International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
IJARST
ISSN: 2457-0362

LANGUAGE IDENTIFICATION FOR MULTILINGUAL MACHINE TRANSLATION

[1]DR.N.SREEKANTH, [2]G.MADHU JYOTHI, [3]G.HARINI REDDY, [4]K.SAI NIKHITHA

[1]Assosiate Professor, Department of Electronics and Communication Engineering,**MALLA REDDY ENGINEERING COLLEGE FOR WOMEN,** Maisammaguda, Dhulapally Kompally, Medchal Rd, M, Secunderabad, Telangana.

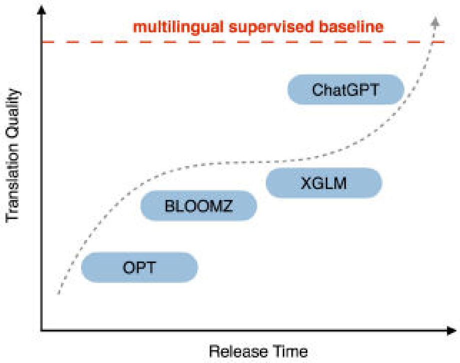
[2,3,4]Student, Department of Electronics and Communication Engineering,**MALLA REDDY ENGINEERING COLLEGE FOR WOMEN,** Maisammaguda, Dhulapally Kompally, Medchal Rd, M, Secunderabad, Telangana.

ABSTRACT

Large language models (LLMs) have shown significant potential in multilingual machine translation (MMT). This paper systematically explores the advantages and challenges of LLMs in this domain by addressing two key questions: 1) How effectively do LLMs translate a wide range of languages? 2) What factors influence their translation performance? We assess popular LLMs, including XGLM, OPT, BLOOMZ, and ChatGPT, across 102 languages. Our empirical findings reveal that even the top-performing model, ChatGPT, falls short of the supervised baseline NLLB in 83.33% of translation tasks. Further analysis uncovers new patterns in LLM behavior during MMT. First, the semantics of prompts can be largely disregarded when using in-context examples, as LLMs maintain strong performance even with nonsensical prompts. Second, cross-lingual examples often provide superior task guidance for low-resource translations compared to same-language pairs. Lastly, we find that BLOOMZ's performance on the Flores-101 dataset may be overestimated, highlighting potential risks associated with using public datasets for evaluation.

I.INTRODUCTION

As large language models (LLMs) grow in scale and complexity, they have developed a remarkable ability to perform a wide range of tasks through human-written instructions and in-context learning (ICL) (Brown et al., 2020). ICL enables these models to learn tasks by using a few provided examples as context. One area where LLMs have shown exceptional promise is machine translation (MT). Previous research has highlighted their surprising effectiveness in high-resource bilingual translations, such as English-German (Vilar et al., 2022; Zhang et al., 2022b), even when models like OPT are not specifically optimized for multilingual data. However, the multilingual translation capabilities of LLMs using ICL remain underexplored. Multilingual machine translation (MMT) is inherently challenging, involving text translation across numerous languages while requiring semantic alignment (Fan et al., 2021; Costa-jussà et al., 2022; Yuan et al., 2022). Furthermore, it is not yet clear which factors influence LLM performance in translation, as previous studies have largely focused on natural language understanding (NLU) tasks (Min et al., 2022; Kim et al., 2022; Zhang et al., 2022a; Wei et al., 2023a).



This paper investigates ICL and LLMs in the context of machine translation, addressing two primary questions: 1) How do LLMs perform in MMT across a wide array of languages? 2) What factors impact their translation performance? To answer the first question, we evaluate and compare prominent LLMs—including the English-centric OPT (Zhang et al., 2022b) and multilingual models like XGLM (Lin et al., 2021), BLOOMZ (Scao et al., 2022), and ChatGPT (OpenAI, 2022)—across 102 languages and 202 translation directions (both X-to-English and English-to-X). ChatGPT significantly outperforms other LLMs, especially in translations into English, addressing earlier concerns regarding LLMs' translation capabilities (Wei et al., 2022a; Chowdhery et al., 2022). However, compared to the widely used supervised baseline NLLB (Costa-jussà et al., 2022), ChatGPT achieves comparable performance in only 16.67% of translation directions. We identify three common types of errors in instances where LLMs struggle: off-target translation, hallucination, and monotonic translation.

