Dissecting In-Context Learning of Translations in GPT-3

Anonymous ACL submission

Abstract

Most of the recent work in leveraging Large Language Models (LLMs) such as GPT-3 for Machine Translation (MT) has focused on selecting the few-shot samples for prompting. In this work, we try to better understand the role of demonstration attributes for the in-context learning of translations through perturbations of high-quality, in-domain demonstrations. We find that asymmetric perturbation of the source-target mappings yield vastly different results. We show that the perturbation of the source side has surprisingly little impact, while target perturbation can drastically reduce translation quality, suggesting that it is the output text distribution that provides the most important learning signal during in-context learning of translations. We propose a method named Zero-Shot-Context to add this signal automatically in Zero-Shot prompting. Our proposed method greatly improves upon the zero-shot translation performance of GPT-3, thereby making it competitive with few-shot prompted translations.

1 Introduction

Recent work has put into question the importance of the correctness of demonstrations for prompting in Large Language Models (LLMs) (Min et al., 2022). One key conjecture is that the latent zero-shot capabilities of LLMs might be considerably higher than their observed zero-shot capabilities for a range of tasks (Min et al., 2022; Kojima et al., 2022). One way to elicit higher zero-shot performance is to qualify the role of demonstration attributes towards task performance and then simulate such in-context learning signals in a zero-shot manner. However, realizing this goal hinges on explicitly dissecting the role of various demonstration attributes (format, inputs, outputs, input-output mapping) towards task performance within few-shot in-context learning. In this work, we explore these questions for the task of Machine Translation (MT). Our line of inquiry is orthogonal to finding the most useful samples for few shot learning, a topic that has received considerable attention for eliciting better translations from LLMs (Vilar et al., 2022; Agrawal et al., 2022). Our contributions are:

1. We explore the role of demonstration attributes within in-context learning of translations in the GPT family of LLMs, through perturbations of the input-output (source-target) mappings. We show that the target text distribution is the most important factor in demonstrations, while the source text distribution provides an inconsequential learning signal.
2. Based on our findings, we propose Zero-Shot-Context prompting, which tries to automatically provide the learning signal corresponding to the target text distribution without any source-target examples. This greatly improves GPT-3's zero-shot performance, even making it competitive with few-shot prompting.

2 Related Work

Our work is related to two key themes, namely prompting LLMs for translation and analysis of in-context learning in LLMs. In this section, we situate our work within these two themes.

LLM Prompting for MT: Most of the work for prompting in MT has focused on selecting the training or development instances to be used as examples during prompting. Vilar et al. (2022) experiment on PaLM (Chowdhery et al., 2022) and find that quality of examples is the most important factor in few-shot prompting performance. Agrawal et al. (2022) experiment with XGLM (Lin et al., 2021) and report that translation quality and the domain of the examples are consequential. Our work builds on these with a different aim, in that we do not explore selecting the examples, rather apply perturbations on high-quality, in-domain examples to better qualify the role of certain demonstration attributes for in-context learning of translations.

Ground Truth	Shuffled Targets	Jumbled Source	Jumbled Target	Reversed Target
English: A B C	English: A B C	English: B A C	English: A B C	English: A B C
German: D E F	German: X Y Z	German: D E F	German: E D F	German: F E D
English: U V W	English: U V W	English: U W V	English: U V W	English: U V W
German: X Y Z	German: D E F	German: X Y Z	German: Y Z X	German: Z Y X

Table 1: Perturbations Applied: The four types of perturbations (shown here as applied on abstract source-target example sequences) manipulate the demonstration attributes differently. For example, while Jumbled Source and Jumbled Target both corrupt the source-target mapping, they modify different learning signals in in-context learning.

