Reproducibility Report of "Context-faithful Prompting for Large Language Models"

*Akashdeep Bhattacherjee, Arya GJ, Murray Kang, Alan Li
Paul G. Allen School of Computer Science and Engineering
{akashb03, agk8, haoqik, lihaoxin}@uw.edu

Reproducibility Summary

Scope of Reproducibility

Zhang, Roller, Goyal, Artetxe, Chen, Chen, Dewan, Diab, Li, Lin, Mihaylov, Ott, Shleifer, Shuster, Simig, Koura, Sridhar, Wang, and Zettlemoyer (2022) focus on improving LLM's context faithfulness by carefully designed prompting template and demonstration. In particular, they claimed that the opinion-based prompting template and counterfactual demonstration are two methods that can effectively improve LLMs' context faithfulness on sub-problems of entity-based knowledge conflict and prediction with abstention. In this work, we try to reproduce Zhou et al. (2023)'s work by conducting same experiments on a wide range of open-source modern LLMs.

Methodology

We try to reproduce the results reported in the original work in an identical environment with the codebase released by Zhou et al. (2023), and further analyze the effectiveness of its proposed strategies on a wider range of open-source models by adding onto the released codebase. We use nlpg hosts, Hyak hosts, and OpenAI API for all experiments mentioned in this work. No additional training or finetuning is needed throughout the project.

Results

Our results show that the strategies proposed by Zhou et al. (2023) are not consistently effective across different families or sizes of LLMs, and the conclusion drawn by the original authors might be too ambitious. In our experiments, the prompting templates are only effective in zero-shot settings, and counterfactual demonstrations do not show obvious improvement on context faithfulness over gold-label demonstrations.

What was Easy

- Preliminary research. The chosen paper is well written. Related works are open and accessible with plenty of well organized open-source tools and packages.
- Data preprocessing, including filtering, sampling, tokenizer, etc.

What was Difficult

- Unable to reproduce experiments from the original paper due to high expense involved in accessing the models and datasets used and lack of compute availability.
- Explaining the randomness that appeared in our results and trying to understand the lack of robustness in our models' responses.

Communication with Original Authors

We contacted Wenxuan Zhou, the main author and the maintainer of the paper's Github, a few times to ask a few questions and get access to the dataset filtering code.

*Alphabetic order.

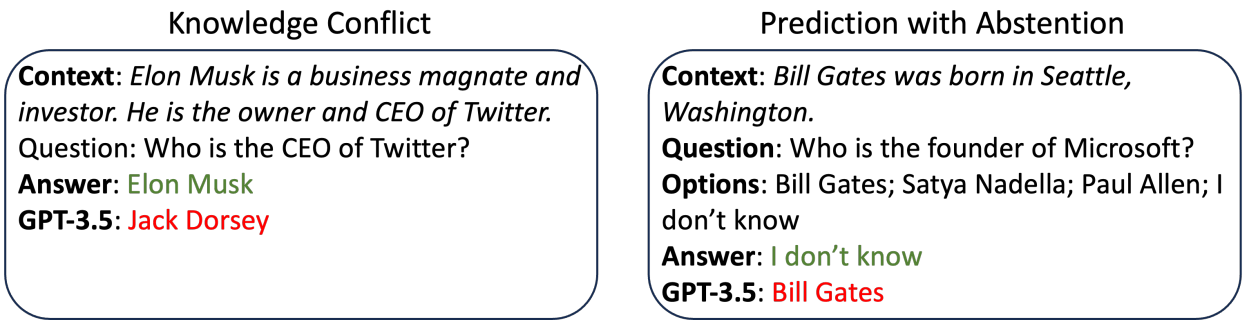

Figure 1: Examples of knowledge conflict and prediction with abstention. LLMs may ignore the provided context and make unfaithful predictions based on their parametric knowledge before Q4 2021.

1 Introduction

Modern large language models (LLMs) have proven to be highly capable of knowledge-driven NLP tasks. However, LLMs may overly rely on parametric knowledge and ignore context cues in prediction. Similar traits have been observed in in-context learning (ICL) settings. Although LLMs are known to be capable of performing a new task via inference alone by conditioning on a few input-label pairs (demonstrations), Min et al. (2022) show that the quality of demonstrations is not essential to inference performance, i.e., randomly replacing labels in demonstrations barely hurts performance on a range of NLP classification and QA tasks in few-shot learning settings. In both context-specific application scenarios, LLMs can easily parrot answers from pretraining without genuinely inducing relations or concepts described in context.