For the second question, we discover several novel working patterns. Notably, LLMs can still perform translations effectively even when provided with unreasonable prompts, as long as appropriate in-context learning examples are given. However, mismatched translation pairs lead to failures, underscoring the importance of exemplars in ICL for machine translation. Surprisingly, random selection of exemplars serves as a strong baseline, while semantically selected exemplars yield only marginal improvements. Additionally, we find that cross-lingual translation pairs can serve as more effective exemplars for low-resource translations than same-language pairs. Lastly, we observe that BLOOMZ's results on the Flores-101 dataset may be overstated; our collected and human-annotated data show consistent performance drops in all human-evaluated directions, highlighting the need for closed-data evaluations to mitigate data leakage risks.

The main contributions of this paper are summarized as follows:

- We benchmark popular LLMs on MMT across 102 languages and 202 English-centric translation directions.
- We systematically report results for LLMs alongside two widely-used supervised baselines (NLLB and M2M-100).
- We identify new ICL patterns in LLMs relevant to machine translation.

II.BACKGROUND

1. Large Language Models

Language modeling has long been a core task in natural language processing, aimed at predicting the probability of the next token in a sequence (Bengio et al., 2000; Mikolov et al., 2010; Khandelwal et al., 2020). The Transformer architecture (Vaswani et al., 2017) serves as the backbone for contemporary LLMs.

LLMs demonstrate significant potential as universal multi-task learners. Radford et al. (2019) discovered that a causal decoderonly language model could function effectively as a multi-task learner with just unsupervised training data. Kaplan et al. (2020) further elucidated the scaling law of LLMs, showing that increasing the number of neural parameters and training data enhances model performance. Wei et al. (2022b) indicated that scaling language models also results in remarkable emergent abilities, such as ICL, which manifests predominantly in larger models. As a result, substantial efforts have been dedicated to scaling up LLMs (Brown et al., 2020; Scao et al., 2022; Vilar et al., 2022; Zeng et al., 2023; Ren et al., 2023). Among these, the GPT family and ChatGPT (OpenAI, 2022) stand out as notable systems, achieving impressive results across a variety of NLP tasks.

III.EXPERIMENT SETUP

1. Dataset

We benchmark multilingual translation using the Flores-101 dataset (Goyal et al., 2022), which facilitates an evaluation of model performance across a diverse range of languages, including low-resource ones. Our experiments focus on translation tasks between English and 101 other languages.

2. LLMs

We assess the translation capabilities of four prominent large language models (LLMs): two pre-trained models, XGLM-7.5B (Lin et al., 2021) and OPT-175B (Zhang et al., 2022b), as well as two instruction-tuned models, BLOOMZ-7.1B (Scao et al., 2022) and ChatGPT (OpenAI, 2022).

3. ICL Strategy

For each model, we evaluate translation performance using eight randomly selected translation pairs from the corresponding development set as in-context exemplars, along with the in-context template “<X>=<Y>.” Here, “<X>” and “<Y>” act as placeholders for the source and target sentences, respectively. We use line breaks as the concatenation symbol. Our analysis indicates that this ICL strategy is a simple yet effective approach. Implementation of ICL is based on OpenICL (Wu et al., 2023).

4. Supervised Baseline

We also report the performance of two widely used supervised models: M2M-100-12B (Fan et al., 2021) and the distillation version of NLLB-1.3B (Costa-jussà et al., 2022), serving as our baselines for many-to-many multilingual translation.

5. Metric

Following the methodology of Goyal et al. (2022), we utilize SentencePiece BLEU (spBLEU) as our evaluation metric. This metric employs a SentencePiece tokenizer (Kudo and Richardson, 2018) with a vocabulary of 256K tokens, allowing for comprehensive evaluation across all languages.

