Dissecting In-Context Learning of Translations

Vikas Raunak **Hany Hassan Awadalla** **Arul Menezes**
Microsoft Azure AI
{viraunak,hanyh,arulm}@microsoft.com

Abstract

Most of the recent work in leveraging Large Language Models (LLMs) such as GPT-3 for Machine Translation (MT) through in-context learning of translations has focused on selecting the few-shot demonstration samples. In this work, we characterize the robustness of LLMs from the GPT family to certain perturbations on few-shot translation demonstrations as a means to dissect the in-context learning of translations. In particular, 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. Further, 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. Based on our findings, 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

Large Language Models (LLMs) such as GPT-3 [Brown et al., 2020], PaLM [Chowdhery et al., 2022] or LLaMA [Touvron et al., 2023] have emerged as general-purpose, *foundation* models capable of addressing many natural language generation or understanding tasks [Bommasani et al., 2022]. For the task of Machine Translation (MT), large-scale evaluations have shown that LLMs from the GPT family match state-of-the-art performance on many language pairs through in-context learning of translations [Hendy et al., 2023]. At the same time, recent work has put into question the importance of the correctness of demonstrations for prompting in Large Language Models (LLMs) [Min et al., 2022]. As such, understanding *how* LLMs leverage the few-shot demonstrations during in-context learning of translations is quite pertinent both from scientific and practical perspectives.

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 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.

37th R0-FoMo: Workshop on Robustness of Few-shot and Zero-shot Learning in Foundation Models at NeurIPS 2023(NeurIPS 2023).

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 as adjudged by an state-of-the-art MT quality metric.

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.

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 towards in-context learning of translations.

Models: We report the results for a range of models in the GPT family of LLMs, ranging from text-davinci-002 [1], text-davinci-003 to gpt-3.5-turbo-instruct, which are among 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 gpt-3.5-turbo-instruct-0914.

Datasets: We experiment with the WMT'21 News Translation task datasets [Akhbardeh 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 $k = 8$, against the WMT21 winning NMT model in Table 1. The perturbation experiments are reported on 100 random samples from the test sets in each case.

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

Translation Model	**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 1: Translation Quality on WMT-21 Test Sets (COMET-QE)

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.

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 2: 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.

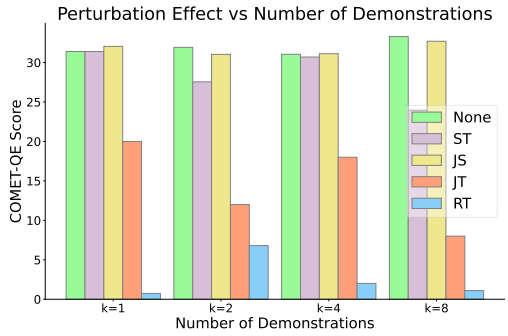

Figure 1: Perturbing the demonstrations for WMT-21 English-German test set. Source and Target perturbations have asymmetric effects despite the input-output mapping getting severely damaged in both cases.

Experiment 1: We apply four perturbations to the demonstrations used for prompting. Table 2 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.

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.

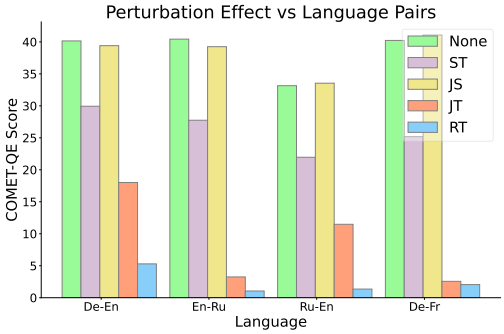

Figure 2: Perturbation effects across different WMT'21 language pairs for text-davinci-002, under few-shot prompting with *k*=8. The asymmetric effect of source and target perturbation holds true throughout the pairs.

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 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.

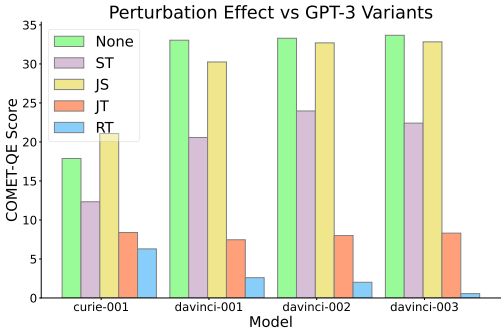

Figure 3: Perturbation effect across GPT-3 model variants for the WMT-21 English-German test set. The asymmetric effect of source and target perturbation holds across different models, suggesting that this is a stable trait of the in-context learning mechanism.

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.