Zhou et al. (2023) propose prompting templates and demonstration strategies that improve LLMs' context faithfulness. In particular, they focus on knowledge conflicts and prediction with abstention subproblems and show the proposed opinion-based template and counterfactual demonstration enhance context faithfulness in both subproblems. In this work, we try to reproduce Zhou et al. (2023)'s work and verify the effectiveness of their methods on a wider range of open-source LLMs across different families and scales. As a stretch goal, we also extend these methods to more datasets used by Min et al. (2022) and see if they can induce consistent improvement under ICL settings. However, according to our experiments, we observe that the proposed templates and demonstration strategies do not show consistent improvement across different LLMs. The counterfactual demonstration provides no obvious improvement in our tests, and prompting templates are only effective in zero-shot setting.

2 Scope of Reproducibility

Zhou et al. (2023) assess and enhance LLMs' context faithfulness on two sub-problems: namely entity-based knowledge conflict (Longpre et al., 2021) and prediction with abstention (Rajpurkar et al., 2018).

Knowledge Conflicts. Given context contains facts different from the pretraining data, LLMs need to answer the question using facts described locally in the context instead of memorized facts. An example from the original paper that shows an LLM fails to predict according to facts described in the context is shown in Figure 1, where `text-davinci-003` identifies Jack Dorsey instead of Elon Musk as the CEO of Twitter, based on its pre-trained data before Q4 2021.

Prediction with abstention. In the case where given context does not provide enough information to answer the question, LLMs should abstain from making predictions and notify users instead of returning incorrect predictions. As the example shows in Figure 1, the context does not provide enough information to infer the answer and therefore the LLM should abstain from answering instead of retrieving parametric knowledge from pretrained data.

In-context learning (stretch goal). Our stretch goal in this project is to conduct ICL experiments, guided by two central hypotheses: 1) the proposed prompting templates can enhance language models' ICL performance with original demonstration, and 2) there will be obvious performance gap between using original and counterfactual demonstration with proposed templates.

2.1 Addressed Claims from the Original Paper

1. "[W]e find that adding counterfactual demonstrations to prompts improves faithfulness in the aspect of knowledge conflict, while using the original (factual) demonstrations leads to limited or negative effects."
2. "We demonstrate that LLMs' faithfulness can be significantly improved using carefully designed prompting strategies. ... [W]e find that reformulating the context and questions to opinion-based question-answering problems (Gupta et al., 2019; Bjerva et al., 2020), where the context is expressed in terms of a narrator's statement, and the question asks about this narrator's opinion, delivers the most gains."

   Besides opinion-based prompting, the original work also proposed attributed prompts and instruction-based prompts and compared model performance with different prompting strategies.

3 Methodology

Zhou, Zhang, Poon, and Chen (2023) focus on context-specific NLP tasks. The input of these tasks can be formulated as (c, q) for free-form generation tasks, where c is the context and q is the question, or (c, q, o) for tasks with close decision spaces (e.g., multi-choice tasks), where o is the set of decisions/choices. The desired output is either a free-form text or a choice. We solve these tasks by prompting LLMs and Zhou et al. (2023) seek to improve the faithfulness of LLMs in two ways, namely prompt engineering and demonstrations, and study ways of designing prompting templates and demonstrations that are dedicated to improving the faithfulness of LLMs. Note that for knowledge conflict task, test examples are given as free-form QA, while for prediction with abstention, we have five choices for each test example, including the "I don't know" option. As the stretch goal, ICL tasks are all binary classification tasks.

Zhou et al. (2023) propose three different templates, as presented below, that have shown to be effective in improving LLMs' context faithfulness in the original paper. {} serves as a placeholder that is filled with specific inputs during prompting.

(1). Given an input of (c, q, o), we have **base** prompting template as a baseline:

Base prompt

$\{c\}$ Q: $\{q\}$? Options: $\{o\}$ A:

(2). Opinion-base prompt, which transforms the context to a narrator's statement and the question to enquire about the narrator's opinion in this statement:

Opinion-based prompt

Bob said, "$\{c\}$" Q: $\{q\}$ in Bob's opinion? Options: $\{o\}$ A:

(3). Instruction-based prompt, which is also giving out natural language instruction, but now we leverage automatic prompt engineering (APE; Zhou et al. (2023) 2022) to generate the prompts:

Instruction-based prompt

Instruction: {*Instr*} $\{c\}$ Q: $\{q\}$? Options: $\{o\}$ A:

(4). Attributed prompt, which instructs LLMs to read the context and answer the question accordingly:

Attributed prompt

$\{c\}$ Q: $\{q\}$ based on the given text? Options: $\{o\}$ A:

Zhou et al. (2023) use the base prompt as our baseline, and compare it against the proposed prompting templates, including attributed prompt (ATTR), instruction-based prompt (INSTR), opinion-based prompt (OPIN), and the combination of opinion-based prompt and instruction-based prompt (OPIN + INSTR).

