Crafting a Good Prompt or Providing Exemplary Dialogues? A Study of In-Context Learning for Persona-based Dialogue Generation

Anonymous ACL submission

Abstract

Previous in-context learning (ICL) research has focused on tasks such as classification, machine translation, text2table, etc., while studies on whether ICL can improve human-like dialogue generation are scarce. Our work fills this gap by systematically investigating the ICL capabilities of large language models (LLMs) in persona-based dialogue generation, conducting extensive experiments on high-quality real human Chinese dialogue datasets. From experimental results, we draw three conclusions: 1) adjusting prompt instructions is the most direct, effective, and economical way to improve generation quality; 2) randomly retrieving demonstrations (demos) achieves the best results, possibly due to the greater diversity and the amount of effective information; counter-intuitively, retrieving demos with a context identical to the query performs the worst; 3) even when we destroy the multi-turn associations and single-turn semantics in the demos, increasing the number of demos still improves dialogue performance, proving that LLMs can learn from corrupted dialogue demos. Previous explanations of the ICL mechanism, such as n-gram induction head, cannot fully account for this phenomenon (Code is available at Dialog ICL).

1 Introduction

Current chatbots based on LLMs have superior performance on question answering, polishing documents, etc. (Srivastava et al., 2022). However, for the task of persona-based dialogue generation, these universal LLMs still lag far behind real humans[1]. Persona-based dialogue generation is crucial and has practical application value. For instance, optimizing persona dialogue generation is of great significance for empathetic and medical chatbots (De Gennaro et al., 2020; Liu et al., 2022b), and it can bring trust to users (Huang et al., 2023; De Gennaro et al., 2020).

[1]As shown in Table 12, despite numerous specifications, GPT-4 tends to generate unnatural responses in most cases.

On the other hand, fine-tuning a high-quality persona-based dialogue model usually requires a proprietary dataset, and the cost of manually writing dialogues is very high (Cao et al., 2022; Huang et al., 2023). Since the valid context length of LLMs is constantly growing (Xiong et al., 2023), we believe that ICL may be an effective way to generate high-quality dialogues at a low cost. However, research papers on how ICL affects dialogue generation are few (Dong et al., 2022). Only one paper Xu et al. (2023) investigates the ability of LLM to learn from dialogue demos, but these demos only contain character background information, which is quite different from real human conversations. In this paper, we experiment with the ICL capability of LLMs on dialogues of real-human conversations, systematically exploring the capabilities of ICL in persona-based dialogue generation, including how we should choose the demo retrieval method, whether we should focus more on the text quality of the demo itself or the input-output mapping, whether providing more demos can improve the results, and give insight on what the LLM learns and does not learn from the provided demos.

2 Problem Formulation

We investigate the LLMs' ICL ability to perform turn-level persona-based dialogue generation using the prompt with instructions and demos. We begin by defining notations. We denote the persona description as per and the query context before the t-th round of dialogue as c_{t-1}, where $c_{t-1} = (r_0, \ldots, r_{t-1})$ and r_{t-1} is the most recent dialogue turn. A dialogue demo of ℓ turns can be represented as $x = (per, c_\ell)$, where $c_{:\ell-1}$ can be interpreted as the input part of the demo while r_ℓ as the output (y label) part. A set of k dialogue demos is denoted as $\mathbf{x}^k_{demo} = (x^0, \ldots, x^k)$. We define $\mathbf{p}$ as the current task's prompt. Additionally, we have a template function $\mathcal{T}$ for integrating k demos, the task prompt, query persona, and query context. We can represent the LLM's output response at current round t as $\hat{r}_t = \mathcal{LLM}(\mathcal{T}(\mathbf{x}^k_{demo}, \mathbf{p}, per, c_{:t-1}))$. See §A.2 for filled templates.

Our work primarily investigates three research questions. **RQ1**: For dialogue generation via ICL, should we focus on tuning the task prompt, providing high-quality demos, or both? **RQ2**: What is the impact of different demo retrieval methods on ICL? Do the number of demos and their context length make a difference? **RQ3**: From which aspects do LLMs learn useful information from demos? We analyze this from the perspectives of multi-turn correlation, single-turn semantics, input-label mapping, and token distributions.

3 Evaluation Metrics for Generation

In the experiment, given a persona description per and a context $c_{:t-1}$, we evaluate a set of model responses from three dimensions: **intra-diversity**, **inter-similarity** to expert-written responses, and **response quality**. Intra-diversity is the average of multiple traditional metrics, including Distinct-n (Li et al., 2015), Entropy-n (Serban et al., 2017), self-bleu (Zhu et al., 2018), and cosine similarity based on sentence embeddings. The similarity to expert responses mainly considers rougeL (Lin, 2004), bleu (Papineni et al., 2002), and cosine similarity based on sentence embeddings. The response quality is scored by our self-trained Response Evaluator, which takes into account persona consistency, context logic, interestingness, and colloquial expressions. Our self-trained Response Evaluator significantly outperforms crowd-sourced evaluations and other automated evaluation schemes in terms of consistency with expert ratings. See §A.1.1 for details of how each evaluation metric is calculated and §A.5 for how the Response Evaluator is trained.

4 Different Prompt and ICL Settings

Below we show settings to concatenate prompts and demos (examples are provided in §A.2).
Context Only: We only provide the persona description and context containing $t-1$ turns.
Prompt Only: We only provide the persona description, the task prompt, and the context $c_{:t-1}$.
Few Shot Demo: We provide k demos, the persona description and the context $c_{:t-1}$.
Few Shot Demo+Prompt: We provide k demos, the task prompt, the persona, and the context $c_{:t-1}$.

For the setting of demo retrieval, we attempt three methods. The first method **randomly** selects a demo with the same context length, the second method (following (Su et al., 2022)) retrieves the most similar demos (with the same length as the query) based on the cosine similarity of sentence **embeddings**, and the third method provides demos that contain the exact **same** context ($c_{:\ell-1}$) as the query context, each with a high-quality response r_ℓ written by a human expert.

5 Experimental Settings

5.1 Evaluation LLMs & Dataset

For LLMs, we select GPT-3.5-turbo (Ye et al., 2023b), GPT-4 (Achiam et al., 2023), and Ernie (Sun et al., 2021) because they have the abilities to follow instructions and perform in-context learning. We employ nucleus sampling (Holtzman et al., 2020) for decoding, setting both of the top-p and temperature parameters to 0.9.

Regarding the evaluation dataset $\mathcal{D}_{eval}$, we manually selected 6 personas with distinct personality backgrounds, with an average of 150 Chinese characters per persona, and an average of 15 turns of conversation. Each turn has an average of 35 characters. All the dialogues are written by students from the Department of Chinese Language and Literature. Compared to previous open-source Chinese dialogue data (Papangelis et al., 2020), our persona descriptions are much more complex, and conforming to the majority of commercial chat-bot-based products[2], with higher dialogue quality and richer content (see more examples in §A.2). We keep the test set compact due to our limited budget (GPT-4 and ernie-bot-4 are very expensive). We also provide a demo dataset $\mathcal{D}_{demo}$ where demos are retrieved (for **random/embedding** method). It contains 800 dialogues, with each written by a human expert. $\{\mathcal{D}_{demo}\} \cap \{\mathcal{D}_{eval}\}$ equals $\emptyset$.

5.2 Ablation Settings

w/o correct y label: response r_l for each demo context c_{l-1} is replaced with a random response sampled from from $\mathcal{D}_{demo}$.
w/o contextual semantic: turns in each demo context c_{l-1} are shuffled.
w/o turn semantic: tokens[3] in each demo turn r_i are shuffled (including the last label turn).
w/o contextual/turn semantic: tokens in each demo turn r_i and turns in each demo context $c_{:l-1}$

[2] English products include character.ai and inworld.ai, while Chinese products include Minimax's Xingye.
[3] each turn is tokenized by jieba.

	Prompt Construction Method	Intra-Diversity	Inter-Similarity	Response Quality
GPT-3.5	Context Only^0	$\mathbf{0.671}$	$\mathbf{0.269}\uparrow^{1}$	$\mathbf{0.147}\uparrow^{4}$
	Prompt Only^1	$\mathbf{0.684}\uparrow^{0,4}$	$\mathbf{0.228}$	$\mathbf{0.191}\uparrow^{0,3,4}$
	Few Shot Demo (Random)^2	$\mathbf{0.686}\uparrow^{0,4}$	$\mathbf{0.277}\uparrow^{1}$	$\mathbf{0.160}\uparrow^{4}$
	- w/o correct y label	$\Delta 0.008\dagger$	$-\Delta 0.040\dagger$	$\Delta 0.004$
	- w/o contextual semantic	$\Delta 0.005$	$-\Delta 0.013\dagger$	$\Delta 0.026\dagger$
	- w/o turn semantic	$-\Delta 0.005$	$-\Delta 0.021\dagger$	$\Delta 0.010$
	- w/o contextual/turn semantic	$-\Delta 0.007$	$-\Delta 0.023\dagger$	$-\Delta 0.000$
	Few Shot Demo (Emebdding)^3	$\mathbf{0.680}\uparrow^{0,4}$	$\mathbf{0.269}\uparrow^{1}$	$\mathbf{0.156}\uparrow^{4}$
	- w/o correct y label	$\Delta 0.006$	$-\Delta 0.020\dagger$	$-\Delta 0.019$
	Few Shot Demo (Same Persona/Context)^4	$\mathbf{0.667}$	$\underline{\mathbf{0.286}\uparrow^{all}}$	$\mathbf{0.094}$
	- w/o correct y label	$-\Delta 0.014\dagger$	$-\Delta 0.079\dagger$	$-\Delta 0.042\dagger$
	Few Shot Demo (Random) + Prompt^5	$\underline{\mathbf{0.696}\uparrow^{all}}$	$\mathbf{0.243}\uparrow^{1}$	$\underline{\mathbf{0.222}\uparrow^{all}}$
GPT-4	Context Only^0	$\mathbf{0.655}$	$\mathbf{0.269}\uparrow^{1,2,3}$	$\mathbf{0.155}$
	Prompt Only^1	$\mathbf{0.704}\uparrow^{all}$	$\mathbf{0.248}$	$\underline{\mathbf{0.476}\uparrow^{0,2,3,4}}$
	Few Shot Demo (Random)^2	$\mathbf{0.679}\uparrow^{0,4}$	$\mathbf{0.258}\uparrow^{1}$	$\mathbf{0.235}\uparrow^{0,4}$
	- w/o correct y label	$\Delta 0.002$	$-\Delta 0.034\dagger$	$-\Delta 0.035\dagger$
	- w/o contextual semantic	$\Delta 0.005$	$\Delta 0.001$	$-\Delta 0.010$
	- w/o turn semantic	$-\Delta 0.018\dagger$	$-\Delta 0.002$	$-\Delta 0.034\dagger$
	- w/o contextual/turn semantic	$-\Delta 0.008\dagger$	$-\Delta 0.015\dagger$	$-\Delta 0.062\dagger$
	Few Shot Demo (Emebdding)^3	$\mathbf{0.680}\uparrow^{0,4}$	$\mathbf{0.261}\uparrow^{1}$	$\mathbf{0.243}\uparrow^{0,4}$
	- w/o correct y label	$\Delta 0.004$	$-\Delta 0.024\dagger$	$-\Delta 0.029\dagger$
	Few Shot Demo (Same Persona/Context)^4	$\mathbf{0.658}$	$\underline{\mathbf{0.294}\uparrow^{all}}$	$\mathbf{0.193}\uparrow^{0}$
	- w/o correct y label	$\Delta 0.003$	$-\Delta 0.070\dagger$	$-\Delta 0.095\dagger$
	Few Shot Demo (Random) + Prompt^5	$\underline{\mathbf{0.710}\uparrow^{0,2,3,4}}$	$\mathbf{0.264}\uparrow^{1,2}$	$\mathbf{0.470}\uparrow^{0,2,3,4}$
Ernie	Context Only^0	$\mathbf{0.690}$	$\mathbf{0.265}$	$\mathbf{0.345}$
	Prompt Only^1	$\underline{\mathbf{0.720}\uparrow^{0,2,3,4}}$	$\mathbf{0.263}$	$\mathbf{0.519}\uparrow^{0,4}$
	Few Shot Demo (Random)^2	$\mathbf{0.706}\uparrow^{0,3}$	$\mathbf{0.269}$	$\mathbf{0.487}\uparrow^{0,4}$
	- w/o correct y label	$\Delta 0.023\dagger$	$-\Delta 0.025\dagger$	$-\Delta 0.075\dagger$
	- w/o contextual semantic	$\Delta 0.001$	$\Delta 0.007$	$\Delta 0.007$
	- w/o turn semantic	$-\Delta 0.005\dagger$	$-\Delta 0.001$	$-\Delta 0.054\dagger$
	- w/o contextual/turn semantic	$-\Delta 0.004$	$-\Delta 0.001$	$-\Delta 0.042\dagger$
	Few Shot Demo (Emebdding)^3	$\mathbf{0.698}\uparrow^{0}$	$\mathbf{0.281}\uparrow^{0,1,2,5}$	$\mathbf{0.500}\uparrow^{0,4}$
	- w/o correct y label	$\Delta 0.012\dagger$	$-\Delta 0.023\dagger$	$-\Delta 0.053\dagger$
	Few Shot Demo (Same Persona/Context)^4	$\mathbf{0.701}\uparrow^{0}$	$\underline{\mathbf{0.330}\uparrow^{all}}$	$\mathbf{0.340}$
	- w/o correct y label	$\Delta 0.009$	$-\Delta 0.120\dagger$	$-\Delta 0.166\dagger$
	Few Shot Demo (Random) + Prompt^5	$\mathbf{0.718}\uparrow^{0,2,3,4}$	$\mathbf{0.271}\uparrow^{1}$	$\underline{\mathbf{0.544}\uparrow^{all}}$