Analyzing In-Context Learning: Theoretical and empirical investigation of in-context learning is an ongoing research endeavor (Xie et al., 2021; von Oswald et al., 2022; Akyürek et al., 2022; Dai et al., 2022). Min et al. (2022) demonstrate that label correctness in demonstrations is of limited importance for open-set classification tasks, while Yoo et al. (2022) show that negated labels do matter. Our experiments differ from these works both on the choice of the task (translation, which has an exponential output space) as well as on the types of perturbations applied to the demonstrations.

3 The Role of Demonstration Attributes

To produce outputs for a specific task, LLMs are typically prompted with demonstrations (input-output examples pertaining to the specific task) appended with the test input. Similar to Min et al. (2022), we posit that there exist four aspects of demonstrations of the translation task that provide a learning signal: the input-output mapping, the input text distribution, the output text distribution and the format. In this section, we conduct an empirical investigation on how LLMs such as GPT-3 leverage the demonstrations provided to them for the task of translation by perturbing the input-output (source-target) mappings provided during prompting. Through these experiments, we hope to compare the importance of three key demonstration attributes – the input text distribution, the output text distribution and their mapping for translation.

Models: In this section, we mainly report results for text-davinci-002 [1], one of the most capable LLM models publically accessible (Liang et al., 2022). We also investigate the veracity of our observations with text-davinci-001 and text-curie-001, two prior LLM versions in the GPT family as well as the more recent text-davinci-003.

[1]LLMs: https://beta.openai.com/docs/models/

Datasets: We experiment with the WMT'21 News Translation task datasets (Barrault et al., 2021), for the following four language pairs: English-German (En-De), German-English (De-En), English-Russian (En-Ru) and Russian-English (Ru-En). On each of these datasets text-davinci-002 achieves highly competitive performance with the WMT-21 winning NMT model (Tran et al., 2021), with eight demonstrations ($k = 8$ in k-shot prompting). We list the full test set performance with text-davinci-002 and text-davinci-003 for $k = 8$ in appendix A, while the perturbation experiments are reported on 100 random samples from the test sets in each case.

Prompt Details: Vilar et al. (2022) report than the choice of the format is inconsequential for few-shot prompting on the translation task. As such, we use the standard prompt used for MT in prior works, namely [Source]: ABC (\n) [Target]: DEF, where Source (e.g., English) and Target (e.g., German) represent the language names. Further, we use high-quality, in-domain sentence pairs sampled from the development set for few-shot prompting.

Evaluation: To minimize reference-bias in evaluation, which has been shown to be detrimental in estimating the LLM output quality in related sequence transduction tasks (Goyal et al., 2022), we make use of a state-of-the-art Quality Estimation (Fomicheva et al., 2020) metric named COMET-QE (Rei et al., 2020) for quality evaluation. Further, one caveat of using the reference-free metric is that it allocates high scores to a translation if it is in the same language as the source sentence, i.e. it doesn't penalize copy errors in translation. To mitigate this evaluation shortcoming, we use a language-id classifier (Joulin et al., 2017) and set the translation to empty if the translation is produced in the same language as the source.

Experiment 1: We apply four perturbations to the demonstrations used for prompting. Table 1

enumerates the four perturbations with abstract source-target sequences: Shuffled Targets (ST) randomizes the mappings between the source and targets in the prompt, Jumbled Source (JS) randomizes the position of the words in the source sentences, Jumbled Ref (JT) randomizes the positions of the words in the target sentences and Reversed Ref (RT) reverses the order of the words in the target sentence. Among these perturbations, ST impacts both the input and output spaces symmetrically, while the other perturbations (JS, JT and RT) perturb only one of the input/output spaces.

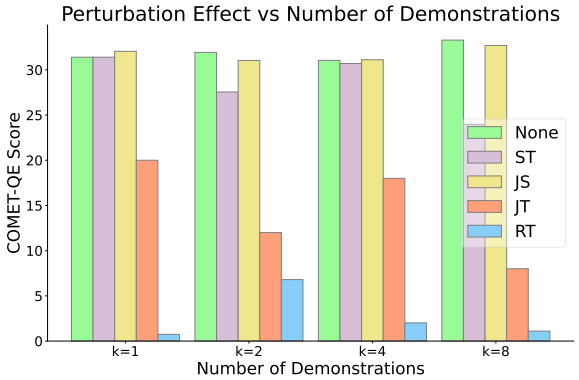

Figure 1: Perturbing the demonstrations for WMT-21 English-German test set. Source and Target perturbations have asymmetric effects despite the input-output mapping getting severely damaged in both cases.