Zhou et al. (2023) evaluate the effectiveness of these templates in both zero-shot and few-shot settings (with demonstrations). We follow the same procedure and filter out instances that exceed the maximum input length for each model separately. For demonstrations, Zhou et al. (2023) test original and counterfactual demonstration in knowledge conflict subproblem. The counterfactual demonstration manipulates original examples such that the facts in the context are substituted with false ones.

3.1 Model Descriptions

All experiments in the original paper are conducted on models from OpenAI. The authors use different sizes of InstructGPTs (Ouyang et al., 2022) on both knowledge conflict and prediction with abstention: 0.3B, 1.3B, 6.7B, and 175B, which are `text-ada-001`, `text-babbage-001`, `text-curie-001`, and `text-davinci-003` respectively.

In our case, we test on `text-davinci-003` for prediction with abstention and ICL settings as sanity check. Due to limited budget, we cannot reproduce all experiments with OpenAI API. Our main goal is to analyze whether the proposed strategies from the original paper are effective for a wider range of open-source LLMs. Therefore we conduct experiments on different sizes of FLAN-t5 (Chung et al., 2022), opt (Zhang et al., 2022), and opt-iml(-max) (Iyer et al., 2023), as well as LLaMA-7B (Touvron et al., 2023), and Alpaca-7B (Taori et al., 2023). No additional training or finetuning is needed for reproduction.

3.2 Datasets

For knowledge conflict, the original paper evaluates using counterfactual datasets that contain incorrect facts, which conflict with what LLMs have already memorized. Natural questions (NQ) (Kwiatkowski et al., 2019) and Re-TACRED (Stoica et al., 2021) are used in this setting. In our case, however, we only use natural questions for MRC since the Re-TACRED is not freely accessible to the public. To create counterfactual data, we follow the original paper and adopt the framework proposed by Longpre et al. (2021), which modifies the context to support a counterfactual answer. Specifically, we follow Longpre et al. (2021) and replace the gold entity answer in the context with a randomly sampled entity of the same entity type from the corpus. This process filters the original NQ test set into a subset of 4747 examples, each with a valid counterfactual instance.

To measure LLMs' ability to update answers, we also need to ensure that LLMs have the knowledge of original answers in the first place. Therefore, we further filter the dataset by only keeping examples that the LLM can answer without any context. In this way, we create a subset for each LLM we test on, and take intersection of the corresponding subsets when we analyze the effectiveness of the strategies on a group of different models.

To generate original and counterfactual demonstrations, the original work uses KATE (Liu et al., 2022) with the NQ train set to filter a set of best demonstration for each test example, which is not feasible to us. Therefore, we filter 16 examples from the NQ train set for each model that the model can correctly answer without any context as original demonstration examples and make their counterfactual instances as counterfactual demonstration.

For prediction with abstention, Zhou et al. (2023) curate their own dataset from RealTime QA (Kasai et al., 2022), which contains 113 examples, and is open to download. Each example can be formulated as *(c, q, o)*, where *c* denotes a paragraph of context, *q* is the question, and *o* is a set of options, including "`I don't know`" option. We use the same curated dataset for reproduction.

In our stretch goal for the In-Context Learning (ICL) settings, following the setup by Min et al. (2022), we employ subsets of the following three binary classification datasets: GLUE-rte (Wang et al., 2018), GLUE-mrpc (Wang et al., 2018), and tweet_eval-hate (Barbieri et al., 2020). Due to limited computation, we sample 800 examples from the GLUE datasets and 1000 examples from the tweet_eval-hate dataset. For each setting, we sample three times with different seeds and take the average to generate final evaluation scores. For the purposes of our study, we manually modify the labels to 'True' and 'False' for the two GLUE datasets, and to 'favor' or 'against' for the tweet_eval-hate dataset.

3.3 Hyperparameters

No additional training is needed in Zhou et al. (2023) and Min et al. (2022)'s work.

3.4 Implementation

We build on codebase released by Zhou et al. (2023) and Min et al. (2022) to adapt prompting on a wider range of models. The original codebase could is publicly available at https://github.com/wzhouad/context-faithful-llm and https://github.com/Alrope123/rethinking-demonstrations. Our adapted code will be released at https://github.com/lihaoxin2020/cse481n-team1 and https://github.com/mk322/rethinking-demonstrations.

3.5 Experimental Setup

For knowledge conflict, we use the same set of evaluation metrics as the original paper despite the memorization ratio. Specifically, we measure the frequency that the LLMs' predictions *contain* an exact match of the original answers (p_o) and the substituted answers (p_s), after both predictions and answers have been normalized by removing stop words and punctuation. We drop memorization ratio (M_R), which is calculated as $M_R = \frac{p_o}{p_o+p_s}$ since some of our models have $p_o = 0$ with low p_s, which leads to an M_R of 0 and therefore invalidates M_R metric. We conduct experiments in three different settings: zero-shot, demonstration using original instances, and demonstration using counterfactual instances. We retrieve demonstrations from the original/counterfactual training set, and evaluate LLMs on the counterfactual test

set. In the few-shot setting, we use 16 demonstration instances as the original work, and some instances are filtered out due to the limited input length of different models.