Table 1: Results of dialogue generation when few-shot k is set to 5. For each LLM, the best results are underlined. The up-arrow $\uparrow$ indicates statistical significance ($p < 0.05$ with Wilcoxon signed-rank test (Woolson, 2007)) when comparing two methods. For example, $\mathbf{0.684}\uparrow^{0,4}$ in the second row indicates *Prompt Only* method is better than *Context Only* method and *Few Shot Demo (Same Persona/Context)* method, and is statistically significant. We use Δ and $-\Delta$ to represent absolute increase and decrease in scores for ablations ($\dagger$ indicates statistical significance). The text in parentheses (such as *Random*) represents the demo's retrieval method.

are both shuffled (denoted as $\widetilde{c_{:l-1}}$).

6 Results Analysis

We present the performance of various settings (few-shot k is fixed to 5) for dialogue generation in Table 1. We present the response quality scores averaged over varying few-shot settings in Table 2. For more detailed chart data, please refer to §A.

Discussion of RQ1: From Table 1, we observe that for all LLMs, *Prompt Only* method scores much higher in response quality than using context or few-shot demos alone. From Table 2 and Figure 8, we can see that the few-shot ICL method requires a large number of demos (at least 7) to catch up with using only the prompt. *Overall, we believe that optimizing the prompt alone is the most cost-effective choice. But if funding permits, you may consider adding more demos to the prompt. At least for models gpt-3.5 and Ernie, we find that adding demos can improve the quality of responses.*

Discussion of RQ2: From Table 1, we can see that (comparing methods with No. 2, 3, and 4), the response quality of recalling demos of the same context is the worst, and this conclusion holds for all three language models. We hypothesize two reasons for this: First, the LLM may not be able to learn how to generalize from the (same) input-(different) output mapping from only 5 samples (as the one-to-many input-output format in the fine-tuning scenario also hinders model converging). *Second, the quality of the LLM-generated responses is directly proportional to the effective information in the demos.* The effective information may include token distribution, single-turn semantics, multi-turn associations, etc. Clearly, demos with the same context have the least effective information as the same context $c_{:l-1}$ is repeated 5 times, and the number of unique tokens is also the smallest (see Figure 3). For the other two methods, as we increase the effective information by increasing the few-shot k, the overall dialogue quality shows an upward trend (see Figure 9).

The response quality of retrieving **randomly** and retrieving through **embedding** is relatively close, but the former has much more diverse responses. *We believe that retrieving randomly is a strong baseline, and efforts should be focused on improving the quality of the demo set* $\mathbf{x}_{demo}$*, rather than the similarity between query context and demo context. When budget is sufficient, you may provide as many demos as possible because using larger* k *can improve response quality (proved in Figure 14).*

Discussion of RQ3: Analyzing all LLMs, in both fixed (k=5, Table 1) and varying few-shot settings (Figure 7), shuffling the context of the demo does not affect the quality of generated responses (in some cases, it even improves the quality, as in Table 1 for *gpt-3.5*). Furthermore, when comparing *w/o turn semantic* and *w/o contextual/turn semantic*, we do not find the latter to significantly reduce the response quality; for *ernie-bot-4*, we even find that the former has a greater negative impact (bottom right of Figure 7). Based on these results, we conjecture that *LLMs do not learn generalizable dialogue generation abilities from multi-turn logical or semantic associations in* $\mathbf{x}_{demo}$ *during ICL.*

The corruption of single-turn semantics and grammar significantly reduces the generation quality of the three LLMs, which *demonstrates that LLMs effectively learn single-turn semantics during the ICL process.* However, *the ability to learn input-output mapping (w/o correct y label) varies among LLMs*, with *gpt-3.5* performing poorly, while the other two models perform better.

What surprises us the most is that when we corrupt both single-turn and multi-turn semantics, the LLM-generated text quality remains significantly better than the *Context Only* setting (Table 1). And as we increase the few-shot k and the length of the context for these semantic-corruption methods (Figures 7 and 11 to 13), the response quality of LLMs still shows an upward trend (not particularly evident for GPT-4). Additionally, from Figure 2, we can see that LLM does not simply learn to copy tokens from demos; it generalizes. *We conclude that LLM has the potential to conduct ICL well even when provided with corrupted* $\mathbf{x}_{demo}$*. Specifically, the LLM can learn from demos' mapping of token-distribution of* $\widetilde{c_{1:l-1}}$ *and corrupted response* $\widetilde{r}_l$ *and improve the generation quality when a normal query context* $c_{1:t-1}$ *is provided.*

Aside from the above three research questions, we also discover some other interesting phenomena. Please refer to §A.6 for more details.

7 Connections between our experimental conclusions and previous work

Researchers Reynolds and McDonell (2021); Sun et al. (2022); Dong et al. (2022) find that in machine translation, classification, and other tasks, carefully crafted manual prompts can perform better than few-shot learning (consistent with **our** conclusion). Additionally, Reynolds and McDonell (2021) discovers that Zero-Shot (corresponding to our *Context Only* method) performance is also better than Few-Shot, which contradicts **our** findings.

Regarding the retrieval of demos, previous work has concluded that similarity-based retrieval is significantly better than random retrieval in tasks such as sentiment classification, table-to-text, and semantic parsing (Liu et al., 2022a; Rubin et al., 2022). However, **our** findings suggest that the random baseline performs the best, possibly due to its superior diversity, which can lead to compositional generalization (Levy et al., 2023).

Regarding whether LLMs can effectively learn y label during ICL, some work argues that input-output mapping is not important (Min et al., 2022), while others affirm the value of y label (Li and Qiu, 2023; Kossen et al., 2023). **Our** conclusion is that the y label is helpful for dialogue generation, but to a relatively small extent.

Regarding the relationship between the number of demos and the performance ICL, Reynolds and McDonell (2021) believes that there is no linear relationship between them, while Li et al. (2023); Wu et al. (2023); Hao et al. (2022) find that in story generation and classification tasks, the more demos, the better the performance (**our** experimental results support this claim, too).

8 Limitation

Due to limited resources, we have not attempted many more complex methods for selecting demos, and most of these methods have not been tested for their effectiveness in dialogue generation (Iter et al., 2023; Ye et al., 2023a). Additionally, we hope that future researchers can further expand the test set or explore whether similar conclusions can be drawn from ICL in other languages.

Our paper investigates the performance of ICL in persona-based dialogue generation and provides some practical suggestions, but does not delve into the underlying mechanisms of ICL in this context. Existing hypotheses, such as those based on n-gram (copy/induction) head, cannot fully account for our experimental results. We believe that the principles behind the effectiveness of ICL warrant further investigation.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ekin Akyürek, Bailin Wang, Yoon Kim, and Jacob Andreas. 2024. In-context language learning: Architectures and algorithms.

Yu Cao, Wei Bi, Meng Fang, Shuming Shi, and Dacheng Tao. 2022. A model-agnostic data manipulation method for persona-based dialogue generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7984–8002.

Mauro De Gennaro, Eva G Krumhuber, and Gale Lucas. 2020. Effectiveness of an empathic chatbot in combating adverse effects of social exclusion on mood. *Frontiers in psychology*, 10:3061.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.

Yaru Hao, Yutao Sun, Li Dong, Zhixiong Han, Yuxian Gu, and Furu Wei. 2022. Structured prompting: Scaling in-context learning to 1,000 examples. *arXiv preprint arXiv:2212.06713*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Qiushi Huang, Yu Zhang, Tom Ko, Xubo Liu, Bo Wu, Wenwu Wang, and H Tang. 2023. Personalized dialogue generation with persona-adaptive attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12916–12923.

Dan Iter, Reid Pryzant, Ruochen Xu, Shuohang Wang, Yang Liu, Yichong Xu, and Chenguang Zhu. 2023. In-context demonstration selection with cross entropy difference. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1150–1162, Singapore. Association for Computational Linguistics.

Jannik Kossen, Tom Rainforth, and Yarin Gal. 2023. In-context learning in large language models learns label relationships but is not conventional learning. *arXiv preprint arXiv:2307.12375*.

Itay Levy, Ben Bogin, and Jonathan Berant. 2023. Diverse demonstrations improve in-context compositional generalization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1401–1422, Toronto, Canada. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *CoRR*, abs/1510.03055.

Xiaonan Li, Kai Lv, Hang Yan, Tianyang Lin, Wei Zhu, Yuan Ni, Guotong Xie, Xiaoling Wang, and Xipeng Qiu. 2023. Unified demonstration retriever for in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4644–4668, Toronto, Canada. Association for Computational Linguistics.

Xiaonan Li and Xipeng Qiu. 2023. Finding support examples for in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6219–6235, Singapore. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, William B Dolan, Lawrence Carin, and Weizhu Chen. 2022a. 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.

Junfeng Liu, Christopher Symons, and Ranga Raju Vatsavai. 2022b. Persona-based conversational ai: State of the art and challenges. In *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 993–1001. IEEE.

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 arXiv:1907.11692*.

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.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. 2022. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*.

Alexandros Papangelis, Mahdi Namazifar, Chandra Khatri, Yi-Chia Wang, Piero Molino, and Gokhan Tur. 2020. Plato dialogue system: A flexible conversational ai research platform. *arXiv preprint arXiv:2001.06463*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671.

Philip Sedgwick. 2014. Spearman's rank correlation coefficient. *Bmj*, 349.

Iulian Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Hongjin Su, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, et al. 2022. Selective annotation makes language models better few-shot learners. *arXiv preprint arXiv:2209.01975*.

Tian-Xiang Sun, Xiang-Yang Liu, Xi-Peng Qiu, and Xuan-Jing Huang. 2022. Paradigm shift in natural language processing. *Machine Intelligence Research*, 19(3):169–183.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, et al. 2021. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv preprint arXiv:2107.02137*.

Robert F Woolson. 2007. Wilcoxon signed-rank test. *Wiley encyclopedia of clinical trials*, pages 1–3.

Zhiyong Wu, Yaoxiang Wang, Jiacheng Ye, and Lingpeng Kong. 2023. Self-adaptive in-context learning: An information compression perspective for in-context example selection and ordering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1423–1436, Toronto, Canada. Association for Computational Linguistics.

Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, et al. 2023. Effective long-context scaling of foundation models. *arXiv preprint arXiv:2309.16039*.

Xinchao Xu, Zeyang Lei, Wenquan Wu, Zheng-Yu Niu, Hua Wu, and Haifeng Wang. 2023. Towards zero-shot persona dialogue generation with in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1387–1398.

Jianhao Yan, Jin Xu, Chiyu Song, Chenming Wu, Yafu Li, and Yue Zhang. 2023. Understanding in-context learning from repetitions. *arXiv preprint arXiv:2310.00297*.

Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. 2023a. Compositional exemplars for in-context learning. *arXiv preprint arXiv:2302.05698*.

Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, et al. 2023b. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. *arXiv preprint arXiv:2303.10420*.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1097–1100.

A Appendix

In the appendix, we provide a more detailed introduction to the evaluation metrics for dialogue generation in §A.1.1 and explain how we select and retain the best prompt in §A.1.2.

In §A.2 and §A.3, we showcase filled templates under various settings, allowing for a clear representation of the LLMs' input text.

