification benchmark to examine the model's reliance on priors and its ability to override them to accurately follow the instructions. We conduct a comprehensive evaluation of four major model families across nine datasets, employing twelve sets of verbalizers for each of them. We observe that the instruction-following abilities of models, across different families and scales, are significantly distinguished by their performance on less natural verbalizers. Even the strongest GPT-4 model struggles to perform better than random guessing on the most challenging verbalizer, emphasizing the need for continued advancements to improve their instruction-following abilities.

## 1 Introduction

Large language models have achieved remarkable success in zero-shot generalization for various natural language processing (NLP) tasks via instruction tuning (Wei et al., 2022a; Ouyang et al., 2022; Sanh et al., 2022; Iyer et al., 2022). One representative model is ChatGPT[1], which has shown promising results in text summarization (Yang et al., 2023), coding (Surameery and Shakor, 2023), healthcare (Sallam, 2023; Zhang et al., 2024), education (Baidoo-Anu and Owusu Ansah, 2023), finance (Dowling and Lucey, 2023) and law (Choi et al., 2023). Existing benchmark datasets (Wang et al., 2019b,a; Cobbe et al., 2021; Hendrycks et al., 2021; Li et al., 2023) primarily focus on common instructions that align well with what models learned during pre-training or instruction-tuning. However, proficiency in responding to these instructions does not necessarily imply strong ability in instruction following as models may rely on memorization of favorable responses rather than genuine generalization due to the vast volume of data they see during training (Tirumala et al., 2022). Nonetheless, instruction following capability plays an important role in task generalization for real-world applications. For example, a user may want models to output answers only when they are certain to reduce hallucinations or control model response length or assign models with specific roles (e.g. tax expert). A natural question arises: *How can we systematically and automatically evaluate instruction-tuned models in terms of instruction-following capability?*

In this paper, we propose to evaluate the instruction-following ability from the aspect of how well models can follow instructions that may not align with their priors and design a novel framework to synthesize them. Specifically, we propose verbalizer manipulation[2] that can be used to construct instructions aligning with model priors to different extents, from *natural*, to *neutral*, to *unnatural*, as shown in Figure 1. In *natural* instructions, we choose multiple verbalizers that align with prior knowledge for each dataset. In *neutral*

* Work was done during Jun's internship at Samsung Research America.

[1] https://chat.openai.com

[2] Following Schick and Schütze (2021), we define a verbalizer as a mapping from golden label names to target ones.



*Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3678–3692
June 16-21, 2024 ©2024 Association for Computational Linguistics

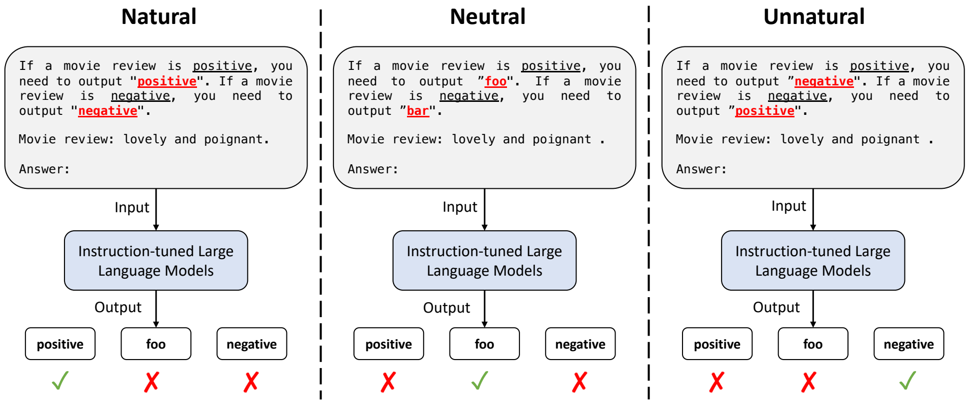


Figure 1: An illustrative example to construct instructions aligning with model priors to different extents, from *natural* (left), to *neutral* (middle), to *unnatural* (right) through verbalizer manipulation for movie review sentiment classification. Levels in terms of aligning with prior knowledge are ranked as *natural* > *neutral* > *unnatural*.

instructions, we select multiple verbalizers that are semantically irrelevant to given tasks. In *unnatural* instructions, verbalizers are flipped from their counterparts in *natural* instructions and contradict with prior knowledge. For example, in a movie review sentiment analysis task, we can use verbalizer “positive|negative”, “1|0”[3], “yes|no” for movie review with positive/negative sentiment to create three sub-evaluation sets for the same dataset in *natural* instructions. The same method can be also used to create multiple sub-evaluation sets for the same dataset in *neutral* and *unnatural* instruction as well. The levels in terms of aligning with prior knowledge of these three instruction groups are ranked as *natural* > *neutral* > *unnatural*. By controlling the level of alignment with prior knowledge and ruling out other factors, we are able to systematically and automatically evaluate the instruction-following capabilities of instruction-tuned models with minimal human efforts.

We evaluate four different model families across various model sizes, namely, Flan-T5 (Wei et al., 2022a), GPT-Series (Ouyang et al., 2022; OpenAI, 2023), Vicuna (Chiang et al., 2023) and OPT-IML (Iyer et al., 2022), on nine benchmark datasets: curated instruction evaluation sets via verbalizer manipulation. First, we compare model performance on *natural*, *neutral* and *unnatural* instructions. We find that larger instruction-tuned models often perform better on both *natural* and *neutral* instructions. Although performance on *neutral* instructions is worse than on *natural* instructions for small models, their performance gap tends to be smaller when model scales and can be (almost) closed for strong OpenAI davinci-003, ChatGPT and GPT-4. On the contrary, the performance of different model families diverge significantly on *unnatural* instructions and there is no clear and consistent trend across model families, showing their significant differences in the ability to follow instructions. Overall, these results indicate that although scaling is an effective way to improve instruction-following ability, it may not be enough when instructions contradict prior knowledge.

Second, we examine verbalizers one by one in both *natural* instructions and their verbalizer-flipped counterparts in *unnatural* instructions. We find that models are not sensitive to verbalizers in *natural* instructions. However, in *unnatural* instructions, performance of the same model diverges significantly and when model further scales, they exhibit scaling-shape (Kaplan et al., 2020) or U-shape (Wei et al., 2022b) or inverse scaling-shape (McKenzie et al., 2022) depending on model family and verbalizers. Even strong ChatGPT and GPT-4 only perform similarly to random guessing when flipped golden label names are used as verlizers in *unnatural* instructions, showing that there still exist fundamental limitations of these models to follow instructions when instructions contradict their prior knowledge.