Robustness of Our Findings: We also conduct experiments on gpt-3.5-turbo-instruct and gpt-3.5-turbo-instruct-0914, two of the more recent LLMs in the GPT family. With gpt-3.5-turbo-instruct on En-De, no perturbation (None in the plots) obtains a COMET-QE score of 34.21, the JS perturbation a score of 35.20 and the JT perturbation obtains a score of 25.45. Similarly, with gpt-3.5-turbo-instruct-0914 on En-De, no perturbation (None in the plots) obtains a COMET-QE score of 33.64, the JS perturbation a score of 34.35 and the JT perturbation obtains a score of 24.42. This observed behavior is agnostic to the choice of the MT quality metric as well: with COMET-KIWI (the state-of-the-art QE metric in the WMT-22 Quality Estimation Shared Task), no perturbation (None in the plots) with gpt-3.5-turbo-instruct obtains a score of 83.75, the JS perturbation a score of 83.94 and the JT perturbation obtains a score of 73.26. Similarly, with COMET-KIWI gpt-3.5-turbo-instruct-0914 obtains a score of 83.94, the JS perturbation a score of 83.85 and the JT perturbation obtains a score of 72.72. These results point to the robustness of our findings.

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.

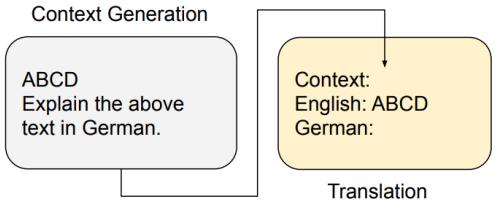

Figure 4: Schematic for Zero-Shot-Context: The Context Generation step provides an automatic learning signal to the LLM about the output text distribution, simulating the most important demonstration attribute.

Proposed Method: We propose a new zero-shot prompting method named Zero-Shot-Context (Figure 4) for obtaining translations, which auto-generates the learning signal pertaining to the output-space specification from the LLM itself (the *Context*) and uses it to condition the translation. The context in our proposed method signals the output space to the model and as such (as we demonstrate

later) it need not be tied to a specific formulation of the prompt. In practice, for the prompt described in Figure 4, we find that the generated context is sometimes a paraphrase or a continuation of the source text in the target language.

Prompting 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 3: 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.

Prompting 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 4: 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 3 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, which is concurrent with the existing literature which show that reference-based metrics such as BLEU or ChrF are inadequate to evaluate translations from LLMs [Garcia et al., 2023, Raunak et al., 2023]. Table 4 presents a similar comparison on the WMT-21 En-Ru test set. Next, we present an ablation on Zero-Shot-Context.

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. However, these scores are lower than Zero-Shot-Context, suggesting that the contextual nature of Zero-Shot-Context is also important.

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

In this work, 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 as well as for better understanding of in-context learning of translations.

References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. In-context examples selection for machine translation. 2022. URL https://arxiv.org/abs/2212.02437.

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.wmt-1.1.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. 2022. URL https://arxiv.org/abs/2211.15661.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models, 2022.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. 2022. URL https://arxiv.org/abs/2204.02311.

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

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555, 2020. doi: 10.1162/tacl_a_00330. URL https://www.aclweb.org/anthology/2020.tacl-1.35.

Xavier Garcia, Yamini Bansal, Colin Cherry, George Foster, Maxim Krikun, Fangxiaoyu Feng, Melvin Johnson, and Orhan Firat. The unreasonable effectiveness of few-shot learning for machine translation, 2023.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. News summarization and evaluation in the era of gpt-3. 2022. URL https://arxiv.org/abs/2209.12356.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. How good are gpt models at machine translation? a comprehensive evaluation, 2023.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 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, April 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-2068.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. 2022. URL https://arxiv.org/abs/2205.11916.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022. URL https://arxiv.org/abs/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. Few-shot learning with multilingual language models. *arXiv preprint arXiv:2112.10668*, 2021. URL https://arxiv.org/abs/2112.10668.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *EMNLP*, 2022. URL https://arxiv.org/abs/2202.12837.

Vikas Raunak, Arul Menezes, Matt Post, and Hany Hassan. Do GPTs produce less literal translations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1041–1050, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.90. URL https://aclanthology.org/2023.acl-short.90.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 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, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.213. URL https://aclanthology.org/2020.emnlp-main.213.

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

Chau Tran, Shruti Bhosale, James Cross, Philipp Koehn, Sergey Edunov, and Angela Fan. Facebook AI's WMT21 news translation task submission. In *Proceedings of the Sixth Conference on Machine Translation*, pages 205–215, Online, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.wmt-1.19.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. Prompting palm for translation: Assessing strategies and performance. 2022. URL https://arxiv.org/abs/2211.09102.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. 2022. URL https://arxiv.org/pdf/2212.07677.pdf.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. 2021. URL https://arxiv.org/abs/2111.02080.

Kang Min Yoo, Junyeob Kim, Hyuhng Joon Kim, Hyunsoo Cho, Hwiyeol Jo, Sang-Woo Lee, Sang-goo Lee, and Taeuk Kim. Ground-truth labels matter: A deeper look into input-label demonstrations. arXiv, 2022. URL https://arxiv.org/abs/2205.12685.