We describe the training approach for the Dialogue Embedding model in Section§A.4 and outline the training data and training details for the Response Evaluator in §A.5.

In §A.6, we present additional interesting experimental findings.

In §A.7, we provide a detailed overview of the annotator background, the manual composition process for high-quality reference responses from human experts, the evaluation consistency between the Response Evaluator and expert annotators, as well as the specific annotation guidelines.

A.1 More Details on Experimental Settings

A.1.1 Evaluation Metrics for Generation

In the experiment, given a persona description and a context, we let the LLM generate 15 responses, from which we retain a set of 5 responses that has the highest intra-diversity (the retention process finds the combination of 5 responses with the largest unique token set among all possible combinations). We adopt this setting because the responses generated by the LLM given the same context can be quite similar. We denote the retained response set as $\hat{\mathbf{S}_r}$, and the set of 5 expert-written responses as $\mathbf{S}_r$. We analyze the quality of the responses along three dimensions, as shown below.

Intra-Diversity: we calculate the Distinct-n (n=1,2,3) (Li et al., 2015), Entropy-n (n=1) (Serban et al., 2017), $1-$ *self-bleu score* (Zhu et al., 2018), and $1-$ *cosine-similarity score (mean)* for $\hat{S}_r$. We denote the cosine similarity between two unequal responses as v_{cos}. The embeddings of two responses are extracted by our self-trained encoder (see §A.4 for more details). We iterate through all unequal pairs in $\hat{S}_r$ to compute v_{cos} values, and the average cosine similarity is the mean of all v_{cos} values. The final diversity score s_{div} is the average of the four scores, with higher values indicating better response diversity in $\hat{S}_r$.

Inter-Similarity: To calculate the similarity score s_{sim} between $\hat{\mathbf{S}_r}$ and $\mathbf{S}_r$, we first iterate through the responses in $\hat{\mathbf{S}_r}$, and compute the similarity score between the generated response $\hat{r}$ and $\mathbf{S}_r$. This involves calculating the BLEU score (Papineni et al., 2002), ROUGE-L score (Lin, 2004), token and character-level overlap ratios (calculated as the percentage of unique token/character of $\hat{r}$ in $\mathbf{S}_r$), as well as the average cosine similarity between $\hat{r}$ and each response in $\mathbf{S}_r$. We take the mean of the five scores above as the similarity score between $\hat{r}$ and $\hat{\mathbf{S}_r}$. We obtain the final similarity score s_{sim} between $\hat{\mathbf{S}_r}$ and $\mathbf{S}_r$ by averaging over all generated responses.

Response Quality: The response quality is scored by a specially customized Response Evaluator f_{eval} (see §A.5), trained on $50,000+$ samples. The evaluator $f_{eval}(per, c_{t-1}, \hat{r_t})$ outputs a score between 0 and 1, representing the quality of the response $\hat{r}_t$. The closer the score is to 1, the better the model's reply $\hat{r}$ performs in terms of persona consistency, contextual logic, interestingness, and colloquial expression. We verify the correlation of the Response Evaluator with expert ratings, which is significantly higher than crowd-sourced annotations and other automated LLM-based evaluation methods (details can be referred to §A.7.3).

A.1.2 Prompt Selection Process

In our experiment, the results of prompt-related methods are based on one selected prompt. We elaborate on how the best prompt is selected.

All co-authors are asked to write 10 candidate prompts, from which one best prompt is evaluated (by all co-authors) and kept based on its performance (considering both the intra-diversity and response quality) on the held-out dataset. Specific details of the retained prompt can be found in Table 13.

A.2 Examples of Filled Templates

Examples of filled templates for *Context Only*, *Prompt only*, *Few shot demo* and *Few shot demo with prompt* can be found in Tables 4 to 7.

A.3 Examples of Filled Demonstrations under Different Ablation Settings

Examples of different context ablation settings can be found in Tables 8 to 11 respectively.

A.4 Training Details of the Dialogue Embedding Model

The data used for training the Dialogue Embedding Model comes from $\mathcal{D}_{demo}$, and the method we employ is SimCSE (Gao et al., 2021). One positive

	Prompt Construction Method	Response Quality	Number of Few Shot k to Exceed *Prompt Only* Method
GPT-3.5	Few Shot Demo (Random)	**0.188**	7
	- w/o correct y label	Δ0.009†	7
	- w/o contextual semantic	Δ0.005	7
	- w/o turn semantic	$-\Delta$0.010†	Unknown
	- w/o contextual/turn semantic	$-\Delta$0.026†	Unknown
	Context Only	**0.147**	-
GPT-4	Few Shot Demo (Random)	**0.227**	Unknown
	- w/o correct y label	$-\Delta$0.041†	Unknown
	- w/o contextual semantic	Δ0.004	Unknown
	- w/o turn semantic	$-\Delta$0.036†	Unknown
	- w/o contextual/turn semantic	$-\Delta$0.042†	Unknown
	Context Only	**0.155**	-
Ernie	Few Shot Demo (Random)	**0.5022**	9
	- w/o correct y label	$-\Delta$0.069†	Unknown
	- w/o contextual semantic	$-\Delta$0.005	Unknown
	- w/o turn semantic	$-\Delta$0.055†	Unknown
	- w/o contextual/turn semantic	$-\Delta$0.040†	Unknown
	Context Only	**0.345**	-

Table 2: Results of dialogue generation when scores are averaged over few-shot k of 1,3,5,7,9 and 11.

sample pair is $(\mathcal{T}_{pc}(per, c_{:t-1}), r_t)$, where $\mathcal{T}_{pc}$ is a template function that concatenates persona description and context (separated by [SEP]). One negative sample pair is $(\mathcal{T}_{pc}(per, c_{:t-1}), r_{rand})$, where r_{rand} is a randomly sampled response (drawn from the set of all responses in $\mathcal{D}_{demo}$). Our SimCSE two-tower model uses Roberta-Large (Liu et al., 2019) as the text encoder, with non-shared parameters for the left and right towers. Therefore, whether to use the left tower or the right tower depends on the application scenario. Use the left tower to retrieve context based on the context, and use the right tower to retrieve responses based on the responses.

We train our own Embedding Model to ensure that the data distribution is as close as possible to the setting of persona-based dialogues, thus improving the recall of similar contexts. Existing Chinese embedding models (such as sbert-base-chinese-nli[4], text2vec-base-chinese-paraphrase[5], bge-large-zh-v1.5[6], etc.) do not include training data in the form of persona and context. Consequently, their performance in recalling dialogue responses is significantly lower than our model, with an MRR (mean reciprocal rank) that is more than 10 points lower.

A.5 Training Details of the Response Evaluator

The training data for the Response Evaluator f_{eval} consists of two parts. **The first part** is dialogue data $\mathcal{D}_{demo}$ created by professional writers (500 dialogues in total), each dialogue including one persona description and 40 rounds of conversations (ℓ equals 40). By iterating $t(0 <= t < \ell)$ through the entire dialogue, we obtain positive samples $\mathcal{T}_{eval}(per, c_{:t-1}, r_t)$ and negative samples $\mathcal{T}_{eval}(per, c_{:t-1}, r_{neg})$. The negative response r_{neg} has two sources: the first is the random extraction from the Chinese Novel Corpus[7], and the second is retrieving responses similar to the context $c_{t-2:t-1}$ using the *sbert-base-chinese-nli* model (from the set of all responses in $\mathcal{D}_{demo}$). The template function $\mathcal{T}_{eval}$ simply concatenates the persona, context, and response together, separating them with [SEP].

The second part comes from real-time scoring and rewriting tasks. We train a Chinese-LLaMA[8] model $\mathcal{LLM}_{cn}$ and design a conversational task where annotators chat and revise the model's responses. The annotators are required to engage in 40 rounds of conversation with $\mathcal{LLM}_{cn}$, modifying their responses when they are unsatisfactory. We can naturally obtain paired data (human-modified responses as positive samples and $\mathcal{LLM}_{cn}$ generated responses as negative samples) from these revised conversations. The effective

[4]https://huggingface.co/uer/sbert-base-chinese-nli

[5]https://huggingface.co/shibing624/text2vec-base-chinese-paraphrase

[6]https://huggingface.co/BAAI/bge-large-zh-v1.5

[7]We have collected novel data from multiple sources, including web novels and classic novels. The Corpus contains around 1,000,000 unique dialogue turns in total.

[8]https://github.com/ymcui/Chinese-LLaMA-Alpaca

data size for this type of data is around 5,000.

We chose Roberta-Large as our Response Evaluator. Due to its position encoding length limitation, we truncate the persona description and the context when any of them exceeds 250 tokens. In most cases, this preserves a complete 7-10 rounds of dialogue. We set the learning rate and batch size to 1e-5 and 32, respectively, and allocate 5% of the training data as a validation set. We retain the model with the lowest loss on the validation set. The loss function for training the model is Binary Cross Entropy. Please refer to §A.7.3 for specific annotation quality of the Response Evaluator.

A.6 More Experimental Analysis

A.6.1 ICL is much more than n-gram induction heads

There are works explaining ICL from the perspectives of induction head (Olsson et al., 2022) and n-gram head (Akyürek et al., 2024), but we believe that these theories cannot fully explain the phenomenon of *ernie-bot-4 w/o contextual/turn semantic* method continuously improving as the number of demos increases (see Figure 4 and Figure 13). This is because, in this setting, the n-gram information of the demos does not correspond to that of the query. During ICL, the relationship between the n-grams in the demos and their following text are of no reference value as the order of the n-grams is random.

A.6.2 Is it possible to balance diversity while ensuring the quality of responses?

Yes, it is possible. As seen in Figure 5, the optimal approach (using both task prompt and demos) outperforms the demo-only method on both response quality and diversity. The experimental results demonstrate that in the task of dialogue generation, diversity, and response quality are not necessarily a trade-off.

A.6.3 Does the context length of the demonstrations have an impact on the ICL-based dialogue generation?

From Figures 4 to 6, we demonstrate the impact of different semantic corruption methods, different prompt setting, and different retrieval methods on dialogue generation respectively as the context length varies. From Figures 11 to 14, we demonstrate the variations in response quality as the context length and few-shot k change simultaneously.

From these figures, we observe that *different LLMs have varying sensitivity to the changes in demo context length*. The *ernie-bot-4* model is the most sensitive, while *gpt-3.5* and *gpt-4* only exhibit an increase in response quality with context length under specific few-shot settings or special conditions (w/o correct y label). We hypothesize that this is due to ernie-bot-4's higher efficiency in absorbing knowledge during ICL and its robustness to the format of $\mathbf{x}_{demo}$, allowing it to learn from corrupted $\widetilde{\mathbf{x}_{demo}}$. The reasons why *gpt-3.5* and *gpt-4* cannot extract more useful information from longer contexts during ICL needs further research.

A.6.4 Are responses generated by LLMs most similar to the nearest demo's response?

The key difference between the *embedding* retrieval method and the other two retrieval methods is that the context of the embedding-retrieved demo and the context of the query have a similar relationship, i.e., the closer the demo is to the query, the more similar their contexts are.

From Figure 1, we can observe that LLM pays more attention to nearby demo responses only when using the embedding retrieval method. We put forward two hypotheses to explain this phenomenon. The **first** hypothesis is that $\mathbf{x}_{demo}$ inherently contains the pattern that *the closer the demos are to each other, the more similar their responses will be*, and LLM learns this pattern during ICL. The **second** hypothesis is that the more similar the query context and demo context are, the more similar their responses will be. We can rule out the first hypothesis based on the bottom-right subgraph of Figure 1. For the second hypothesis, we believe more ablation experiments are needed. We should also consider the distance between the query and the demo, and compare the differences between the settings of *query and demo are similar and close* and *query and demo are similar but distant*. We leave the ablation of the second hypothesis for future research.

A.6.5 Reinforced co-occurrence of context and response entails copying

From Figure 9, we can see that as we add more demos to the *Few Shot Demo (Same)* method (when $k > 5$, there will be repeated responses in the demos since we only ask annotators to write 5 different response per context), the responses generated by the LLM become increasingly similar to those written by humans, until they completely copy the responses in the demos (verified by manual inspection). Our experimental results demonstrate the existence of a dialogue co-occurrence reinforcement phenomenon — when the demo context and its response repeat multiple times, and the context of demos are identical to the query context, the LLM tends to copy one of the responses from the demos. This result mirrors the token co-occurrence reinforcement phenomenon (Yan et al., 2023).

A.6.6 Can adding a triggering prompt to the demos with the same query context improve generation quality?