Results: The results of applying these perturbations on En-De are presented in Figure 1, across different number of demonstrations ($k = 1, 2, 4, 8$). The results show that while ST and JT both significantly disrupt the source-target mappings in the demonstrations, they have greatly different impact. Translation quality declines by a large value for JT, an effect that becomes larger with increasing k, e.g., for JT perturbation at $k = 8$, the translation quality is considerably worse. On the other hand, JS produces very little to no effect on the quality of translations. Further, owing to the nature of the perturbation ST becomes more disruptive at higher values of k, while yielding no impact for $k = 1$.

Experiment 2: We repeat the same experiment as above (Experiment 1) with four different language pairs from WMT-21 and text-davinci-002.

Results: The results are reported in Figure 2. We find that the trends are similar to the first experiment (Figure 1). Across the language pairs, JS and JT have asymmetric impact on translation quality, showing that in each case the critical learning signal arrives from the target text distribution, while

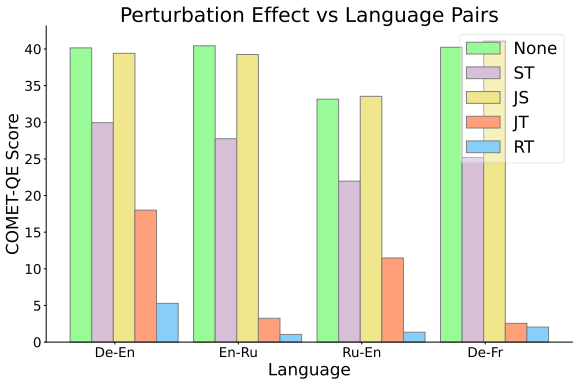

Figure 2: Perturbation effects across different WMT'21 language pairs for text-davinci-002, under few-shot prompting with k=8. The asymmetric effect of source and target perturbation holds true throughout the pairs.

the source text distribution is an inconsequential factor with respect to the output translation quality.

Experiment 3: We repeat Experiment 2, by keeping the language pair fixed to En-De and varying the LLMs. We report results in Figure 3 for three other models from the GPT family, namely text-curie-001, text-davinci-002 and text-davinci-003.

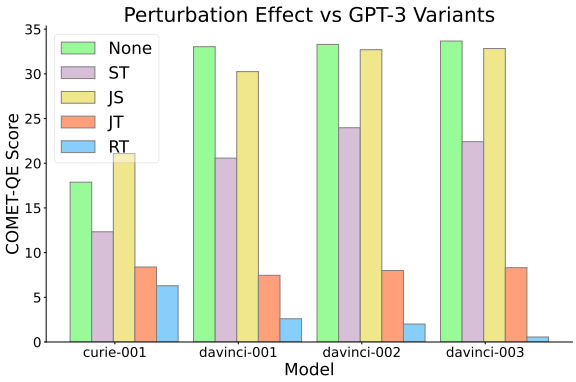

Figure 3: Perturbation effect across GPT-3 model variants for the WMT-21 English-German test set. The asymmetric effect of source and target perturbation holds across different models, suggesting that this is a stable trait of the in-context learning mechanism.

Results: We find that across different models, JS and JT have asymmetric impact on the translation quality, consistent with the prior two experiments.