Finally, we explore whether zero-shot chain of thought (zero-shot-CoT) prompting (Kojima et al., 2022) can improve model performance in *unnatural* instructions that utilize flipped golden label names as verbalizers. We find that although it is helpful when model scales, there still exist large performance gaps compared to corresponding re-

[3]Different from Wei et al. (2023b), we hypothesize that “1”/“0” align more with “positive”/“negative”, respectively, during pre-training or instruction-tuning. This hypothesis is supported by our results on small models in Section 4.2.

sults in *natural* instructions. Only strong ChatGPT and GPT-4 can outperform random guessing while other three model families (Flan-T5, Vicuna, OPT-IML) consistently perform worse than random guessing baseline. In a nutshell, when model scales to larger sizes, they still have difficulty in following instructions contradicting to prior knowledge even though they are allowed to output intermediate reasoning steps. We hope that our work can inspire future research to focus more on instruction-following capability.

## 2 Related Work

**Instruction-tuned Large Language Models.** Large language models have revolutionized the field of NLP and they can perform well in many NLP tasks without any parameter update by only being given several demonstrations in their prompts (Brown et al., 2020). These models are pre-trained with next token prediction or other pre-training objectives, and hence, may not be good at following instructions from humans (Ouyang et al., 2022). To bridge this gap, there have been growing interests in NLP community to train models that can follow human instructions. Mishra et al. (2022); Wei et al. (2022a); Iyer et al. (2022); Sanh et al. (2022) collect standard NLP datasets, write templates for them and transform them into text-to-text format (Raffel et al., 2020) and show that models can generalize to unseen tasks if they are trained on many seen tasks. Chung et al. (2022) studies the scaling effects of instruction-tuning and systematically study what factors are important for unseen test generalizations. Longpre et al. (2023) further finds that task balancing and enrichment techniques are important for instruction-tuning. This line of work mainly focuses on standard NLP tasks and does not reflect how language models are used in many real-world applications (Ouyang et al., 2022). To bridge this gap, Ouyang et al. (2022) collects instructions from humans including their customers to train an instruction-following models like ChatGPT and has achieved remarkable successes. However, collecting large-scaling instruction-following data is time-consuming and expensive, and researchers have been working on utilizing ChatGPT-like models as data generators or human-in-the-loop to generate instruction-following data. Taori et al. (2023) utilizes GPT 3.5 to generate 52K instruction-following data and uses it to train Alpaca. Xu et al. (2024) further explores to evolve instructions from Alpaca (Taori et al., 2023) to generate more complicated instruction-following data to train WizardLM. However, both Alpaca and WizardLM only utilize single-turn data. To alleviate this issue, Xu et al. (2023) utilizes ChatGPT to chat with itself to generate high-quality conversations to train Baize. Chiang et al. (2023) train Vicuna with ShareGPT dialogue data, which are multi-turn conversation dialogues between human users and ChatGPT.

**Language Model Evaluation.** Language models before the era of instruction-tuning (Devlin et al., 2019; Liu et al., 2019; Raffel et al., 2020; Brown et al., 2020) mainly focus on perplexity[4] or results on standard benchmark datasets (Wang et al., 2019b,a), as well as challenging test sets focusing on robustness or generalization (Ribeiro et al., 2020; Yan et al., 2022). However, as models become more and more capable in the era of instruction-tuning, they become harder and harder to evaluate. Hendrycks et al. (2021) collects MMLU dataset including elementary mathematics, US history, computer science, law, etc., to measure knowledge and problem solving capabilities of language models. Liang et al. (2023) instead proposes HELM, a framework to comprehensively evaluate their reasoning, knowledge, robustness, fairness, etc. Chia et al. (2023) introduces InstructEval to comprehensively evaluate instruction-tuned language models. Recently, there have been growing interests in leveraging GPT-4 to evaluate weaker language models (Xu et al., 2024, 2023) although it has been found to be unfair (Wang et al., 2023). However, this line of work mainly focuses on evaluating their general capabilities. Instead, our work focuses on automatic instruction-following evaluation with minimum human efforts. There have been several works sharing a similar focus as ours. Min et al. (2022) finds demonstration with random labels often have comparable performance than using golden labels. We instead focus on instruction-only setting without any demonstration where models are instructed to output specific label names according to their golden labels. Si et al. (2023) measures the inductive biases of large language models via different features, we instead focus on the same task but manipulate different verbalizers to evaluate their instruction-following capability. Webson and Pavlick (2022) finds that models tend to be sensitive to templates and verbalizes for natural

[4] https://paperswithcode.com/sota/language-modelling-on-wikitext-2

language inference (NLI) tasks for small models while our work goes beyond NLI and finds sufficiently large models can perform similarly under different verbalizers. Even when label names are flipped, they can still perform very well under certain tasks, e.g. sentiment classification. The closest work to ours are probably Jang et al. (2023), Wei et al. (2023b) and Wei et al. (2023a). Jang et al. (2023) evaluates instruction-tuned language models with negated prompts while our work utilizes verbalizer manipulations from different groups to control the level of alignment with prior knowledge to follow instructions and have different conclusions. Wei et al. (2023b) finds that large instruction-tuned language models can strengthen their priors and cannot effectively learn to flip labels from given demonstrations. We instead show that if instructions are provided, they do have the ability to flip labels for some tasks due to their strong instruction-following capabilities. Wei et al. (2023a) proposes symbol tuning to force models to learn in-context by changing their label names with symbols to better leverage examples in demonstrations while our work aims to utilize verbalizer manipulation to evaluate the instruction-following capabilities of large language models. Contemporary to our work, Wu et al. (2023) evaluates models' task-level generalizablity by manually designing counterfactual task variants. On the contrary, we propose verbalizer manipulation as a unified evaluation protocol that can be applied to any classification tasks with minimum human efforts.

# 3 Experimental Setup

## 3.1 Datasets

We conduct experiments on nine different binary classification benchmark datasets[5]. Specifically, we utilize **SST-2** ((Socher et al., 2013); Movie review sentiment classification), **FP** ((Malo et al., 2014); Financial phrase sentiment classification), **EMOTION**((Saravia et al., 2018); Twitter message emotion classification), **SNLI** ((Bowman et al., 2015); Stanford natural language inference), **SICK** ((Marelli et al., 2014); Sentence pair entailment analysis), **RTE** ((Dagan et al., 2006); Textual entailment recognition), **QQP** ((Chen et al., 2017); Quora question duplicate detection), **MRPC**((Dolan and Brockett, 2005); Paraphrase identification) and **SUBJ** ((Conneau and Kiela, 2018); Subjective/objective movie description classification). For each dataset and each verbalizer, we use 100 examples to construct our evaluation sets. We defer more details to Appendix B.