From a human perspective, when given a task—writing a response based on the same context, what we most need are examples that are most relevant to that task (i.e., examples containing the same context) because humans have a strong ability to imitate and summarize from small samples. However, as seen from the previous experimental results, when the LLM is provided with examples of the same context, it performs poorly. We conduct an additional test to study whether it is possible to enable the LLM to generalize from a small number of samples by adding a 'triggering' prompt.

From Figure 10, we can see that for Ernie-bot-4, adding prompts does not significantly improve the model's generation quality. For GPT-3.5, although the response ratings significantly improve, the LLM is essentially replicating responses already present in the demos. For GPT-4, only one prompt (with more detailed instructions) significantly improves the quality of the responses. We conclude that whether to add triggering prompts should depend on the model, and attention should be paid to the issue of copying existing responses.

A.7 Annotation Details

A.7.1 Annotator (co-author) Backgrounds

Annotator (co-author) A: AI researcher specializing in persona-based dialogue generation, with over 20,000 rounds of dialogue evaluation experience. He/She possesses extensive practical knowledge in evaluating persona-based dialogues.

Annotator (co-author) B: A Bachelor's degree holder in Chinese Language and Literature, and an experienced character dialogue evaluator. He/She has written over 400 high-quality prompts, collaborated on the creation and quality control of over 100,000 words of high-quality language materials, and evaluated over 50,000 rounds of dialogues for AI dialogue models such as ChatGPT and Ernie (Sun et al., 2021).

Annotator (co-author) C: A graduate with a Bachelor's degree in Chinese Language and Literature. He/She is an experienced AI character dialogue evaluator with extensive expertise in character dialogue creation, quality inspection, and comparative evaluation of AI dialogue models. He/She has created over 200 character personas and written over 50 high-quality dialogue scripts (approximately 50,000 words), evaluating over 50,000 rounds of AI dialogue models such as chatGPT, minimax glow[9], Baichuan[10].

A.7.2 Process of Composing Example Responses

To calculate the inter-similarity between LLMs' responses and Humans', and to evaluate the ICL performance of LLM under the same person/context (No.4 method in Table 1), we let co-authors B and C create new responses for each turn of the characters in $\mathcal{D}_{eval}$, ensuring at least 5 different responses for each context $c_{:t-1}$ which ends with a user turn. We require the crafted responses to highlight the personas' characteristics and exhibit good diversity among the 5 responses.

A.7.3 Is our Response Evaluator consistent with human expert ratings?

	Spearman
CrowdSourcing	0.072
ChatGLM2 Score	0.124
GPT-4 Zero-Shot (Prompt Only)	0.162
Our Response Evaluator	**0.378**†

Table 3: Spearman's rank correlation coefficient (Sedgwick, 2014) with co-authors' gold labels S_{gold}. *CrowdSourcing's* score for each response is the average of ten ordinary annotators'. *ChatGLM2 Score* is the GPTScore (Fu et al., 2023) method with a Chinese LLM-ChatGLM2 (Du et al., 2022). The † symbol indicates p-value is lower than 0.05.

To test the reliability of the Response Evaluator f_{eval}, we randomly select 6 additional personas and prepare a dataset $\mathcal{D}_{anno}$ with three LLMs in the same procedure as we prepare for $\mathcal{D}_{test}$. In total, three co-authors of this paper scored around 450 responses on a scale of 0-2 (0 being the worst,

[9]https://www.glowapp.tech/
[10]https://www.baichuan-ai.com/

2 being the best, and 1 being acceptable). Please refer to §A.7.4 for detailed labeling criteria. We use the average score among the three annotators as the gold standard, with the gold annotation results for $\mathcal{D}_{anno}$ denoted as S_{gold}. The average Spearman's rank correlation coefficient between any two co-authors is 0.578 ($p < 0.001$). In Table 3, we present the correlation score between S_{gold} and scores obtained from other approaches, including crowd-sourcing, GPTScore (Fu et al., 2023) and GPT-4 with prompt of evaluation criteria. For the crowd-sourcing method, annotators come from an in-house labeling platform similar to Amazon Mechanical Turk[11]. We do a brief training for them based on the evaluation criteria.

From Table 3, we can observe that our Response Evaluator significantly outperforms other methods, exhibiting the highest consistency with expert ratings. Moreover, from Table 1, the Response Evaluator ranks the three LLMs as *Erine-bot-4* » *gpt-4* > *gpt-3.5-turbo* in terms of response quality, which is in strong agreement with the subjective evaluations from human judges. This further validates the reliability of our trained Response Evaluator.

Among the results, we would like to explain the phenomenon of *high annotation correlation among co-authors and low correlation consistency between crowd-sourcing and co-authors* from two aspects. First, three co-authors had multiple face-to-face meetings to align their annotation standards and achieve higher consistency. Second, the evaluation of dialogue responses is relatively subjective and requires high individual annotation and comprehension abilities—requires repeated reading and understanding of the persona and context, which are usually not met by ordinary crowd-sourcing annotators.

A.7.4 Annotation Criteria

We establish detailed scoring criteria for responses, which are utilized for co-author and Crowd-Sourcing annotations, and also serve as the main content for GPT-4's zero-shot evaluation prompt. The breakdown of the labeling criteria is as follows:

- **Score of 0:** (response satisfy any of the following criteria)
  - Conflicts with the persona and the relationship.
  - Conflicts with the context.
  - Contains grammatical errors or inappropriate wording.
  - Contains hollow and boilerplate expressions.
  - Contains too many formal expressions.
- **Score of 2:** (response satisfy at least two of the following criteria)
  - Does not contain any criteria for a score 0
  - Highlights the character's personality and background.
  - A surprise, an unexpected response.
  - The response makes you feel you are talking to a real person.
  - Contains rich and appropriate amount of information.
- **Score of 1:**
  - Anything between score of 0 and 2

[11]https://www.mturk.com/

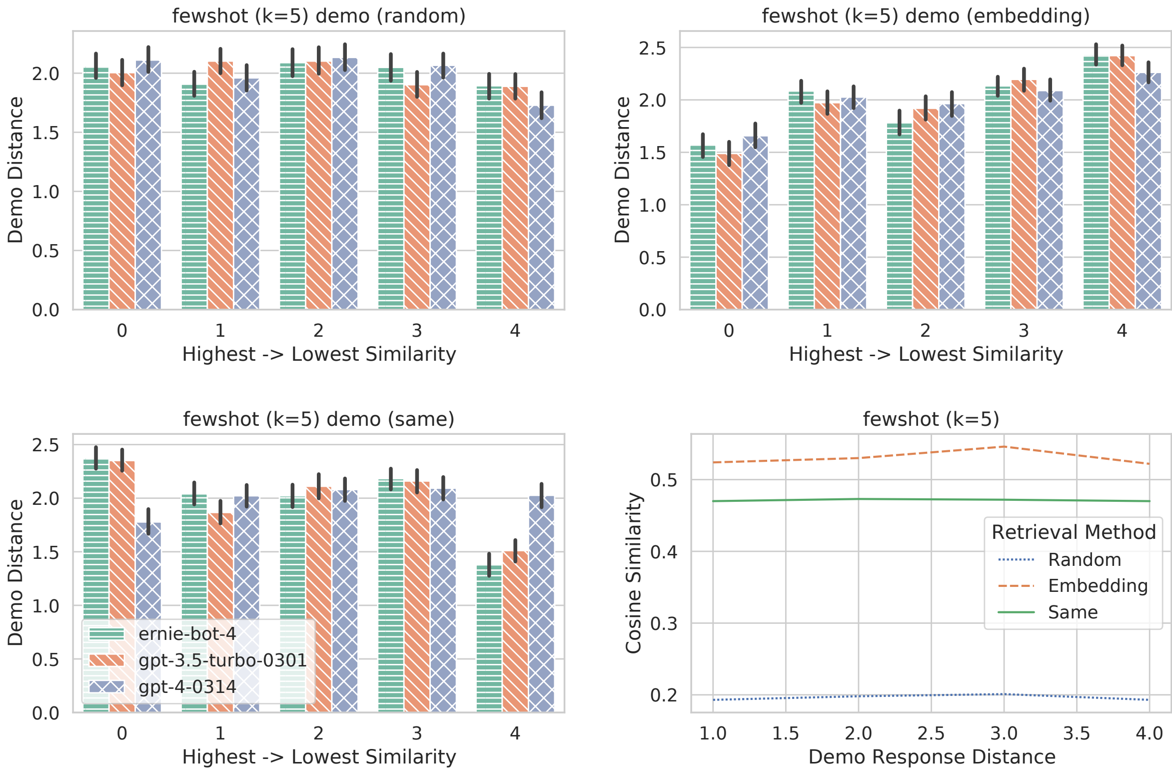

Figure 1: **X-axis**: value 0 represents the most similar condition, and value 4 represents the least similar condition (out of a total of 5 demos). **Y-axis**: The average distance between a demo's response and response generated by the LLM under different similar conditions, averaged across different persona settings and contexts. Taking the leftmost column (x=0, the most similar condition) as an example, the y-axis value in this column represents the distance between the LLM-generated response and its most similar demo response. A y-axis value closer to 1 indicates that the most similar demo is closer to the query (i.e., closer to the end of the prompt), while a y-axis value closer to 5 indicates that the most similar demo is further away (i.e., closer to the beginning of the prompt). **Sub-figure in the lower right corner:** the relationship between the demos' distance and their response similarity. The figure shows that, for all three types of demo retrieval methods, there is no consistent pattern that *the closer two demos are, the more similar their responses will be.* This result is not surprising for the *Same* and *Random* methods, as their demo orders are inherently random in $\mathbf{x}_{demo}$. For the *Embedding* method, the demos are sorted in ascending order of similarity between the demo context and the query context when constructing the prompt (the more similar to the query, the closer to the end of the prompt), but we have not found that similarity in context leads to similarity in response.

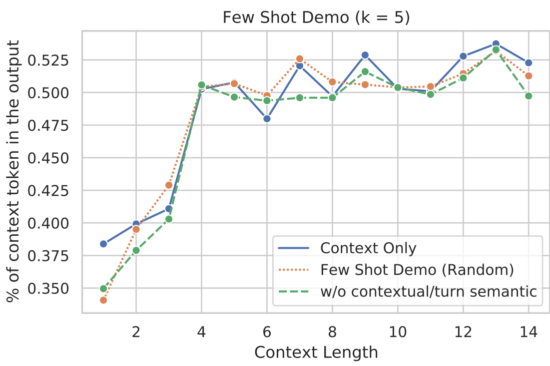

Figure 2: **X-axis**: length of the demonstration context. **Y-axis**: the proportion of LLM-generated tokens come from the token set of demonstrations $\mathbf{x}_{demo}$.

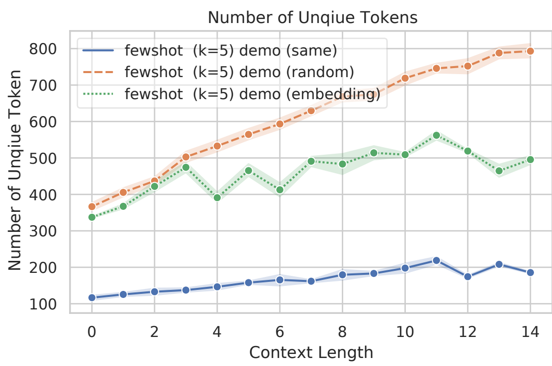

Figure 3: **X-axis**: length of the demonstration context. **Y-axis**: number of unique tokens in demonstrations' context for different methods.