Analysis: Compared to Min et al. (2022), wherein the randomization of the input-output mappings in the demonstrations leads to *better* performance than no demonstrations (zero-shot prompting) for open-set classification tasks, our results are quite different. We find that *depending on the type of perturbation*, in-context translation learning results can be vastly different *even when all the*

perturbations break the correct input-output mapping. For some perturbations (e.g., JT and RT) the translation quality is much worse than zero-shot. To reconcile these results, we hypothesize that the difference arises from the increased complexity of the auto-regressive search in the case of translation, i.e., a clear specification of the output space in the demonstrations becomes much more critical to constrain the search space.

Further, the ST results in Figures 2 & 3 show that source-target mapping is also a critical demonstration attribute, a fact consistent with prior results emphasizing the importance of example quality (Vilar et al., 2022; Agrawal et al., 2022). However, we show that it is not the primary learning signal in in-context learning of translations and even therein the source word order matters for little, suggesting that only an approximation of the input text distribution is sufficient for effective in-context learning.

4 Zero-Shot-Context for Translation

Previously, we demonstrated that the most important demonstration attribute for in-context learning of translations is the output text distribution. In this section, we present a method of providing this learning signal in a zero-shot manner. Our experiment here represents an inverse of experiments in section 3, i.e., here we *add a useful learning signal to zero-shot prompting*, rather removing learning signals from few-shot prompting to gauge their importance. We present a method named 'Zero-Shot-Context' and show that it greatly improves upon zero-shot performance for GPT-3, eliciting performance competitive even with few-shot prompting.

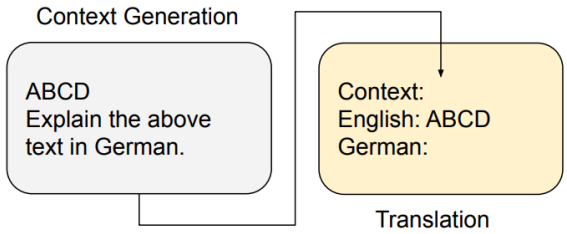

Figure 4: Schematic for Zero-Shot-Context: The Context Generation step provides an automatic learning signal to the LLM about the output text distribution, simulating the most important demonstration attribute.

Proposed Method: We propose a new zero-shot prompting method named Zero-Shot-Context (Figure 4), which auto-generates the output space specification learning signal from the LLM itself (the *Context*) and uses it to condition the translation.

Method	CQE↑	BLEU↑	ChrF↑	TER↓
Zero-Shot	*32.29*	22.6	54.3	71.4
Zero-Shot-Context	*37.65*	23.1	55.4	68.5
Few Shot (k=1)	39.92	22.4	54.1	71.8
Few Shot (k=2)	39.04	24.7	56.6	64.8
Few Shot (k=4)	40.36	24.0	55.7	65.4

Table 2: Zero-Shot-Context vs Baselines on WMT-21 En-De: Zero-Shot-Context greatly improves upon Zero-Shot Translations, gaining +5 QE points in quality.

Method	CQE↑	BLEU↑	ChrF↑	TER↓
Zero-Shot	*35.39*	19.8	49.4	74.3
Zero-Shot-Context	*40.67*	18.8	48.7	75.6
Few Shot (k=1)	37.92	20.5	50.1	72.3
Few Shot (k=2)	39.35	19.3	50.0	72.7
Few Shot (k=4)	39.25	20.2	50.1	72.3

Table 3: Zero-Shot-Context vs Baselines on WMT-21 En-Ru: Zero-Shot-Context greatly improves upon Zero-Shot and is even competitive with few-shot translations.

Experiment and Results: In Table 2 we compare Zero-Shot-Context with Zero-Shot prompting, as well as few-shot prompting (for k=1, 2, 4) with high-quality, in-domain examples sampled from the development set, on En-De WMT-21 test set with text-davinci-002. The results show that Zero-Shot-Context greatly improves upon Zero-Shot translation quality as measured by COMET-QE (CQE). Note that the gains are not visible in reference-based evaluation. Table 3 presents a comparison on the WMT-21 En-Ru test set. We present an ablation on Zero-Shot-Context in appendix B.