## 3.2 Verbalizer Manipulation

For each dataset, we have an instruction template to manipulate its verbalizers. Our templates to manipulate labels for each dataset are deferred to Appendix C. Specifically, for each dataset in *natural* / *neutral* / *unnatural* instructions, we have multiple verbalizers, as shown in Table 1. For example, for SST-2, golden label names are "positive"|"negative" and in *natural* instructions, they will be mapped to "positive"|"negative", "1"|"0", "yes|no". In *neutral* instructions, they will be mapped to "foo"|"bar", "bar"|"foo", "sfo"|"lax", "lax"|"sfo", "lake"|"river","river"|"lake". In *unnatural* instructions, we map them to "negative"|"positive", "0"|"1", "no"|"yes". An illustrative example of three different instruction groups to manipulate verbalizers for SST-2 dataset is shown in Figure 1. For each dataset and each verbalizer (mapping), we generate an evaluation set variant, leading to 2700 examples (9 datasets $\times$ 3 mappings $\times$ 100 examples/dataset) in both *natural* and *unnatural* instructions, and 5400 examples (9 datasets $\times$ 6 mappings $\times$ 100 examples/dataset) in *neutral* instructions.

## 3.3 Instruction-tuned Models

We evaluate state-of-the-art instruction-tuned large language models, namely Flan-T5, GPT-Series, Vicuna and OPT-IML, on datasets in Section 3.1 via verbalizer manipulation in Section 3.2 across various model sizes. For Flan-T5, we evaluate its `small` (80M), `base` (250M), `large` (780M), `xl` (3B) and `xxl` (11B) versions. For GPT-Series, we evaluate `text-ada-001` (ada), `text-babbage-001` (babbage), `text-curie-001` (curie), `text-davinci-003` (davinci), ChatGPT and GPT-4 via official OpenAI APIs[6]. For Vicuna, we evaluate its 7B (`vicuna-7b-1.1`) and 13B (`vicuna-13b-1.1`) versions. For OPT-IML, we utilize its 1.3B (`opt-iml-max-1.3b`) and 30B (`opt-iml-max-30b`) versions (Iyer et al., 2022)). Since our work focuses on evaluating instruction-

[5] Our method can also be used in multi-class classification problems as long as one clarifies how golden labels are manipulated in the instruction. For simplicity, we focus on binary classification tasks in this work.

[6] For experiments in the main body of the paper, we used `gpt-3.5-turbo-0301` for ChatGPT and `gpt-4-0314` for GPT-4, respectively.

| Dataset | Golden label name | Natural | Neutral | Unnatural |
|---|---|---|---|---|
| SST-2 | positive<br>negative | positive, 1, yes<br>negative, 0, no | foo, bar, sfo, lax, lake, river<br>bar, foo, lax, sfo, river, lake | negative, 0, no<br>positive, 1, yes |
| FP | positive<br>negative | positive, 1, yes<br>negative, 0, no | foo, bar, sfo, lax, lake, river<br>bar, foo, lax, sfo, river, lake | negative, 0, no<br>positive, 1, yes |
| EMOTION | joy<br>sadness | joy, 1, yes<br>sadness, 0, no | foo, bar, sfo, lax, lake, river<br>bar, foo, lax, sfo, river, lake | sadness, 0, no<br>joy, 1, yes |
| SNLI | entailment<br>contradiction | entailment, 1, yes<br>contradiction, 0, no | foo, bar, sfo, lax, lake, river<br>bar, foo, lax, sfo, river, lake | contradiction, 0, no<br>entailment, 1, yes |
| SICK | entailment<br>contradiction | entailment, 1, yes<br>contradiction, 0, no | foo, bar, sfo, lax, lake, river<br>bar, foo, lax, sfo, river, lake | contradiction, 0, no<br>entailment, 1, yes |
| RTE | entailment<br>not entailment | entailment, 1, yes<br>not entailment, 0, no | foo, bar, sfo, lax, lake, river<br>bar, foo, lax, sfo, river, lake | not entailment, 0, no<br>entailment, 1, yes |
| QQP | duplicate<br>not duplicate | duplicate, 1, yes<br>not duplicate, 0, no | foo, bar, sfo, lax, lake, river<br>bar, foo, lax, sfo, river, lake | not duplicate, 0, no<br>duplicate, 1, yes |
| MRPC | equivalent<br>not equivalent | equivalent, 1, yes<br>not equivalent, 0, no | foo, bar, sfo, lax, lake, river<br>bar, foo, lax, sfo, river, lake | not equivalent, 0, no<br>equivalent, 1, yes |
| SUBJ | subjective<br>objective | subjective, 1, yes<br>objective, 0, no | foo, bar, sfo, lax, lake, river<br>bar, foo, lax, sfo, river, lake | objective, 0, no<br>subjective, 1, yes |

Table 1: Golden label name mapping for verbalizer manipulation in three different groups.

following capability, we focus on instruction-only setting without any demonstration. We explore the effect of adding few-shot demonstrations in Appendix A. For all experiments, we set temperature as 0 during decoding. We parse predictions from decoded strings and use accuracy (%) as the evaluation metric.

## 4 Experimental results

### 4.1 Results on Instructions with Different Naturalness

We evaluate four model families in Section 3.3 on *natural*, *neutral* and *unnatural* instructions and report results for each instruction group that are averaged over datasets and verbalizers. Results are shown in Figure 2.

**Larger models generally perform better on both *natural* and *neutral* instructions.** For Flan-T5, GPT-series[7] and OPT-IML, we find that model performance improves as they scale to larger sizes on both *natural* and *neutral* instructions. These results are encouraging since it seems that larger models can have better instruction-following capabilities even though instructions do not align with prior knowledge on *neutral* instructions. Further comparing model performance on *natural* and *neutral* instructions, we find that smaller models (model size $\leq$ 30B) perform worse on *neutral* instructions. These performance gaps indicate that smaller models still have difficulty in following instructions. However, their performance gap tends to be smaller when model scales and can be (almost) closed for strong OpenAI davinci, ChatGPT and GPT-4, demonstrating their strong instruction-following capabilities. These results show that simply scaling model size is an effective method to improve model instruction-following capabilities.