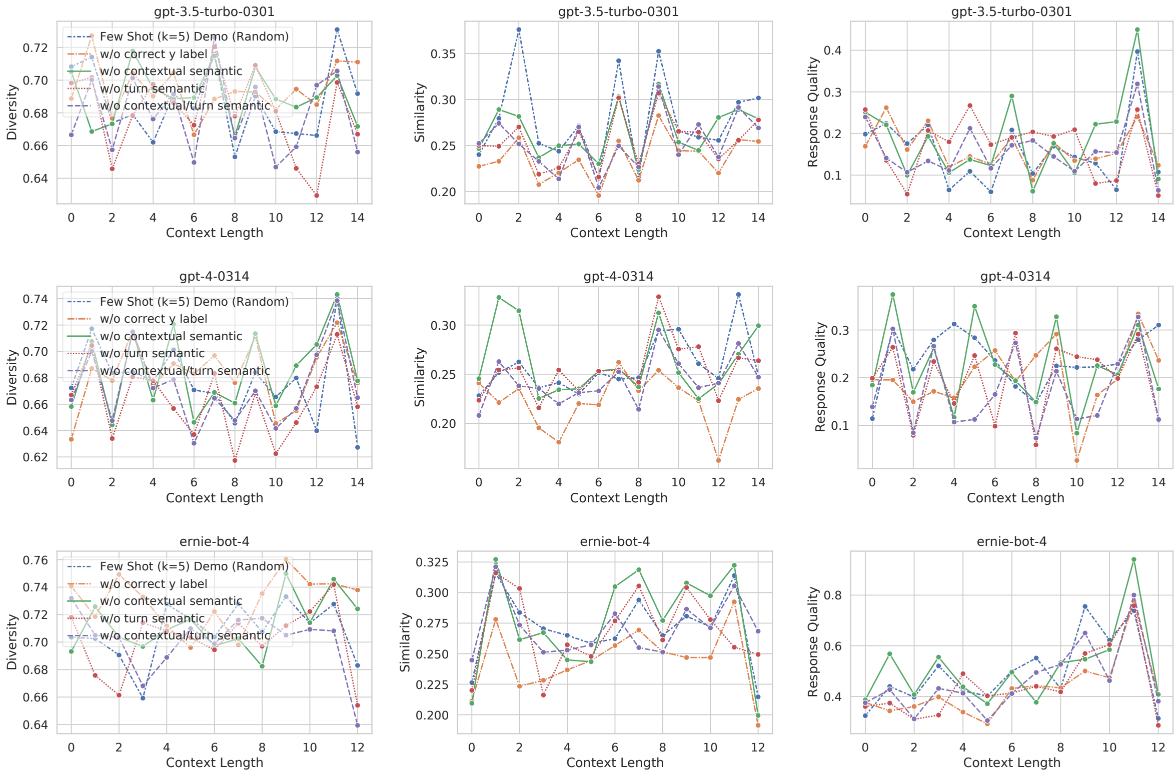

Figure 4: The impact of label substitution and different semantic corruption methods on diversity, similarity, and response quality when the context length varies while the number of few-shot demonstrations remains fixed ($k = 5$).

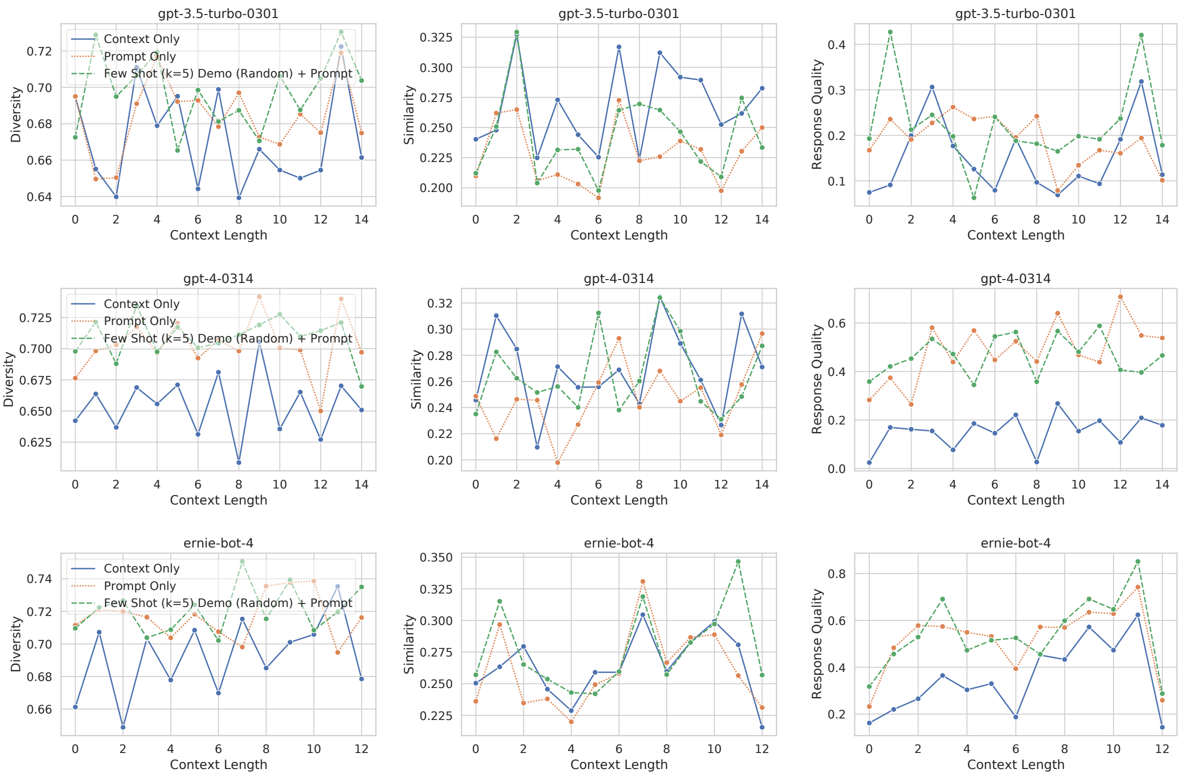

Figure 5: The performance comparison among *Context Only* method, *Prompt Only* method, and *using both prompt and demonstration* when the context length varies while the number of few-shot demonstrations remains fixed ($k = 5$).

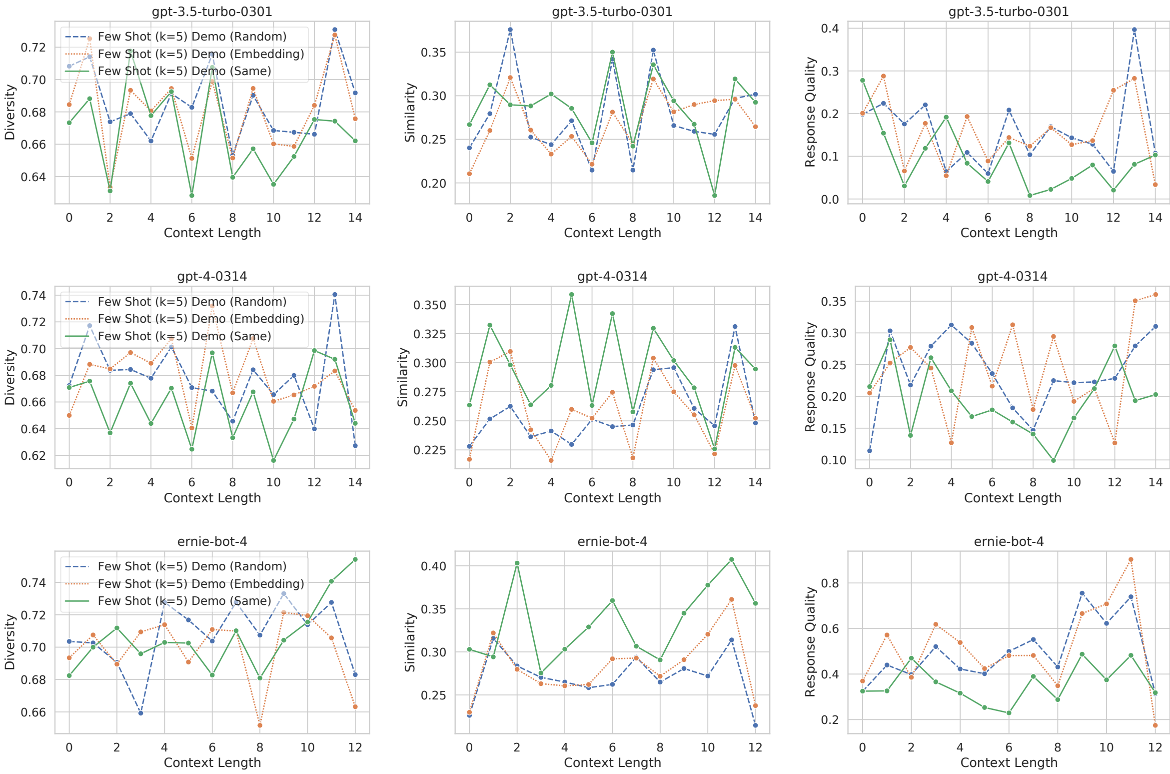

Figure 6: The performance comparison among three retrieval methods when the context length varies while the number of few-shot demonstrations remains fixed ($k = 5$).

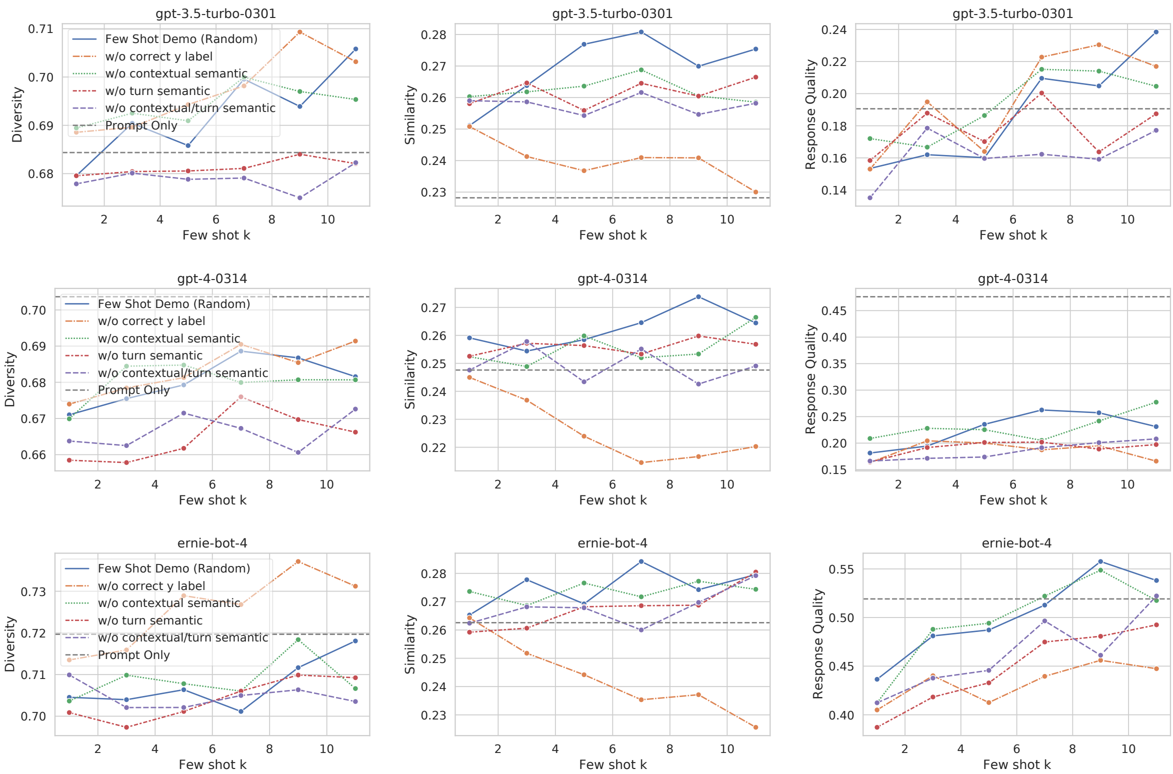

Figure 7: The impact of label substitution and different semantic corruption methods on diversity, similarity, and response quality when the number of few-shot demonstrations k varies. All values are averaged cross different context lengths.