We extend the aforementioned experiments by running them on models of different sizes (from the same family) and models with different training routines (only pre-trained vs instruction-finetuned) to see the impact of each of those factors. To compare size, we take an intersection of all the "known" examples of all the models, then test the performance of each model on this intersected collection of examples. We use a similar setup to compare the effect of the training routine.

For prediction with abstention, Zhou et al. (2023) report accuracy on the entire dataset (All), accuracy on the subset of questions that can be answered based on retrieved documents (HasAns), and accuracy on questions that cannot be answered based on retrieved documents (NoAns). They also use the Brier score to evaluate the accuracy of the estimation, which measures the mean squared difference between the estimation and the true binary outcome of answerability. The probability of each choice is calculated by $P(\text{choice}|\text{prompt})$ followed by normalization across all choices, and we take the choice with the largest probability as the prediction. Probability scaling for length does not help the performance, as mentioned in the original paper.

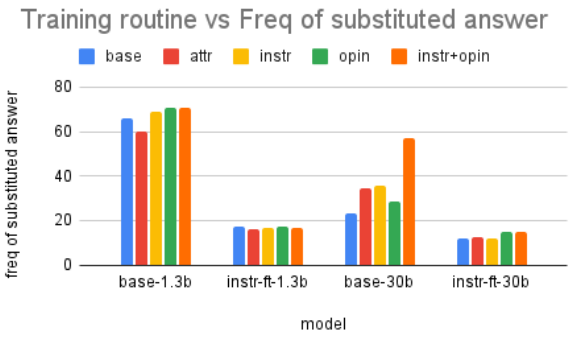

Figure 2: Results of training routine's impact on p_s in (%) using OPT family of models (opt and opt-iml). Experiments were conducted in zero-shot setting.

In our case, we use the same metrics and predicting method for experiments conducted on GPT-3.5, and we add the metric that measures the frequency that LLMs generated text *contain* an exact match of the correct choice after normalization. Furthermore, we observed that open-source LLMs suffer from long context in few-shot settings due to the long context given for this subproblem. The original work uses three demonstrations for each instance in the few-shot setting and filters out instances if the input length exceeds the maximum input length of each model. However, for LLMs we test on, most models have a maximum input length of only 2048 tokens, i.e., half of `text-davinci-003`'s input length. Filtering would keep only around 40 examples, which is not enough for analysis, and therefore we drop few-shot experiments.

Inspired by advice given by classmates from the NLP capstone session, we introduce 5 new prompting templates, each building on the aforementioned 5 templates by adding the sentence "`Answer "I don't know" if information is not given.` " before the options list. We call them explicit as opposed to the original implicit templates which do not elaborate in which case the model should predict "`I don't know`".

For the In-Context Learning (ICL) settings, we apply several manipulations to the datasets, notably in terms of label arrangements and prompt configurations. To assess the model's reliance on original and counterfactual demonstration, each model is associated with 16 training examples (16-shot learning) for each random seed as original demonstration, and the labels are deliberately flipped to generate counterfactual demonstration. Furthermore, we integrate proposed prompting strategies into the prompt template to evaluate their impact on the model's performance. These strategies, which include BASE, ATTR, INSTR, OPIN, and OPIN+INSTR, are designed to guide the model's focus and enhance its learning within the context. The prompts, imbued with these strategies, serve as the models' input, and we carefully monitor how effectively each model utilizes these prompts to improve in-context learning. To gauge the performance of the models under these varying conditions, we use binary F1 scores as our evaluation metric. The F1 score, a harmonic mean of precision and recall, provides a balanced measure of the models' performance, particularly in scenarios with imbalanced classes. Thus, it offers a robust means of evaluating the models' in-context learning abilities across different scenarios and prompting strategies.

3.6 Computational Requirements

Resources we have include nlpg hosts, Hyak hosts, and a limited OpenAI API budget paid from our own pocket. **For knowledge conflict**, the most computationally expensive part is filtering the dataset for each model from 4747 examples. We manage to run models up to 30B parameters on Hyak and it requires about 10 hours to run one pass of the dataset. Fortunately, filtering only needs one pass, and more than half of the original data will be filtered out. Therefore, one trial of experiment on one model would take at most 5 hours. In the original codebase, they mention that an estimated cost using OpenAI API would be around $150 for each trial using `text-davinci-003`.

For prediction with abstention, due to the small sample size, it usually takes less than 15 minutes to inference on the whole dataset once. The estimated OpenAI API cost would be $30 for each trial using `text-davinci-003`. We also test on the latest GPT-4, which cost around $20 per trial. Note that GPT-4 is less costly than InstructGPT due to the output and evaluation difference mentioned above.