**Different model families diverge significantly on *unnatural* instructions.** Although larger models generally perform better on both *natural* and *neutral* instructions, this is not true for *unnatural* instructions. Different model families diverge significantly on *unnatural* instructions and there is no clear and consistent trend across model families. For Flan-T5, results are U-shaped when model scales (Wei et al., 2022b), while for OPT-IML, results follows inverse scaling-shape (McKenzie et al., 2022). In fact, results on these two model families are significantly worse than random guessing (50%). Although Vicuna and GPT-Series follow scaling-shape (Kaplan et al., 2020), their performance still has large gaps compared to results on *natural* instructions, and these gaps seem not to be smaller when they scale. For example, the performance gap for ChatGPT is 11.8% while stronger GPT-4 has 15.7%, making it unclear if further scaling them can bridge this performance gap. This is surprising since these clear and valid instructions can be easily followed by humans but remain difficult for GPT-4, which has shown near human-

[7] Since exact model sizes in GPT-Series are unknown for some of them, we assume that ada $\leq$ babbage $\leq$ curie $\leq$ davinci $\leq$ ChatGPT $\leq$ GPT-4.

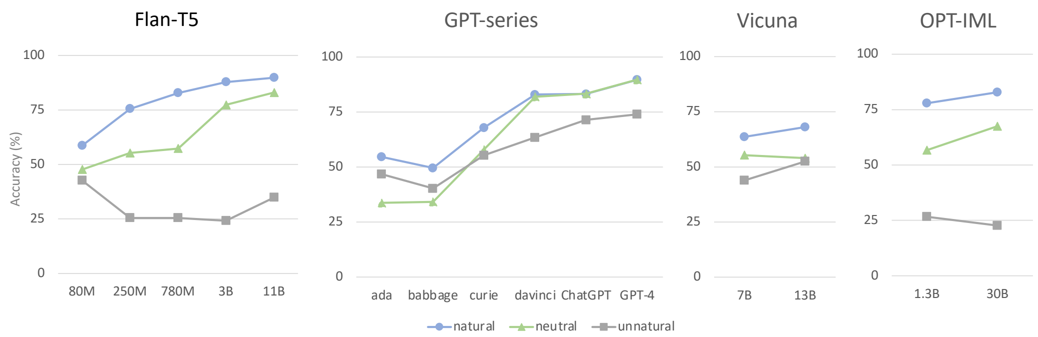


Figure 2: Results comparison under *natural*, *neutral* and *unnatural* instructions across different model families.

level performance on many tasks (Bubeck et al., 2023). Overall, these results indicate that although scaling is an effective way to improve instruction-following, it does not seem to be enough when instructions contradict prior knowledge.

### 4.2 Results of Different Verbalizers in Natural and Unnatural Instructions

Previous discussions focus on average results across different verbalizers for each instruction group. However, it is possible that verbalizers even in the same instruction group align or contradict with prior knowledge differently. For example, it is hard to know if "yes" aligns with prior knowledge more than "1" in SST-2 dataset for *natural* instructions with positive golden labels. Therefore, we further delve into the results of different verbalizers for *natural* instructions and its flipped version in *unnatural* instructions. Average results over nine different datasets are summarized in Figure 3.

**Models perform similarly for different verbalizers in *natural* instructions.** We find that models across four families perform similarly for different verbalizers in *natural* instructions and larger models often perform better than their smaller counterparts. However, we do observe that verbalizers where models perform the best may change in different model sizes and families. For example, for Flan-T5 780M, *natural-golden verbalizers* $>$ *natural-1*$|$*0* $>$ *natural-yes*$|$*no* while for Flan-T5 11B, the order is reversed. In addition, for Vicuna, the best performing verbalizer is *natural-1*$|$*0*, while for OPT-IML, *natural-golden verbalizers* performs better. These results show different models can have different prior knowledge. However, for strong davinci, ChatGPT and GPT-4, their differences are almost not noticeable. This is non-trivial since larger models often have a better understanding about world knowledge and hence store more prior knowledge (Wei et al., 2023b). More consistent results on larger models again show that scaling is an very important factor for instruction-following capability.

**Models diverge significantly for different verbalizers in *unnatural* instructions.** Although previous discussion has shown that models perform similarly for different verbalizers in *natural* instructions, results on their flipped verbalizers in *unnatural* instructions show that they diverge significantly. In Figure 3, we find that verbalizers in *unnatural* group shows very different behaviors when they scale and this behavior also changes in different model families. For example, on *unnatural-no*$|$*yes* and *unnatural-0*$|$*1*, Vicuna achieves better performance when model sizes are larger but degrades on *unnatural-flipped golden verbalizers*. However, for OPT-IML on *unnatural no*$|$*yes*, model performance decreases when it scales to be larger. These results further strengths our finding that different models can have different prior knowledge. On the other hand, it also shows that scaling is not the only factor influencing instruction following although it is important. Further more, we find that for the largest model in each family, performance is ranked as *unnatural 0*$|$*1* $>$ *unnatural no*$|$*yes* $>$ *unnatural-flipped golden verbalizers*. These results show that although they may have different prior knowledge, the difficulty level of overriding their prior knowledge to follow instructions seems consistent. Finally, we find that even the best ChatGPT and GPT-4 only perform similar to random guessing, showing that these models still have fundamental limitations to follow instructions when instructions contradict to their prior knowledge.

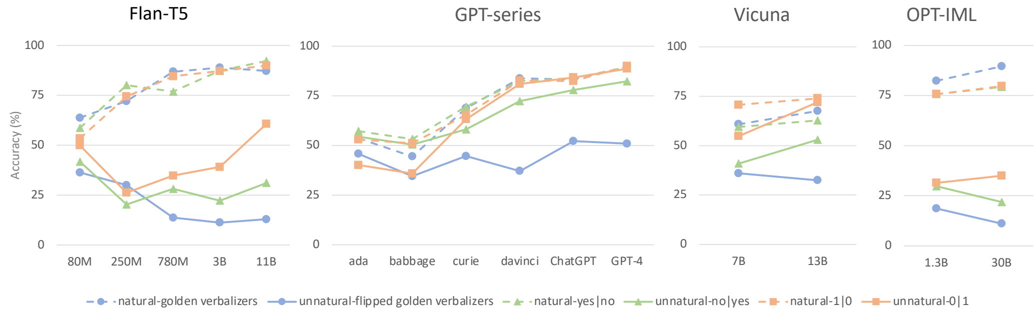


Figure 3: Results of different verbalizers in *natural* and *unnatural* instructions.