Further Analysis: Our findings indicate that the latent zero-shot GPT-3 performance for translations could indeed be higher than currently reported and that it is possible to leverage *direct computation* to improve LLM translation performance instead of manually retrieving or selecting examples.

5 Summary and Conclusions

We analyzed the relative importance of demonstration attributes as learning signals within few-shot in-context learning of translations in LLMs from the GPT family. We demonstrated that the critical learning signal arrives from the output text distribution, followed by the input-output mapping, while the input text distribution matters for little. We use this finding to propose Zero-Shot-Context, a method that tries to automatically generate the critical learning signal. Zero-Shot-Context greatly improves upon zero-shot translation quality in GPT-3, further validating our findings. We hope that our work could serve as a useful contribution towards better zero-shot utilization of LLMs for translation.

6 Limitations

Our work experiments with high-quality, in-domain examples for few-shot prompting. It is conceivable that perturbations could have different impacts on examples with varying quality. Also, while our proposed zero-shot method does not consume any manual examples, it suffers from the limitation that it involves two passes over a LLM. While this is partially mitigated by the method presented as an ablation in appendix B, we believe that a single pass method could also be derived by pre-computing the singular context once for the entire test set, a proposal we didn't investigate.

References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2022. In-context examples selection for machine translation.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. 2022. What learning algorithm is in-context learning? investigations with linear models.

Loic Barrault, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz, editors. 2021. *Proceedings of the Sixth Conference on Machine Translation*. Association for Computational Linguistics, Online.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Zhifang Sui, and Furu Wei. 2022. Why can gpt learn in-context? language models secretly perform gradient descent as meta optimizers. *arXiv preprint arXiv:2212.10559*.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners.

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

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2021. Few-shot learning with multilingual language models. *arXiv preprint arXiv:2112.10668*.

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*.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Chau Tran, Shruti Bhosale, James Cross, Philipp Koehn, Sergey Edunov, and Angela Fan. 2021. Facebook AI's WMT21 news translation task submission. In *Proceedings of the Sixth Conference on Machine Translation*, pages 205–215, Online. Association for Computational Linguistics.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2022. Prompting palm for translation: Assessing strategies and performance.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. 2022. Transformers learn in-context by gradient descent.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2021. An explanation of in-context learning as implicit bayesian inference.

Kang Min Yoo, Junyeob Kim, Hyuhng Joon Kim, Hyunsoo Cho, Hwiyeol Jo, Sang-Woo Lee, Sang-goo Lee, and Taeuk Kim. 2022. Ground-truth labels matter: A deeper look into input-label demonstrations. arXiv.

A WMT-21 Test Set Performance

In this section, we list the translation quality evaluations of text-davinci-002 and text-davinci-003, as compared to the WMT-21 winning system (Tran et al., 2021) to establish the competitiveness of the translations from the GPT-3 models, using the evaluation protocol in section 3. Both the translations from the WMT-21 winning system as welll as the GPT-3 translations were obtained through greedy decoding. The results are presented in Table 4.

Method	**En-De**	**De-En**	**Ru-En**	**En-Ru**
Facebook-WMT-21	39.36	39.88	35.25	46.41
davinci-002 (k=8)	39.57	40.28	35.67	39.06
davinci-003 (k=8)	40.31	41.31	36.03	41.82

Table 4: Translation Quality on WMT-21 Test Sets

B Ablation on Zero-Shot Context

We consider the following experiment: we pick a random target-side sentence from the development set and replace the Context-Generation step's output with the random target-side sentence. Ideally, an in-domain, high-quality target-side sentence should also be able to provide a learning signal regarding the output text distribution. We find that this is indeed the case, and simply replacing the context generation step with the random target-side sentence also improves upon zero-shot performance, achieving 36.10 COMET-QE score for WMT-21 En-De test set and 37.86 COMET-QE score for WMT-21 En-Ru. These scores are lower than Zero-Shot-Context, suggesting that the contextual nature of Zero-Shot-Context is also important.