For the ICL experiments, we used both OpenAI API to run GPT-3.5 and three GPUs with 72GB memory in total to run the two 30B models. The estimated OpenAI API cost would be $500 in total for `text-davinci-003`. The total estimated GPU time for running each 30B model is around 50 hours.

4 Results

4.1 Knowledge Conflict

We used the different prompting strategies presented to verify if they really do make the model context-faithful. Our results seemed to be a mixed bag. The original paper's results showed that using their prompting templates and counterfactual demonstration, the models predicted the substituted answers considerably more compared to just the base prompt, but we do not see this result reliably in our data. The original authors also found that OPIN and OPIN+INSTR prompts performed better than other strategies, but this is another result we cannot corroborate with our data. One result from the paper that we were able to see in our result as well was that providing demonstrations (both original and counterfactual) seemed to help with increasing the frequency of substituted answers and reducing the frequency of original answers, although it must be mentioned that in the original paper, the counterfactual demonstrations performed much better than the original demonstrations. We can also infer from Table 1 that the prompting templates suggested by the original seems to predominately work only in the zero-shot scenarios and not so much in both the few-shot scenarios.

We also tried to understand the impact of factors such as a model's size and training routine on its performance being context faithful when in a knowledge conflict. We found that base models performed better than their instruction-finetuned counterparts (shown in Figure 2). Figure 3 and Figure 4 show the impact of increasing model size which results in improvements in the in-context learning performance, particularly in the few-shot setting with both original and counterfactual demonstrations.

	Method	$p_s \uparrow$					$p_o \downarrow$				
		base	**attr**	**instr**	**opin**	**instr+opin**	**base**	**attr**	**instr**	**opin**	**instr+opin**
Zero-shot	FLAN-Large	34.03	33.25	33.25	**39.79**	22.55	6.81	7.33	7.33	**6.02**	3.71
	FLAN-XL	25.70	27.02	31.25	20.45	**31.75**	11.26	9.94	9.47	8.44	**8.37**
	OPT-1.3B	29.53	28.36	29.82	26.02	**33.33**	20.76	**15.50**	18.13	18.42	18.71
	OPT-30B	72.45	72.65	64.95	73.52	**80.97**	0.87	0.87	**0.68**	0.97	0.78
	Llama-7B	10.31	14.35	39.60	42.99	**64.45**	75.52	61.41	32.89	42.32	**15.21**
	Alpaca-7B	8.80	12.16	6.96	33.08	**48.66**	81.06	74.58	76.59	54.83	**34.54**
Original	FLAN-Large	**46.66**	39.99	39.99	**46.66**	**46.66**	13.33	6.67	6.67	**0.00**	**0.00**
	FLAN-XL	66.67	60.00	**73.33**	66.67	**73.33**	**6.67**	**6.67**	**6.67**	**6.67**	**6.67**
	OPT-1.3B	23.08	15.38	**30.77**	7.69	7.69	0.00	7.69	7.69	**0.00**	**0.00**
	OPT-30B	**20.88**	19.69	20.79	8.42	4.88	0.29	0.29	**0.19**	3.99	2.88
	Llama-7B	48.65	57.47	**64.45**	53.45	64.16	37.03	27.34	25.59	29.00	**24.48**
	Alpaca-7B	36.05	43.83	46.34	38.36	**51.64**	52.32	44.00	38.13	50.09	**35.16**
Counter	FLAN-Large	**46.67**	40.00	40.00	**46.67**	**46.67**	13.33	6.67	6.67	**0.00**	**0.00**
	FLAN-XL	66.67	60.00	**73.33**	66.67	**73.33**	**6.67**	**6.67**	**6.67**	**6.67**	**6.67**
	OPT-1.3B	23.08	15.38	**30.77**	7.69	7.69	**0.00**	7.69	7.69	**0.00**	**0.00**
	OPT-30B	**19.79**	17.59	10.64	7.77	4.88	**0.29**	0.49	9.97	4.66	2.66
	Llama-7B	61.05	67.69	**73.10**	**73.10**	72.73	29.96	21.05	18.52	18.51	**17.48**
	Alpaca-7B	39.61	47.48	**52.79**	43.42	59.70	49.34	39.98	33.78	44.40	**26.66**

Table 1: Results (in %) in the knowledge conflict setting. The best results for each model in each setting are highlighted in **bold**. The best result in each setting is highlighted in green

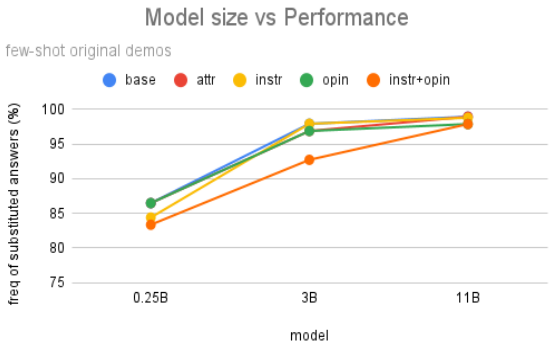

Figure 3: Model size vs Performance using FLAN-t5 with original demonstrations.