### 4.3 Results Comparison between Direct and Zero-Shot Chain-of-Thought Prompting

Previous results have shown that even the best ChatGPT and GPT-4 only perform similar to random guessing on *unnatural-flipped golden verbalizers* and these results are obtained via direct prompting. In this subsection, we further explore if outputting chain-of-thought (CoT) (Wei et al., 2022c) on *unnatural-flipped golden verbalizers* evaluation subset can make models perform better. Therefore, we design another template for each dataset and add *Let's think step by step.* in the prompt following Kojima et al. (2022). We summarize results on *natural-golden verbalizers* and *unnatural-flipped golden verbalizers* via direct prompting, and *unnatural-flipped golden verbalizers* via zero-shot CoT in Figure 4.

For Vicuna and OPT-IML, inverse scaling-curves in *unnatural-direct prompting* become scaling curves in *unnatural-zero shot CoT* prompting. For Flan-T5, results are much more U-shaped in *unnatural-zero shot CoT* compared to those in *unnatural-direct prompting*. Further more, ChatGPT and GPT-4 can significantly outperform random guessing in *unnatural-zero shot CoT* prompting while their counterparts in *unnatural-direct prompting* only have similar performance to random guessing. This is encouraging since it shows that scaling is an effective method to improve instruction-following capabilities along with more advanced prompting techniques. However, they still show large performance gaps compared to results under *natural-direct prompting* setting. For example, Flan-T5 11B, Vicuna 13B and OPT-IML 30B still significantly underperform random guessing. Even strong ChatGPT still has 16.8% accuracy gap to *natural-direct prompting* and for GPT-4, this gap is surprisingly larger and becomes 24.3%. In a nutshell, zero-shot CoT prompting can make models better instruction-followers when instructions contradict prior knowledge, but the models still have a large performance gap with instructions that align with prior knowledge.

### 4.4 Per Dataset Analysis

The previous subsection focuses on average results across different datasets and only ChatGPT and GPT-4 can outperform random guessing on *unnatural* instructions with flipped golden verbalizers in zero shot CoT prompting. In this subsection, we further delve into each dataset by comparing their results using direct prompting with golden verbalizers in *natural* instructions, direct and zero shot CoT prompting with flipped golden verbalizers on *unnatural* instructions. We group results of datasets according to their tasks (e.g., EMOTION, FP and SST-2 are sentiment classification datasets) and results are shown in Figure 5.

**ChatGPT and GPT-4 perform comparably on majority of datasets in both *natural* and *unnatural* instructions.** ChatGPT performs similarly on majority of datasets (6/9, 6/9) compared to GPT-4 ($\leq$ 10% performance gap) on both *natural* and *unnatural* instructions, respectively. GPT-4 outperforms ChatGPT $>$ 10% on RTE and SUBJ in *natural* settings but underperforms it in *unnatural* setting. Another outlier dataset is MRPC, where GPT-4 outperforms ChatGPT 13% and 53% in *natural* and *unnatural* setting, respectively. Overall, these results show that they share more similarity than difference via direct prompting.

**ChatGPT and GPT-4 retain performance on sentiment classification task in *unnatural* direct prompting compared to *natural* counterpart but drop significantly on natural language inference task.** Surprisingly, we find that ChatGPT and GPT-4 can retain their performance on sentiment classification task (FP, EMOTION, SST-2) but drop significantly on natural language inference (NLI)

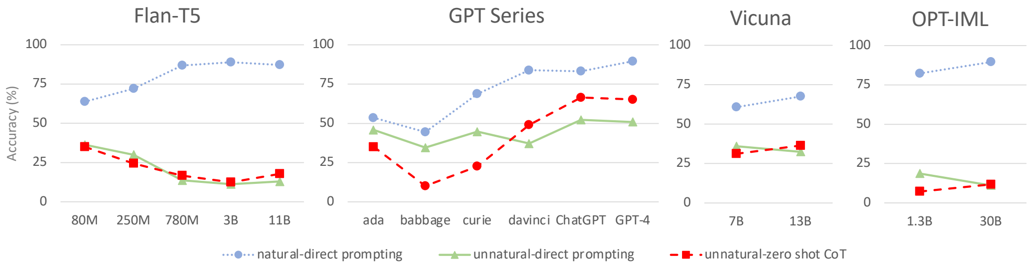


Figure 4: Results comparison between natural-direct prompting with golden verbalizers, unnatural direct prompting and unnatural zero-shot chain-of-thought prompting with flipped golden verbalizers.

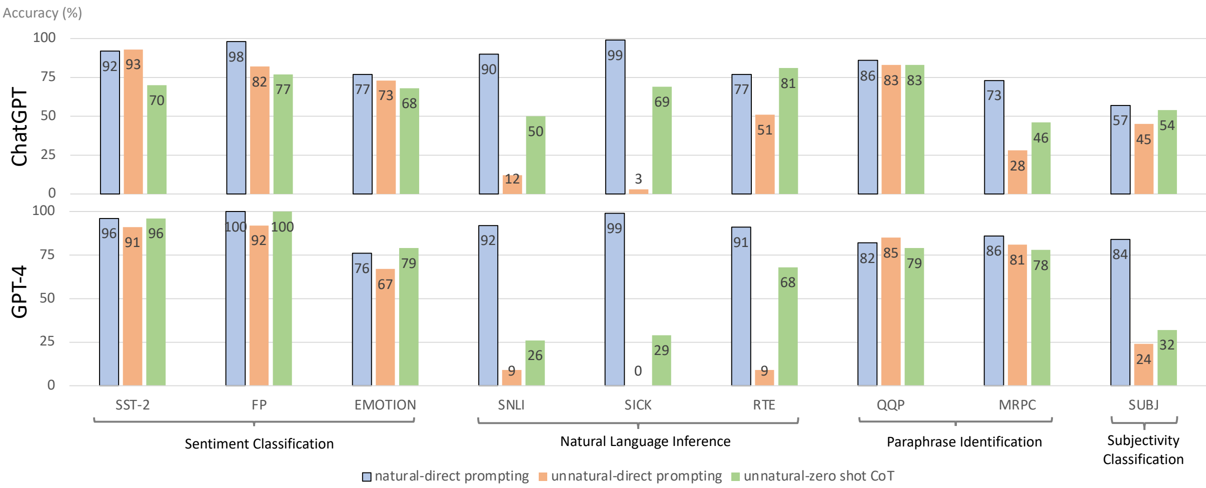


Figure 5: Results comparison between natural-direct prompting with golden verbalizers, unnatural direct and zero-shot chain-of-thought prompting with flipped golden verbalizers for each dataset on ChatGPT and GPT-4.

task (SNLI, SICK, RTE). As an example, on SST-2, ChatGPT outperforms 1% and GPT-4 only decreases 5% with *unnatural* direct prompting while for SICK, ChatGPT and GPT-4 decrease 96% and 99%, respectively. We hypothesize that the discrepancy is because sentiment classification requires less reasoning while NLI requires more, making flipping golden verbalizers much more difficult. One may wonder if they show similar trend on other tasks. For paraphrase identification task, QQP has similar performance after verbalizer flipping for both ChatGPT and GPT-4 while for MRPC, only ChatGPT drops a lot and GPT-4 retains its performance. This result shows that task can be an important factor but not the only one. Models can be sensitive to data distribution.