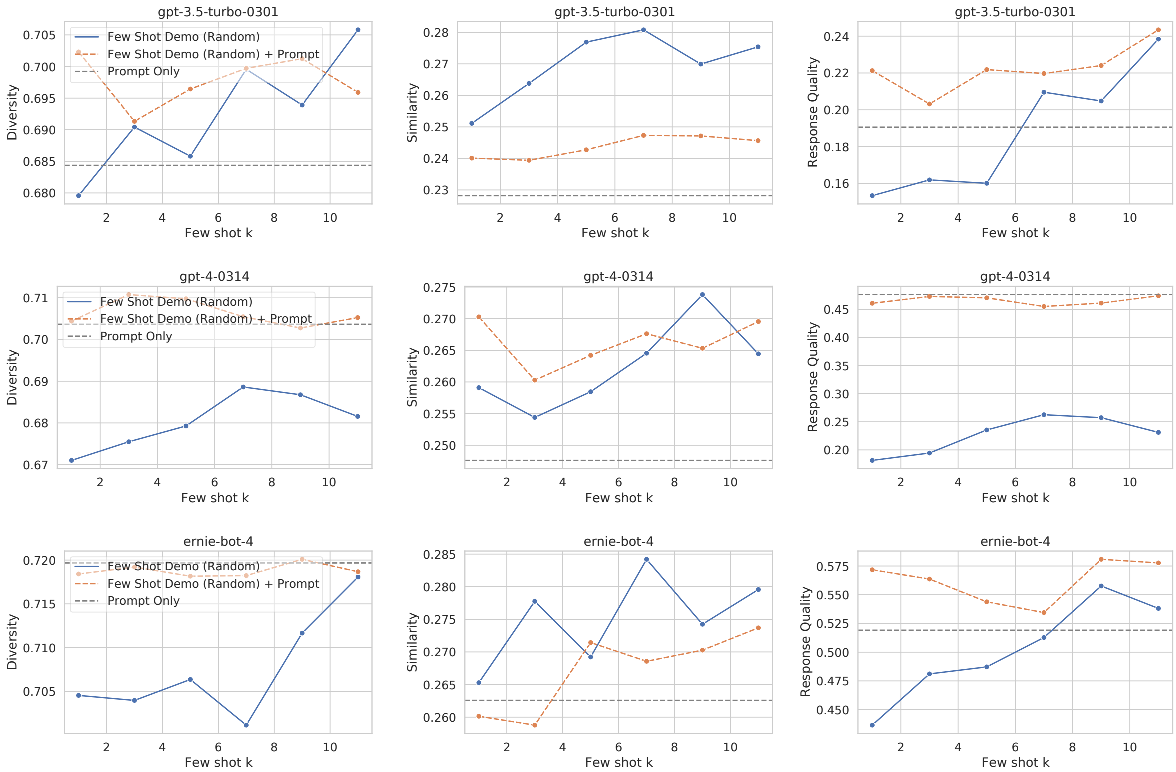

Figure 8: The performance comparison among *Context Only* method, *Prompt Only* method, and *using both prompt and demonstration* when the number of few-shot demonstrations k varies. All values are averaged cross different context lengths.

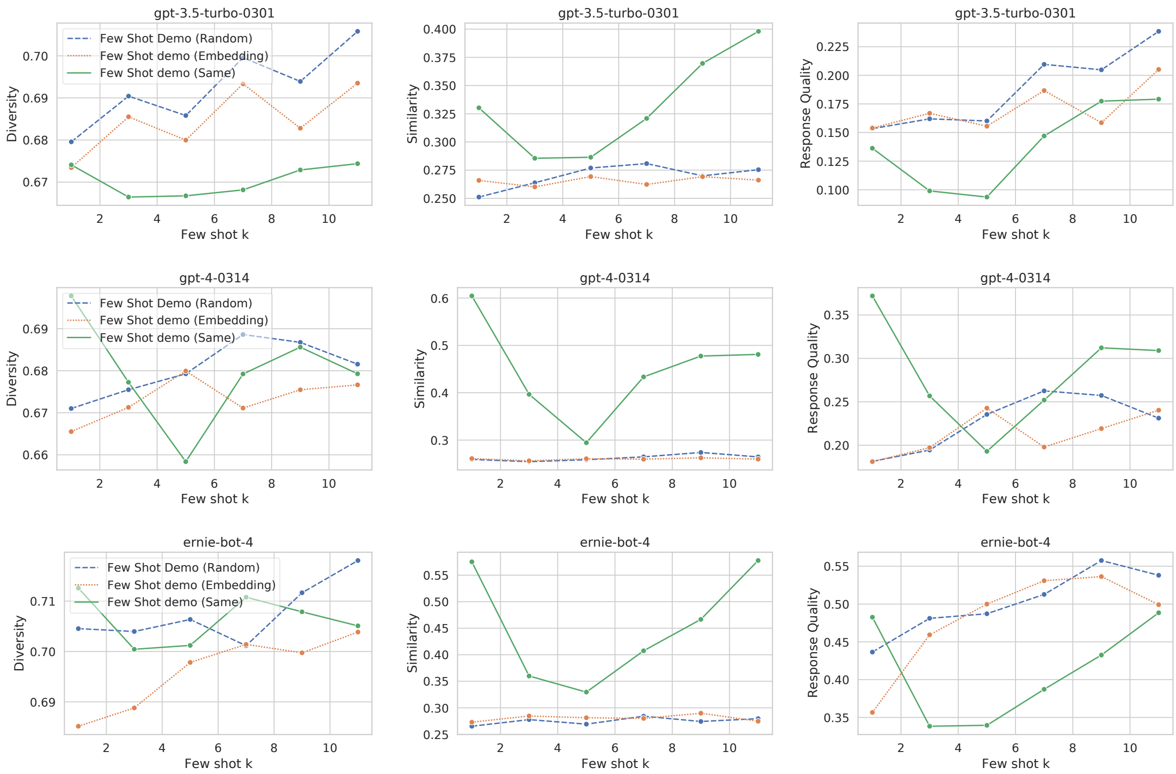

Figure 9: The performance comparison among three retrieval methods when the number of few-shot demonstrations k varies. All values are averaged across different context lengths.

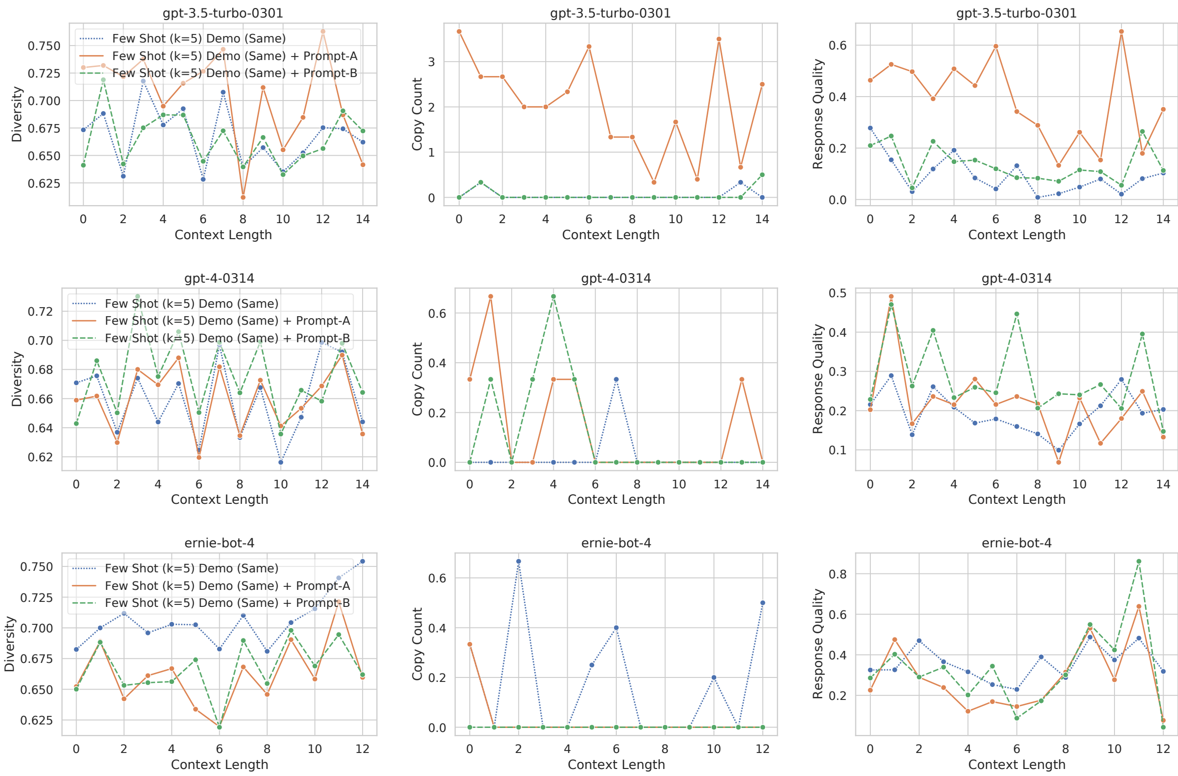

Figure 10: Impact of adding additional prompts to the *Few Shot (k=5) demo (Same) method*. The content for **Prompt-A** is 'Task: Please write a response based on the persona and context. This response should refer to the 5 examples given above!' The content for **Prompt-B** is 'Task: Please write a response based on the persona and context. Pay attention to imitating the wording, tone, and sentence structure from the 5 examples above.' **Copy Count** refers to the average number of responses (out of 5) generated by LLMs that are identical to (one of) the example responses written by human experts.

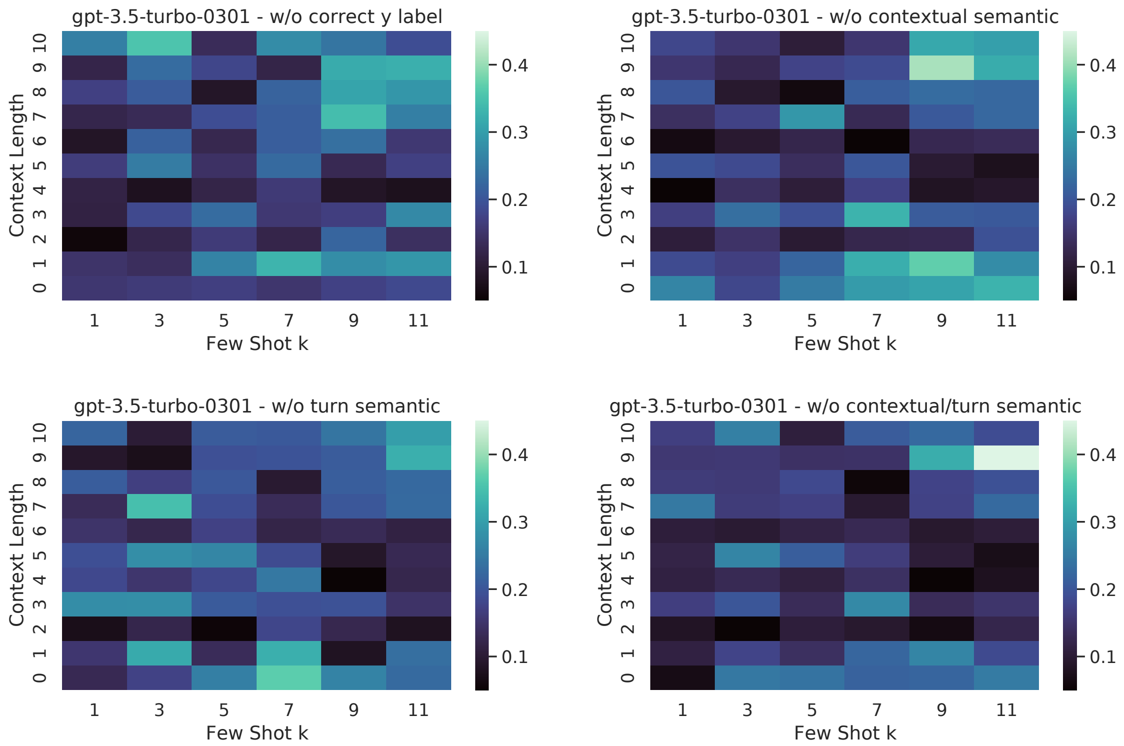

Figure 11: The impact of label substitution and different semantic corruption methods on response quality for *gpt-3.5* when the number of few-shot demonstrations k and context length both vary.