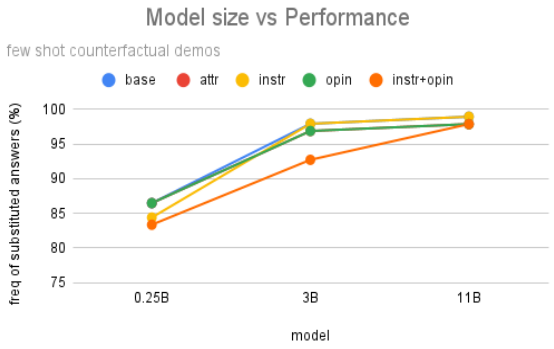

Figure 4: Model size vs Performance using FLAN-t5 counterfactual demonstrations.

4.2 Prediction with Abstention

We test different prompting strategies on prediction with abstention subproblems on different models. We first set up our experiments on GPT-3.5 with the exact same settings as the original authors and manage to reproduce the scores reported in the paper. Then we move on to different families of open-source LLMs with different sizes.

Table 2 presents results tested on FLAN-t5-XXL and predicted choice is determined by maximizing $P(\text{choice}|\text{prompt})$. FLAN-t5-XXL is the best-performing model on HasAns subset among all open-source models we tested on, which indicates that it has a strong ability to contextualize and extract information from the given prompt. The results demonstrate that the model is unable to recognize unanswerable questions with base prompts under an implicit setting. In this case, the OPIN and OPIN + INSTR templates boost accuracy on zero-shot settings for NoAns subset by 14.3% and 16.3% respectively, and trigger the LLM's ability to distinguish not answerable queries from answerable queries. When we remind the model of the "I don't know" option explicitly in the prompt, even the BASE prompt can achieve better performance than the best scores achieved by implicit prompting. All proposed templates, excluding the INSTR template, boost accuracy on NoAns subset by more than 10%, and the largest improvement is made by the OPIN template with 16.6% improvement.

Extending to a wider range of LLMs, we do not observe consistent improvement in different LLMs with proposed prompting templates. Figure 5 shows the Brier score across different sizes of FLAN-t5 models (FLAN-t5-BASE, FLAN-t5-LARGE, FLAN-t5-XL, and FLAN-t5-XXL corresponding to parameter size 0.25B, 0.78B, 3B, and 11B respectively) under the implicit zero-shot setting of curated RealTime QA. On smaller models, proposed templates achieve similar, or even higher Brier scores than the BASE template. Zhou et al. (2023) observe the same phenomenon in their test across different sizes of InstructGPTs, and they hypothesize that the reason is smaller LLMs have inferior reading comprehension ability.

However, unlike the test results shown in the original paper, the largest model does not indicate the best answerability judgment overall. The FLAN-t5-LARGE seems to be the best-performing model on answerability judgment across all 5 prompting templates, although it is smaller than FLAN-t5-XL and FLAN-t5-XXL.

We also observe the some recently released LLMs perform really bad on this task with almost no capability to recognize unanswerable questions with $P(\text{choice}|\text{prompt})$ prediction accuracy of 0% on NoAns subset and low accuracy on HasAns subset. We conducted various sanity checks and turn out no prompting strategies that Zhou et al. (2023) proposed can leverage the performance. An example would be Alpaca-7B performance shown in Table 3. We therefore try to evaluate via text generated from the model, which results in much better accuracy on HasAns subset, but, still, the proposed templates are unable to trigger the model to predict "I don't know" for unanswerable questions.

4.3 In-Context Learning Stretch Goal

In this study, OPT-iml-max-30b and GPT-3.5 are both instruction-finetuned models, while OPT-30b is only pre-trained without additional finetuning. Looking at the 'Correct labels' scenario results in the Figure 6, it becomes apparent that the instruction-finetuned models, particularly OPT-iml-max-30b, show noticeable improvements in performance when applied with different prompting strategies such as INSTR and the OPIN+INSTR. In stark contrast, the performance

	Method	Acc↑ NoAns	Acc↑ All	Brier↓ All
Implicit	Base	0.0	55.9	41.0
	Attr	2.0	56.8	37.7
	Instr	0.0	55.0	40.9
	Opin	14.3	61.3	34.0
	Opin + Instr	**16.3**	**62.2**	**33.1**
Explicit	Base	22.4	64.9	33.2
	Attr	36.7	71.2	26.6
	Instr	20.4	64.0	34.0
	Opin	**38.8**	**72.1**	**25.4**
	Opin + Instr	32.7	69.4	26.9

Table 2: Results from FLAN-t5-XXL (in %) on zero-shot RealTime QA. The overall best results in each setting are highlighted in **bold**. As all prompts achieve *nearly* perfect accuracy (> 98%) on the **HasAns** subset, it is not included in the table.