**ChatGPT and GPT-4 with unnatural-zero shot CoT improve significantly in NLI task but it has much less effect on sentiment classification.** Both ChatGPT and GPT-4 with *unnatural-zero shot CoT* improve significantly in NLI datasets, and ChatGPT can outperform GPT-4 after zero-shot CoT. On the other hand, *unnatural-zero shot CoT* has much less effect on sentiment classification task and even hurts performance across three datasets for ChatGPT. This is probably because unnatural-zero shot CoT is mainly useful for reasoning tasks and sentiment classification requires much less reasoning compared to NLI tasks, making zero shot CoT less useful.

## 5 Conclusion

In this paper, we design a framework to evaluate the instruction-following capabilities of instruction-tuned language models via verbalizer manipulations. We design three instruction-following evaluation sets, namely *natural*, *neural* and *unnatural* instructions, which align with prior knowledge to different extents. We evaluate four different model families on nine datasets across scales. Our results show that although larger instruction-tuned models generally perform better on both *natural* and *neutral* instructions, their performance diverges significantly in *unnatural* instructions. We further examine verbalizers one by one in *unnatural* instructions, and find that the same model family performs significantly different on instructions with different verbalizers, even with more advanced zero shot CoT prompting. These results show there still exist fundamental limitations within state-of-the-art instruction-tuned large language models in following human instructions. We hope that our work can inspire future research to focus more on instruction-following capabilities.

## Limitations

This paper acknowledges certain constraints and identifies avenues for subsequent research endeavors. Firstly, while we aimed for comprehensive assessments across all models, constraints in resources prevented the examination of larger language models like Llama 2 70B, Bard, and Claude. Secondly, our current evaluations are centered on classification tasks; future investigations may explore the application of verbalizer manipulation within generative tasks. Thirdly, the present study is limited to the English language; we intend to broaden our analysis to include multiple languages in future research. Lastly, the OpenAI API is non-deterministic, which may lead to different results for the same input.

## Ethics Statement

For the acquisition of various verbalizers, our approach did not involve the utilization of human annotations; rather, we developed the mapping rules ourselves. Although it is improbable, there is still a chance that these self-devised mapping rules might contain latent biases. During an initial analysis, we detected no instances indicative of such biases. Nevertheless, the possibility of these unintended biases is significant and warrants attention for a more meticulous and in-depth future analysis.

## References

David Baidoo-Anu and Leticia Owusu Ansah. 2023. Education in the era of generative artificial intelligence (ai): Understanding the potential benefits of chatgpt in promoting teaching and learning. *Available at SSRN 4337484*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *ArXiv preprint*, abs/2303.12712.

Zihang Chen, Hongbo Zhang, Xiaoji Zhang, and Leqi Zhao. 2017. Quora question pairs.

Yew Ken Chia, Pengfei Hong, Lidong Bing, and Soujanya Poria. 2023. Instructeval: Towards holistic evaluation of instruction-tuned large language models. *ArXiv preprint*, abs/2306.04757.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Jonathan H Choi, Kristin E Hickman, Amy Monahan, and Daniel Schwarcz. 2023. Chatgpt goes to law school. *Available at SSRN*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *ArXiv preprint*, abs/2210.11416.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *ArXiv preprint*, abs/2110.14168.

Alexis Conneau and Douwe Kiela. 2018. SentEval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In *First PASCAL Machine Learning Challenges Workshop*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Michael Dowling and Brian Lucey. 2023. Chatgpt for (finance) research: The bananarama conjecture. *Finance Research Letters*, 53:103662.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Srinivas Iyer, Xiaojuan Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, Xian Li, Brian O'Horo, Gabriel Pereyra, Jeff Wang, Christopher Dewan, Asli Celikyilmaz, Luke Zettlemoyer, and Veselin Stoyanov. 2022. Opt-iml: Scaling language model instruction meta learning through the lens of generalization. *ArXiv preprint*, abs/2212.12017.

Joel Jang, Seonghyeon Ye, and Minjoon Seo. 2023. Can large language models truly understand prompts? a case study with negated prompts. In *Transfer Learning for Natural Language Processing Workshop*, pages 52–62. PMLR.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *ArXiv preprint*, abs/2001.08361.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacaeval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*. Featured Certification, Expert Certification.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *ArXiv preprint*, abs/1907.11692.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 22631–22648. PMLR.

P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology (JASIST)*.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

Ian McKenzie, Alexander Lyzhov, Alicia Parrish, Ameya Prabhu, Aaron Mueller, Najoung Kim, Sam Bowman, and Ethan Perez. 2022. The inverse scaling prize.

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*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. 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)*, pages 3470–3487, Dublin, Ireland. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *ArXiv preprint*, abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll 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. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Malik Sallam. 2023. Chatgpt utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. In *Healthcare*, volume 11, page 887. MDPI.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Chenglei Si, Dan Friedman, Nitish Joshi, Shi Feng, Danqi Chen, and He He. 2023. Measuring inductive biases of in-context learning with underspecified demonstrations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11289–11310, Toronto, Canada. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Nigar M Shafiq Surameery and Mohammed Y Shakor. 2023. Use chat gpt to solve programming bugs. *International Journal of Information Technology & Computer Engineering (IJITC) ISSN: 2455-5290*, 3(01):17–22.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Kushal Tirumala, Aram H. Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. 2022. Memorization without overfitting: Analyzing the training dynamics of large language models. In *Advances in Neural Information Processing Systems*.

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

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large language models are not fair evaluators. *ArXiv preprint*, abs/2305.17926.

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344, Seattle, United States. Association for Computational Linguistics.

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, Yi Tay, and Quoc V Le. 2022b. Inverse scaling can become u-shaped. *ArXiv preprint*, abs/2211.02011.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022c. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Jerry Wei, Le Hou, Andrew Lampinen, Xiangning Chen, Da Huang, Yi Tay, Xinyun Chen, Yifeng Lu, Denny Zhou, Tengyu Ma, and Quoc Le. 2023a. Symbol tuning improves in-context learning in language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 968–979, Singapore. Association for Computational Linguistics.