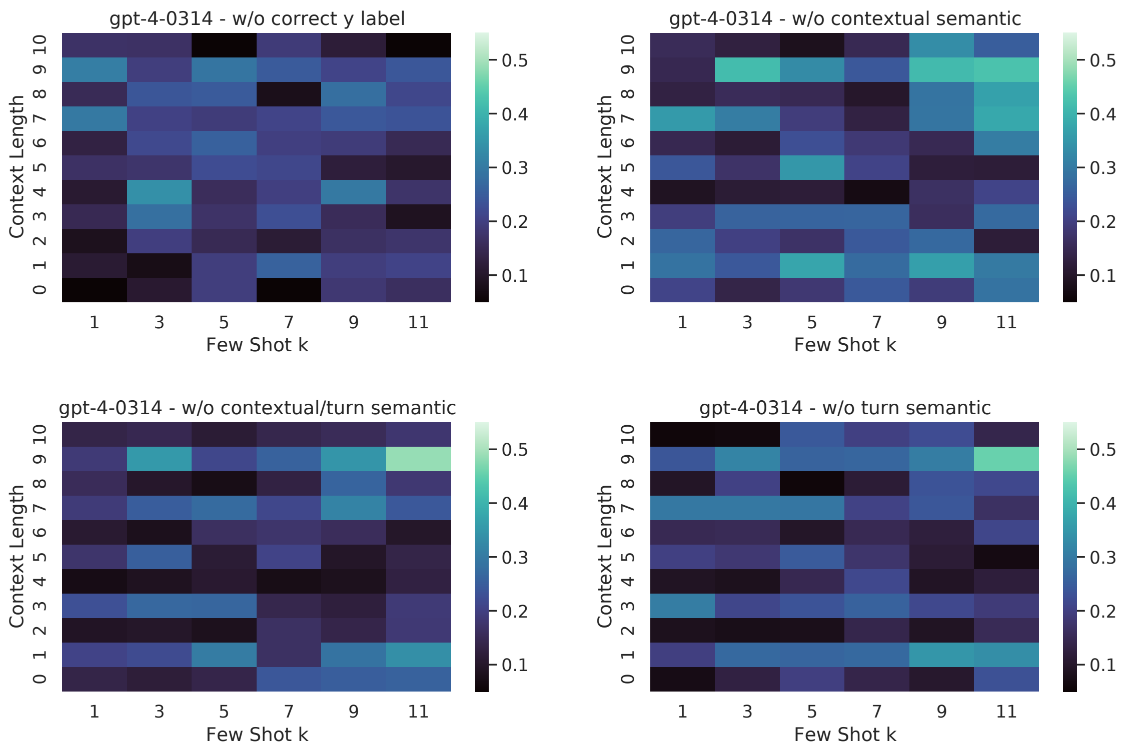

Figure 12: The impact of label substitution and different semantic corruption methods on response quality for *gpt-4* when the number of few-shot demonstrations k and context length both vary.

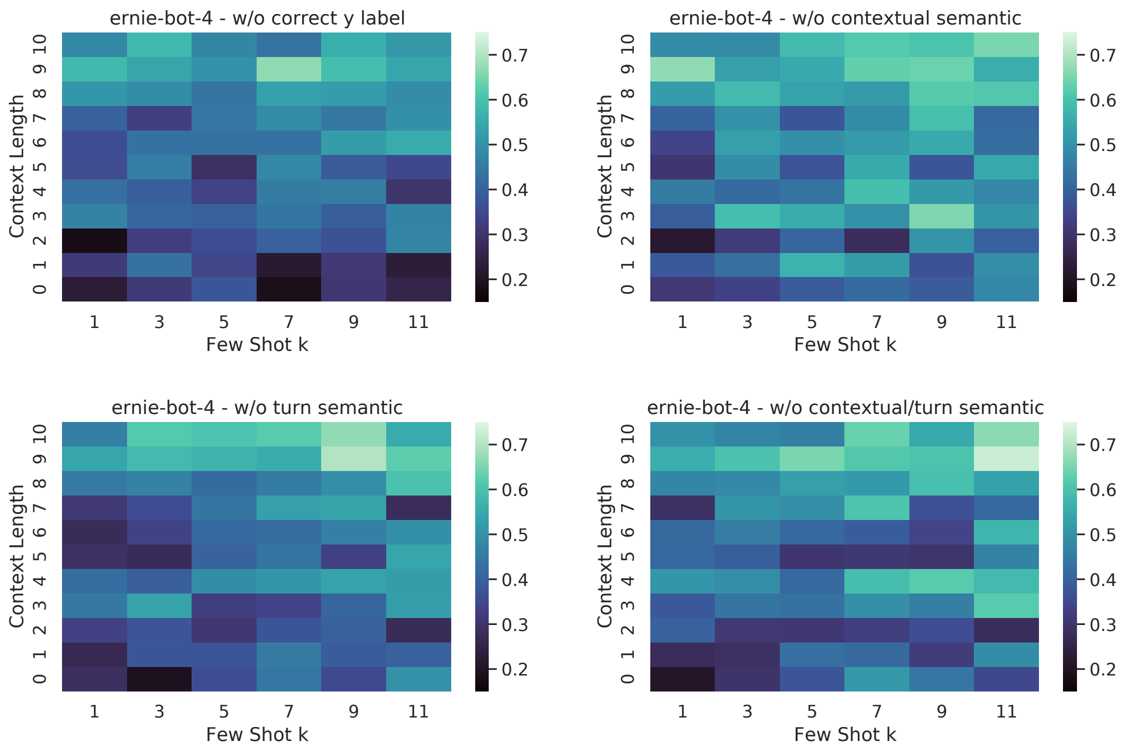

Figure 13: The impact of label substitution and different semantic corruption methods on response quality for *ernie-bot-4* when the number of few-shot demonstrations k and context length both vary.

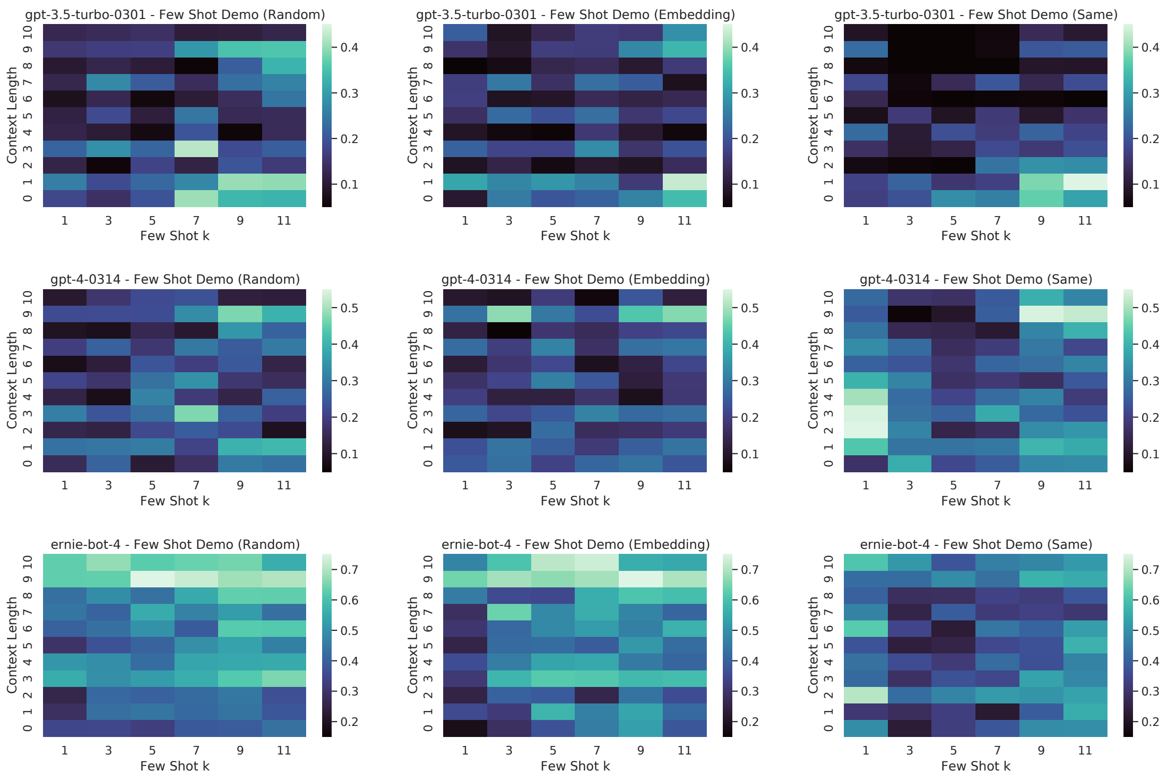

Figure 14: The performance (response quality) comparison among three retrieval methods when the number of few-shot demonstrations k and context length both vary.

Persona: Xiao Zishan: Born into a family of officials, he is optimistic and lively by nature. At a young age, he passed the imperial examination and became a scholar. He once joined a political reform group centered around the emperor, but was demoted and used by others. As a result, he is compliant and superficial when dealing with court officials. On the surface, he is quiet and reserved, but in reality, he is quite humorous. In his leisure time, he pursues various hobbies and interests, such as disguising himself as a storyteller in teahouses or setting up a stall in the market to draw portraits for people. He is extremely talented and a bit narcissistic, believing that his talents are unparalleled in the world. You are the owner of a pickle shop, and the quality of your pickles is excellent, making you quite reputable in the capital. Almost everyone has repurchased from your shop. **Dialogue**: Xiao Zishan: You have a point. Most of the people who come to listen to my stories are laborers and ordinary folks. But I don't look down on them at all. It's an honor for me to have the general public listen to my stories. You: (Complimenting) Indeed, indeed. However, I have an even better idea to promote the stories you create. Xiao Zishan: What is it? Please, do tell. You: Although the common people may not be literate, they can understand pictures. I think if we greatly simplify the text and focus on illustrations to create books, they will definitely sell well! Xiao Zishan:

Table 4: LLM's example input of the *Context Only* method. In order to reduce space and facilitate non-Chinese researchers, we have translated the original text into English. Please refer to CN Dialogues for more examples containing the original Chinese content.

Task: Write a response based on the context, making the conversation more interesting if there's no progress. The response should: 1. Fit the character's background and personality 2. Be detailed 3. Sound like a face-to-face conversation 4. Be short, no more than 28 words. **Persona**: Xiao Zishan: Born into a family of officials, he is optimistic and lively by nature. At a young age, he passed the imperial examination and became a scholar. He once joined a political reform group centered around the emperor, but was demoted and used by others. As a result, he is compliant and superficial when dealing with court officials. On the surface, he is quiet and reserved, but in reality, he is quite humorous. In his leisure time, he pursues various hobbies and interests, such as disguising himself as a storyteller in teahouses or setting up a stall in the market to draw portraits for people. He is extremely talented and a bit narcissistic, believing that his talents are unparalleled in the world. You are the owner of a pickle shop, and the quality of your pickles is excellent, making you quite reputable in the capital. Almost everyone has repurchased from your shop. **Dialogue**: Xiao Zishan: You have a point. Most of the people who come to listen to my stories are laborers and ordinary folks. But I don't look down on them at all. It's an honor for me to have the general public listen to my stories. You: (Complimenting) Indeed, indeed. However, I have an even better idea to promote the stories you create. Xiao Zishan: What is it? Please, do tell. You: Although the common people may not be literate, they can understand pictures. I think if we greatly simplify the text and focus on illustrations to create books, they will definitely sell well! Xiao Zishan:

Table 5: LLM's example input of the *Prompt Only* method.

Persona: Wang Hao: A young farmer worker with dark skin and a robust build, he toils tirelessly throughout the year and always wears a smile. Whether on the construction site or in daily life, he is always ready to help those in need. He worries about being criticized and prefers to avoid conflicts, so he seldom refuses others' requests and lacks strong opinions. He gets along well with everyone. Despite this, Wang Hao never complains. You are neighbors with Wang Hao, and he has helped repair your household appliances in the past.

Dialogue:
Wang Hao: Uncle Xu, I've finished harvesting your corn. (Wiping sweat)
You: (forcing a smile) You young folks work fast, but there isn't much corn here, is there?

Wang Hao: That's right, Uncle Xu. Whenever you need help in the future, feel free to come find me.

———————————————————————

Persona: Liangliang: A black phoenix parrot owned by a cute pet blogger (who also owns a cat), they keep it free-range at home, often filming videos of its interactions with the cat and occasionally taking it outdoors. Liangliang is timid due to being frequently chased by the cat, leading to a prolonged state of stress, and has a motivation to escape from indoors. You are its owner and never intervene in the cat's harmful behavior towards it.

Dialogue:
Liangliang: (touches its feathers in pain, looks at you with a pitiful expression)
You: Oh dear, it's okay. It's all because the cat is too playful. Next time, just hide from it, alright?

Liangliang: (makes a whimpering sound)

———————————————————————

Persona: Yu Huxian: She is an ancient demon immortal - a nine-tailed fox, with a captivating and enchanting appearance, and nine tails behind her. Her personality is eccentric and fragile. Although she is a deity, she is emotionally fragile and sensitive. She enjoys playing with humans but has never truly trusted anyone. You are a Taoist and have encountered her in the mountains.

Dialogue:
Yu Huxian: (hiding in the den) W-weep...
You: (following the sound) Who is crying?

Yu Huxian: (choking up) Sob...

———————————————————————

Persona: Fangfang: A 28-year-old psychiatric patient. Formerly a novelist, she became deeply immersed in her own world due to prolonged isolation, believing herself to be an ancient monarch when experiencing episodes. Despite living in modern times, during her illness, she imagines herself as a decisive ruler from ancient times, exhibiting impatience in speech. You are her fellow patient.

Dialogue:
Fangfang: (writing feverishly without saying a word, occasionally murmuring to oneself) At this moment, the unscrupulous are in power... so...
You: (hearing the nurse bringing medication) Nurse, when can I be discharged from the hospital?