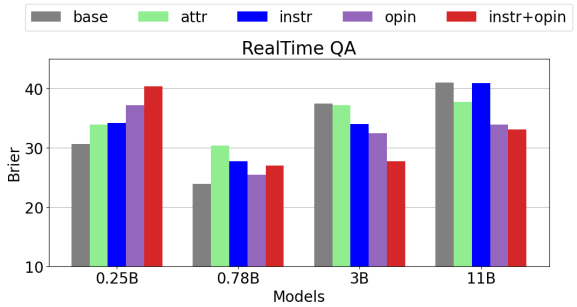

Figure 5: Brier scores across different sizes of FLAN-t5, evaluated in implicit zero-shot setting of RealTime QA.

	Method	Acc↑ HasAns	Acc↑ NoAns	Acc↑ All	Brier↓ All
Implicit	Base	32.3	0.0	18.2	44.8
	Attr	27.4	0.0	15.5	44.4
	Instr	29.0	0.0	16.4	45.2
	Opin	30.6	0.0	17.3	44.8
	Opin + Instr	29.0	2.1	17.3	44.3

	Method	Acc↑ HasAns	Acc↑ NoAns	Acc↑ All
Implicit	Base	79.0	6.3	47.3
	Attr	85.5	6.3	50.9
	Instr	82.3	6.3	49.1
	Opin	79.0	6.3	47.3
	Opin + Instr	85.5	6.3	50.9

Table 3: Results from Alpaca (in %) on RealTime QA. The **left** shows scores measured on the probability of each choice and the **right** shows scores measured on exact matches found within the generated text, so no Brier score is available.

of the pre-trained model OPT-30b remains constant across all prompting strategies, suggesting that it may not utilize these strategies as effectively. As presented in Table 4, GPT-3.5 also performs remarkably well across all prompting strategies, further supporting the observation that instruction-finetuned models are better at exploiting these prompts to enhance in-context learning. From these findings, it can be inferred that instruction-finetuning might be a valuable step in preparing language models for more effective use of prompting strategies, leading to better performance in in-context learning tasks.

With flipped labels as counterfactual demonstration, shown in Figure 6, the performance of both OPT-30b and OPT-iml-max-30b drops significantly on the BASE template, illustrating that accurate input-label mappings are critical for optimal in-context learning.

However, the effect of counterfactual demonstration combined with proposed prompting templates is unclear. It can be observed that the flipped BASE score is lower than the scores obtained with other strategies, such as flipped INSTR, flipped OPIN, and flipped OPIN+INSTR. This suggests that the prompting strategies, although beneficial under correct labeling conditions, do not necessarily make the models more faithful to the context when faced with flipped labels. Therefore, the robustness of these strategies under varying label conditions might need further exploration and refinement.

5 Discussion

Based on our experiments conducted on a wide range of modern LLMs with different context-specific tasks, the prompting templates proposed by Zhou et al. (2023) tend to be more effective under a zero-shot setting. However, we observe no evidence that shows counterfactual demonstration provides consistent improvement across modern LLMs of different families or sizes compare to original demonstration. Due to limited resources, we are not able to conduct experiments on model as huge as `text-davinci-003`, and even for `text-davinci-003`, we are only able to conduct

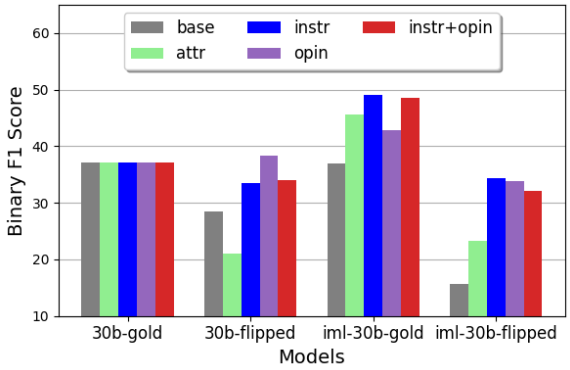

Figure 6: Results of binary F1 scores (%) in the in-context learning experiments on the two OPT-30b language models for the five prompting strategies.

	Method	GPT3-davinci-003
Gold labels	Base	66.7
	Attr	68.3
	Instr	67.7
	Opin	69.3
	Opin + Instr	68.6

Table 4: Results of binary F1 scores (%) in the in-context learning experiments on the GPT3 model for the five prompting strategies with the gold input-label mappings.

very limited amount of tests on that. Therefore, it is unclear that whether proposed strategies are more effective on giant models with more than 30B parameters.