Jerry W. Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. 2023b. Larger language models do in-context learning differently. *ArXiv preprint*, abs/2303.03846.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. 2023. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. *ArXiv preprint*, abs/2307.02477.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. 2024. WizardLM: Empowering large pre-trained language models to follow complex instructions. In *The Twelfth International Conference on Learning Representations*.

Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2023. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. *ArXiv preprint*, abs/2304.01196.

Jun Yan, Yang Xiao, Sagnik Mukherjee, Bill Yuchen Lin, Robin Jia, and Xiang Ren. 2022. On the robustness of reading comprehension models to entity renaming. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 508–520, Seattle, United States. Association for Computational Linguistics.

Xianjun Yang, Yan Li, Xinlu Zhang, Haifeng Chen, and Wei Cheng. 2023. Exploring the limits of chatgpt for query or aspect-based text summarization. *ArXiv preprint*, abs/2302.08081.

Xinlu Zhang, Shiyang Li, Xianjun Yang, Chenxin Tian, Yao Qin, and Linda Ruth Petzold. 2024. Enhancing small medical learners with privacy-preserving contextual prompting. In *The Twelfth International Conference on Learning Representations*.

# A Results Comparison between Zero-Shot and Few-Shot Prompting

| Verbalizer | Natural | | Unnatural | |
|---|---|---|---|---|
| Prompting | zero-shot | 4-shot | zero-shot | 4-shot |
| ChatGPT | 81.8 | 85.1$_{(\uparrow \mathbf{3.3})}$ | 52.1 | 56.4$_{(\uparrow 4.3)}$ |
| GPT-4 | 89.4 | 91.7$_{(\uparrow 2.3)}$ | 58.4 | 88.4$_{(\uparrow \mathbf{30.0})}$ |

Table 2: Averaged accuracy comparison between zero-shot prompting and 4-shot prompting with natural golden verbalizers and unnatural flipped golden verbalizers for each dataset on ChatGPT and GPT-4.

Besides plain task instruction, in-context demonstrations also help the model to understand the task. In this section, we explore adding few-shot demonstrations in addition to plain task instruction for improving the model performance in following instructions. Specifically, we evaluate the best-performing ChatGPT and GPT-4 models[8] with *natural-golden verbalizers* and *unnatural-flipped golden verbalizers*. We compare the model performance using zero-shot prompting and 4-shot prompting. The four demonstrations in 4-shot prompting are drawn from corresponding training sets and have balanced label distribution. They demonstrate the expected behaviors of following the specified verbalizers to complete the tasks. We fix the demonstrations for each dataset. In Table 2, we show the averaged accuracy across nine evaluation datasets. We show the per-dataset results in Figure 6. We find that on natural verbalizers, few-shot prompting shows marginal improvement over zero-shot prompting. On unnatural verbalizers, ChatGPT benefits little from few-shot demonstrations on most datasets. However few-shot demonstrations significantly boost the performance of GPT-4, almost closing the gap between natural and unnatural verbalizers. This suggests that GPT-4 has stronger in-context learning abilities than ChatGPT under the unnatural verbalizers. It implies that the in-context learning abilities of the models can also be better distinguished when evaluating with tasks that do not align with models' prior knowledge.

# B Dataset Preprocessing

For each dataset, we utilize their available versions in Huggingface Datasets (Lhoest et al., 2021). Specifically, for FP and EMOTION, we choose their SENTENCES_ALLAGREE and SPLIT subsets, respectively. For FP dataset, as it only has training set, we randomly split it into 80/20 as our in-house training/test set. In addition, for FP, EMOTION, SICK and SNLI datasets, they have multiple classes and we only choose examples whose corresponding labels are shown in Table 1. For SST-2, QQP, RTE and MRPC within GLUE benchmark (Wang et al., 2019b), we randomly sample 100 examples for each dataset from their validation sets while for other five datasets, we randomly sample 100 examples for each dataset from their test sets.

# C Prompt Template

Our instruction templates for verbalizer manipulation in direct prompting setting and zero-shot chain-of-thought prompting is shown in 7 and 8, respectively. Fields with red colors are replaced with verbalizers in Table 1 and fields with blue color will be substituted with input examples in each dataset in text format.

[8]Due to the deprecations of old models by OpenAI, the experiments in this section were performed with `gpt-3.5-turbo-0125` for ChatGPT and `gpt-4-0613` for GPT-4, respectively.

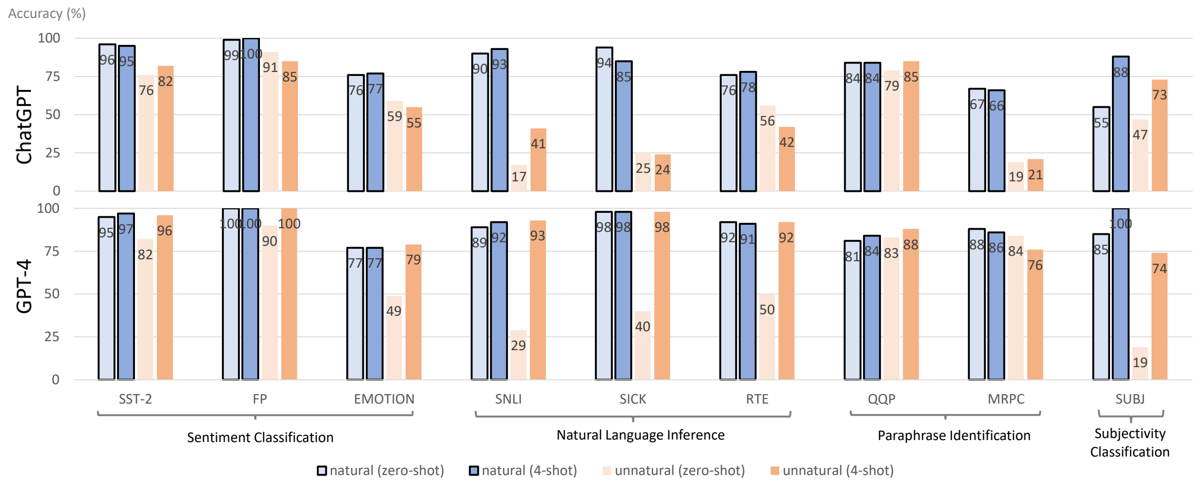


Figure 6: Per-dataset accuracy comparison between zero-shot prompting and 4-shot prompting with natural golden verbalizers and unnatural flipped golden verbalizers for each dataset on ChatGPT and GPT-4.

```
You are a helpful assistant judging the sentiment of a movie review. If the movie review is positive, you need to output "{positive}". If the movie review is negative, you need to output "{negative}". You are only allowed to output "{positive}" or "{negative}".\n\nMovie review: {input}\n\nAnswer:
```