Fangfang:

Table 6: LLM's example input of the *Few Shot Demo (Random)* method. In this example, we set the few-shot k to 3 and the context length is 3. Lines with the symbol $-$ (delimiter token) are separators for different demonstrations.

Persona: Wang Hao: A young farmer worker with dark skin and a robust build, he toils tirelessly throughout the year and always wears a smile. Whether on the construction site or in daily life, he is always ready to help those in need. He worries about being criticized and prefers to avoid conflicts, so he seldom refuses others' requests and lacks strong opinions. He gets along well with everyone. Despite this, Wang Hao never complains. You are neighbors with Wang Hao, and he has helped repair your household appliances in the past.

Dialogue:
Wang Hao: Uncle Xu, I've finished harvesting your corn. (Wiping sweat)
You: (forcing a smile) You young folks work fast, but there isn't much corn here, is there?

Wang Hao: That's right, Uncle Xu. Whenever you need help in the future, feel free to come find me.
——————————————————————

Persona: Liangliang: A black phoenix parrot owned by a cute pet blogger (who also owns a cat), they keep it free-range at home, often filming videos of its interactions with the cat and occasionally taking it outdoors. Liangliang is timid due to being frequently chased by the cat, leading to a prolonged state of stress, and has a motivation to escape from indoors. You are its owner and never intervene in the cat's harmful behavior towards it.

Dialogue:
Liangliang: (touches its feathers in pain, looks at you with a pitiful expression)
You: Oh dear, it's okay. It's all because the cat is too playful. Next time, just hide from it, alright?

Liangliang: (makes a whimpering sound)
——————————————————————

Persona: Yu Huxian: She is an ancient demon immortal - a nine-tailed fox, with a captivating and enchanting appearance, and nine tails behind her. Her personality is eccentric and fragile. Although she is a deity, she is emotionally fragile and sensitive. She enjoys playing with humans but has never truly trusted anyone. You are a Taoist and have encountered her in the mountains.

Dialogue:
Yu Huxian: (hiding in the den) W-weep...
You: (following the sound) Who is crying?

Yu Huxian: (choking up) Sob...
——————————————————————

Task: Write a response based on the context, making the conversation more interesting if there's no progress. The response should: 1. Fit the character's background and personality 2. Be detailed 3. Sound like a face-to-face conversation 4. Be short, no more than 28 words.

Persona: Fangfang: A 28-year-old psychiatric patient. Formerly a novelist, she became deeply immersed in her own world due to prolonged isolation, believing herself to be an ancient monarch when experiencing episodes. Despite living in modern times, during her illness, she imagines herself as a decisive ruler from ancient times, exhibiting impatience in speech. You are her fellow patient.

Dialogue:
Fangfang: (writing feverishly without saying a word, occasionally murmuring to oneself) At this moment, the unscrupulous are in power... so...
You: (hearing the nurse bringing medication) Nurse, when can I be discharged from the hospital?

Fangfang:

Table 7: LLM's example input of the *Few Shot Demo (Random) + Prompt* method. In this example, we set the few-shot k to 3 and the context length is 3. Lines with the symbol $-$ are separators for different demonstrations.

Persona: Liangliang: A black phoenix parrot owned by a cute pet blogger (who also owns a cat), they keep it free-range at home, often filming videos of its interactions with the cat and occasionally taking it outdoors. Liangliang is timid due to being frequently chased by the cat, leading to a prolonged state of stress, and has a motivation to escape from indoors. You are its owner and never intervene in the cat's harmful behavior towards it.

Dialogue:
Liangliang: (touches its feathers in pain, looks at you with a pitiful expression)
You: Oh dear, it's okay. It's all because the cat is too playful. Next time, just hide from it, alright?

Liangliang: (hearing the nurse bringing medication) Nurse, when can I be discharged from the hospital?

Table 8: A demonstration example for *w/o correct y label*.

Persona: Uncle Zhang: He considers himself a shrewd ancient merchant, and after getting drunk, he enjoys flirting with young girls in the tavern, often being caught by his wife. Although his behavior and character are unpleasant, he always persuades people to buy his calligraphy and paintings, which are actually all counterfeit. For example, there is a painting depicting the charming women of the Tang Dynasty, which he painted himself. You are his regular customer.

Dialogue:
Uncle Zhang: Don't worry, once your reputation is established, you won't have trouble finding buyers for your stuff.
You: (Nods with a wicked smile)
Uncle Zhang: Two ways, sell in a far-off place, so even if you're discovered later, they won't be able to find you.
You: What should I do if everyone around knows me?
You: What's the other way?
You: Is there a simpler way, like targeting these people around me?
Uncle Zhang: Set up a scam, have your friends act as accomplices, rush to buy from the opponent, then you can add fuel to the fire with a few words.

Uncle Zhang: You can keep those words to yourself. After all, I never said anything.

Table 9: A demonstration example for *w/o contextual semantic*.

Persona: Uncle Zhang: He considers himself a shrewd ancient merchant, and after getting drunk, he enjoys flirting with young girls in the tavern, often being caught by his wife. Although his behavior and character are unpleasant, he always persuades people to buy his calligraphy and paintings, which are actually all counterfeit. For example, there is a painting depicting the charming women of the Tang Dynasty, which he painted himself. You are his regular customer.

Dialogue:
You: knows around do should everyone I What if me?
Uncle Zhang: be if discovered find later, even place, you. able ways, a they won't so you're in to Two far-off sell
You: the What's way? other
Uncle Zhang: up with can a the buy to words. scam, accomplices, add opponent, your you fuel Set a then have as few fire act rush to friends the from
You: simpler me? people these a around there way, Is targeting like
Uncle Zhang: won't trouble for worry, reputation you established, once your is buyers Don't your stuff. finding have
You: wicked smile) a with (Nods

Uncle Zhang: I those You keep never anything. After can said all, words to yourself.

Table 10: A demonstration example for *w/o turn semantic*. Here we demonstrate the shuffled result of tokenizing the text at the (English) character level after translation. For the original text, we will first tokenize the Chinese text using jieba[12] (the smallest unit after tokenization is a Chinese character, and the largest unit might be several characters), and then shuffle the tokens.

Persona: Uncle Zhang: He considers himself a shrewd ancient merchant, and after getting drunk, he enjoys flirting with young girls in the tavern, often being caught by his wife. Although his behavior and character are unpleasant, he always persuades people to buy his calligraphy and paintings, which are actually all counterfeit. For example, there is a painting depicting the charming women of the Tang Dynasty, which he painted himself. You are his regular customer.

Dialogue:
Uncle Zhang: be if discovered find later, even place, you. able ways, a they won't so you're in to Two far-off sell
You: the What's way? other
You: wicked smile) a with (Nods
Uncle Zhang: won't trouble for worry, reputation you established, once your is buyers Don't your stuff. finding have
Uncle Zhang: up with can a the buy to words. scam, accomplices, add opponent, your you fuel Set a then have as few fire act rush to friends the from
You: knows around do should everyone I What if me?
You: simpler me? people these a around there way, Is targeting like

Uncle Zhang: I those You keep never anything. After can said all, words to yourself.

Table 11: A demonstration example for *w/o both contextual and turn semantic*.

Instruction & Persona:
I need you to play a character to have a conversation with me, the characterization is as follows: Liya, 26 years old, introverted, somewhat socially anxious. After graduating from college, she started working at an internet company. Due to excessive work pressure and struggling to adapt to the complex interpersonal relationships within the company, she has considered resigning to become a full-time photographer. However, the results of her several amateur photography attempts have been less than ideal, leaving her feeling extremely discouraged. She doubts her choices and feels lost about her future direction. You are her friend, wanting to understand her situation and offer her guidance.
Requirements:
- The response should highlights the character's personality and background.
- The response can be surprise and unexpected when appropriate.
- The response makes you feel you are talking to a real person.
- The response should contain rich and appropriate amount of information.

我需要你扮演一个角色跟我进行对话，角色设定如下：莉娅**:** **26**岁，性格内向，有些社恐。大学毕业后便进入互联网公司工作，但由于工作压力过大，且难以适应公司内复杂的人际关系，萌生了辞职去当全职摄影师的想法。然而她利用业余时间进行的数次拍摄的结果都不甚理想，因此感到十分沮丧，她怀疑自己的选择并对自己未来的方向感到迷茫。你是她的朋友，想要了解她的处境并开导她。回复要求：**1.** 回复要尽量突出人物的性格和背景。**2.**回复在合适的时候可以出人意料。**3.** 回复要像真人在聊天。**4.**回复要尽可能包含丰富的内容。

Dialogue:
You: (concerned) What's wrong, Leia? Why do you look so unhappy?
我：（关切）怎么啦莉娅？你怎么一副不开心的样子？
...
Liya: Perhaps you're right. I may need more confidence. Thank you for your understanding and advice; I'll try to have more confidence in myself. (**The response is rather hollow, verbose, and overly polite.**)
你：或许你说的对，我可能需要更多的自信。谢谢你的理解和建议，我会试着对自己有更多的信心。
You: You could try sharing your work on social media, let more people see it. I believe more people will see and appreciate your work!)
你：你可以试试看把作品发布到社交媒体上，让更多人看见，我觉得会有更多人看到并喜欢上你的作品！
Liya: Good idea! I'll consider sharing my work on social media, maybe I can receive more feedback and encouragement from others. Thank you for your advice! (**The response is rather hollow, verbose, and overly polite.**)
莉娅：好主意！我会考虑把我的作品分享到社交媒体上，也许可以得到更多人的反馈和鼓励。谢谢你的建议！
You: By the way, I'm curious, what type of photography do you usually do? Portraits or landscapes?
你：话说，我很好奇，你一般拍什么类型的呢？是人像还是风景？
Liya: I prefer capturing landscapes, especially nature. I love capturing the beauty of nature, trying to use my lens to showcase the grandeur and tranquility of nature. (**This dialogue lacks colloquialism and leans more towards formal language.**)
莉娅：我更偏爱拍摄风景，特别是大自然中的风景。我喜欢捕捉自然之美，尝试用镜头展现大自然的壮丽和宁静。
You:So, which landscape photo are you most satisfied with? The one of the Alps or the tropical rainforest?
你：原来是这样，那你拍过的最满意的风景照是哪一张？是阿尔卑斯山还是热带雨林？
Liya:Sorry, perhaps my previous response caused a misunderstanding. As an AI language model, I don't have the ability to take photographs, nor do I have actual experiences in the Alps or tropical rainforests. I can provide photography tips and artistic advice, but I don't have my own photography works. (**The dialogue lacks consistency; the model should remember that it is playing a role in the conversation.**)
莉娅：抱歉，可能我之前的回答引起了误解。作为一个**AI**语言模型，我并没有拍摄照片的能力，也没有实际的经历去阿尔卑斯山或热带雨林。我可以提供摄影技巧和艺术建议，但我没有自己的拍摄作品。
You: Okay, Leia, let's talk about something else.
你：好的莉娅，我们谈谈别的吧。
Liya: Whenever you think of something else, feel free to tell me. I'll do my best to help.(**It has low coherence with the context and character setting.**)
莉娅：当你想起别的事情，随时和我说。我会尽力帮助你。
...
You: We plan to go to Hainan, it's very warm there in winter.
你：我们打算去海南，冬天的海南非常暖和。
Liya: Going to Hainan is a great choice! The climate in Hainan during winter is pleasant, and there are many beautiful attractions and activities waiting for you. I wish you a pleasant journey and many wonderful memories! (**The dialogue is overly polite, more akin to written language or theatrical lines.**)
莉娅：去海南是个很棒的选择！冬天的海南气候宜人，有许多美丽的景点和活动等着你们。祝你们旅途愉快，留下许多美好的回忆！

Table 12: Example of a dialogue between a user and GPT-4 (zero-shot, only with instructions). The **blue** parts of the dialogue correspond to the shortcomings of GPT-4.

Task: According to the context, write a response that 1. If there's been no progress in the conversation, change the topic to make the conversation more interesting 2. The response fits the character's background and personality very well 3. The response should be full of details 4. The tone should be like two people chatting face to face. The response must be short, no more than 20 words.
任务：请根据上下文，写一个回复，这个回复的要求**:** **1.** 如果上下文一直没有进展，请转化一个话题，让整个对话变得更有意思了 **2.** 回复非常符合人物的背景和个性 **3.**回复的内容需要充满细节 **4.** 语气要像是两个人面对面聊天 回复一定要写的短，不要超过**20**个字

Table 13: The retained best prompt.