5.1 What was Easy

- Preliminary research stage was easy and fun. We chose a paper that is well written with plenty of open resources that help us along the way.
- Data pre-processing occurred smoothly without much difficulty. Thanks to Longpre et al. (2021) and Zhou et al. (2023) for publishing their good work.

5.2 What was Difficult

- We couldn't reproduce a lot of the experiments in the original paper because of their high expense (using GPT-3, paid datasets, and lack of compute).
- Trying to understand the randomness in our data as they did not really follow the general trends associated with language models and spent lots of time on sanity check. For example, we would expect larger models to perform better than their smaller counterparts on the same set of examples, but this wasn't always the case as evidenced by the results in Figure 2 and in Table 5
- Following on from the previous point, we experienced a lack of robustness in almost all of our models, more so, in the instruction-finetuned variants. There were many instances where the base model would predict the right/expected answer in the knowledge conflict setting, but their instruction-finetuned variants outputted meaningless text or blank predictions. We struggled to explain this behavior and in the end and unable to reach a compelling theory for this phenomenon.

5.3 Recommendations for Reproducibility

The original paper concluded that they proposed "two methods, opinion-based prompts and counter-factual demonstrations, are effective in improving LLMs' faithfulness to contexts". In our work, the proposed strategies have been shown to be not consistently effective across a wide range of modern open-source LLMs. We appreciate the straightforward idea that improves context-faithfulness of InstructGPT without expensive training, but the methods are not generalizable to a wide range of LLMs and we think the conclusion in the original paper requires more experiments to be settled.

Communication with Original Authors

One of the original authors and the maintainer of the paper's GitHub repository, Wenxuan Zhou, was particularly helpful to us. They responded in a timely manner to any doubts we had. They were also kind enough to provide the code they used to filter the datasets used the Knowledge Conflict experiments. Overall, we feel that we had good support from their side to carry out the reproduction of their paper.

References

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Johannes Bjerva, Nikita Bhutani, Behzad Golshan, Wang-Chiew Tan, and Isabelle Augenstein. 2020. SubjQA: A Dataset for Subjectivity and Review Comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5480–5494, Online. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Mansi Gupta, Nitish Kulkarni, Raghuveer Chanda, Anirudha Rayasam, and Zachary C Lipton. 2019. Amazonqa: A review-based question answering task.

Srinivasan Iyer, Xi Victoria 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 Ves Stoyanov. 2023. Opt-iml: Scaling language model instruction meta learning through the lens of generalization.

Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. 2022. Realtime qa: What's the answer right now?

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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 *EMNLP*.

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, Melbourne, Australia. Association for Computational Linguistics.

George Stoica, Emmanouil Antonios Platanios, and Barnabás Póczos. 2021. Re-tacred: Addressing shortcomings of the tacred dataset.

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`.

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

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

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. Context-faithful prompting for large language models. *arXiv preprint arXiv:2303.11315*.

Appendices

A Model size impact in the knowledge conflict setting

We also investigated how the model size affects the knowledge conflict tests for zero-shot settings in FLAN-T5 models.

Model	$p_s \uparrow$				
	base	**attr**	**instr**	**opin**	**instr+opin**
0.25B	85.57	92.30	82.99	84.61	87.99
0.75B	82.69	77.88	48.99	74.99	28.99
3B	89.42	82.69	50.99	71.15	49.99
11B	99.03	94.23	97.99	93.27	96.99

Table 5: Zero shot performance results for different models sizes from the FLAN-T5 family of models (in %) in the knowledge conflict setting.

B Prediction with Abstention Results on Different Models

Here we show the results of prediction with abstention tests on different families and sizes of open-source LLMs.

Method	NoAns Acc↑					Brier↓				
	base	**attr**	**instr**	**opin**	**instr+opin**	**base**	**attr**	**instr**	**opin**	**instr+opin**
opt-1.3B	6.1	6.1	4.1	2.0	0.0	38.0	39.4	38.7	40.1	41.0
opt-30B	6.1	4.1	4.1	4.1	2.0	38.2	37.2	38.8	39.8	40.3
opt-iml-1.3B	8.2	8.2	8.2	8.2	8.2	41.4	41.4	41.3	41.4	41.1
opt-iml-30B	0.0	0.0	0.0	0.0	0.0	43.9	43.9	44.0	43.7	44.1
opt-iml-max-30B	0.0	0.0	0.0	0.0	0.0	43.1	43.2	43.4	43.3	43.2
LLaMA-7B	8.3	8.3	6.3	6.3	4.2	41.9	42.0	42.9	42.7	43.1

Table 6: Results (in %) in the abstention with prediction under zero-shot setting using $P(\text{choice}|\text{prompt})$ prediction.