(a) SST-2

```
You are a helpful assistant judging the sentiment of a financial phrase. If the financial phrase is positive, you need to output "{positive}". If the financial phrase is negative, you need to output "{negative}". You are only allowed to output "{positive}" or "{negative}".\n\nFinancial phrase: {input}\n\nAnswer:
```

(b) FP

```
You are a helpful assistant judging the emotion of a Twitter message. If the emotion of a Twitter message is joy, you need to output "{joy}". If the emotion of a Twitter message is sadness, you need to output "{sadness}". You are only allowed to output "{joy}" or "{sadness}".\n\nTwitter message: {input}\n\nAnswer:
```

(c) EMOTION

```
You are a helpful assistant judging if sentence 1 entails sentence 2. If sentence 1 entails sentence 2, you need to output "{entailment}". If sentence 1 contradicts sentence 2, you need to output "{contradiction}". You are only allowed to output "{entailment}" or "{contradiction}".\n\nSentence 1: {sentence_1}\nSentence 2: {sentence_2}\n\nAnswer:
```

(d) SNLI

```
You are a helpful assistant judging if sentence 1 entails sentence 2. If sentence 1 entails sentence 2, you need to output "{entailment}". If sentence 1 contradicts sentence 2, you need to output "{contradiction}". You are only allowed to output "{entailment}" or "{contradiction}".\n\nSentence 1: {sentence_1}\nSentence 2: {sentence_2}\n\nAnswer:
```

(e) SICK

```
You are a helpful assistant judging if sentence 1 entails sentence 2. If sentence 1 entails sentence 2, you need to output "{entailment}". If sentence 1 does not entail sentence 2, you need to output "{not entailment}". You are only allowed to output "{entailment}" or "{not entailment}".\n\nSentence 1: {sentence_1}\nSentence 2: {sentence_2}\n\nAnswer:
```

(f) RTE

```
You are a helpful assistant judging if two given questions from Quora are semantically equivalent. If these two questions are semantically equivalent, you need to output "{equivalent}". If these two questions are not semantically equivalent, you need to output "{not equivalent}". You are only allowed to output "{equivalent}" or "{not equivalent}".\n\nQuestion 1: {question_1}\nQuestion 2: {question_2}\n\nAnswer:
```

(g) QQP

```
You are a helpful assistant judging if two sentences from online news sources are semantically equivalent. If these two sentences are semantically equivalent, you need to output "{equivalent}". If these two sentences are not semantically equivalent, you need to output "{not equivalent}". You are only allowed to output "{equivalent}" or "{not equivalent}".\n\nSentence 1: {question_1}\nSentence 2: {question_2}\n\nAnswer:
```

(h) MRPC

```
You are a helpful assistant judging if the given input is a subjective or objective description of a movie. If the movie description is subjective, you need to output "{subjective}". If the movie description is objective, you need to output "{objective}". You are only allowed to output "{subjective}" or "{objective}".\n\nMovie description: {input}\n\nAnswer:
```

(i) SUBJ

Figure 7: Instruction templates for verbalizer manipulation in direct prompting.

You are a helpful assistant judging the sentiment of a movie review. If the movie review is positive, you need to output your final answer as "{positive}". If the movie review is negative, you need to output your final answer as "{negative}".\n\nMovie review: {input}\n\nAnswer: Let's think step by step.

(a) SST-2

You are a helpful assistant judging the sentiment of a financial phrase. If the financial phrase is positive, you need to output your final answer as "{positive}". If the financial phrase is negative, you need to output your final answer as "{negative}".\n\nFinancial phrase: {input}\n\nAnswer: Let's think step by step.

(b) FP

You are a helpful assistant judging the emotion of a Twitter message. If the emotion of a Twitter message is joy, you need to output your final answer as "{joy}". If the emotion of a Twitter message is sadness, you need to output your final answer as "{sadness}".\n\nTwitter message: {input}\n\nAnswer: Let's think step by step.

(c) EMOTION

You are a helpful assistant judging if sentence 1 entails sentence 2. If sentence 1 entails sentence 2, you need to output your final answer as "{entailment}". If sentence 1 contradicts sentence 2, you need to output your final answer as "{contradiction}".\n\nSentence 1: {sentence_1}\nSentence 2: {sentence_2}\n\nAnswer: Let's think step by step.

(d) SNLI

You are a helpful assistant judging if sentence 1 entails sentence 2. If sentence 1 entails sentence 2, you need to output your final answer as "{entailment}". If sentence 1 contradicts sentence 2, you need to output your final answer as "{contradiction}".\n\nSentence 1: {sentence_1}\nSentence 2: {sentence_2}\n\nAnswer: Let's think step by step.

(e) SICK

You are a helpful assistant judging if sentence 1 entails sentence 2. If sentence 1 entails sentence 2, you need to output your final answer as "{entailment}". If sentence 1 does not entail sentence 2, you need to output your final answer as "{not entailment}".\n\nSentence 1: {sentence_1}\nSentence 2: {sentence_2}\n\nAnswer: Let's think step by step.

(f) RTE

You are a helpful assistant judging if two given questions from Quora are semantically equivalent. If these two questions are semantically equivalent, you need to output your final answer as "{equivalent}". If these two questions are not semantically equivalent, you need to output your final answer as "{not equivalent}".\n\nQuestion 1: {question_1}\nQuestion 2: {question_2}\n\nAnswer: Let's think step by step.

(g) QQP

You are a helpful assistant judging if two sentences from online news sources are semantically equivalent. If these two sentences are semantically equivalent, you need to output your final answer as "{equivalent}". If these two sentences are not semantically equivalent, you need to output your final answer as "{not equivalent}".\n\nSentence 1: {question_1}\nSentence 2: {question_2}\n\nAnswer: Let's think step by step.

(h) MRPC

You are a helpful assistant judging if the given input is a subjective or objective description of a movie. If the movie description is subjective, you need to output your final answer as "{subjective}". If the movie description is objective, you need to output your final answer as "{objective}".\n\nMovie description: {input}\n\nAnswer: Let's think step by step.

(i) SUBJ

Figure 8: Instruction templates for verbalizer manipulation in zero-shot chain-of-thought prompting.