Benchmarking LLMs for Massively Multilingual Machine Translation

In this section, we present our findings on multilingual machine translation, focusing on the translation capabilities of large language models (LLMs).

ChatGPT is the Best Translator Among Evaluated LLMs

the evaluation results organized by language family. Detailed results for each translation direction can be found in Appendix A. Both XGLM and OPT demonstrate strong multilingual translation abilities, suggesting that alignment across multiple languages is achievable even with unsupervised data (Garcia et al., 2023). Instruction-tuned models like BLOOMZ and ChatGPT frequently outperform their pre-trained counterparts. Notably, BLOOMZ surpasses the supervised baseline in seven groups of translation directions, while ChatGPT achieves the highest average BLEU score across most evaluated directions.

LLMs Perform Better on Translating into English than from English

Previous research has shown that LLMs excel in translations into English but struggle more with translations from English (Wei et al., 2022a; Chowdhery et al., 2022). Our findings align with this observation for XGLM, BLOOMZ, and OPT. Interestingly, ChatGPT exhibits more balanced performance; however, it still struggles with translating from English to low-resource languages.

LLMs Lag Behind the Strong Supervised Baseline, Particularly for Low-Resource Languages

the translation performance of the supervised baseline (NLLB) compared to the best-performing LLM (ChatGPT) for each language. On the left side of the figure, ChatGPT shows comparable BLEU scores to NLLB. Conversely, on the right side, ChatGPT significantly underperforms NLLB, particularly for low-resource languages. Overall, ChatGPT falls short of NLLB in 83.33% of translation directions.

For the cases where ChatGPT struggles, we identify three common types of translation errors: 1) off-target translation, 2) hallucination, and 3) monotonic translation. Table 2 provides example cases for each error type, which also frequently occur in other LLMs.

Instruction-Tuned LLMs Can Still Benefit from In-Context Learning

While instruction-tuning enhances the performance of LLMs, they can further benefit from in-context learning strategies. This highlights the importance of leveraging context to improve translation accuracy across various tasks and languages.

IV.FINDINGS ON IN-CONTEXT EXEMPLARS

Semantically Selected Exemplars Do Not Provide More Benefits than Randomly Picked Exemplars

Selecting in-context exemplars is a critical step in implementing in-context learning. In this study, we utilize a development set for exemplar selection, which has proven to be a high-quality candidate pool (Vilar et al., 2022). We compare four methods for selecting in-context exemplars:

- Random: Exemplars are chosen at random.
- BM25: Exemplars are selected based on similarity to the test case's source sentence, evaluated using the BM25 algorithm.
- TopK: Exemplars are selected according to the similarity of sentence embeddings to the test case's source sentence.
- Oracle: Exemplars are chosen based on similarity of their target sentences to the test case's target sentence, serving as an upper bound for selection strategy.

The effects of varying the number of in-context exemplars across these different selection methods. Generally, as the number of exemplars increases from 1 to 8, BLEU scores rise rapidly. However, beyond this point, translation performance plateaus regardless of the selection strategy. When more exemplars are added (e.g., 32), BLEU scores often decline, contrasting with observations in natural language understanding tasks (Li et al., 2023).

Interestingly, randomly picked exemplars yield comparable translation performance to semantically selected ones, with oracle selection showing results similar to random selection. These findings suggest that while translation exemplars can guide LLMs, they may struggle to extract beneficial knowledge from semantically selected exemplars.

ICL Exemplars Teach LLM the Core Features of the Translation Task

To gain further insights into how in-context learning (ICL) exemplars impact LLMs' understanding of translation tasks, we analyze their behavior when exposed to abnormal in-context exemplars (Table 4).

When mismatched translations are used as exemplars, LLMs fail entirely, highlighting the model's reliance on context to maintain semantic consistency between source and target sentences. Additionally, using word-level or document-level translation exemplars negatively affects performance, indicating that the granularity of exemplars is crucial. An intriguing finding is that LLMs perform worse when given duplicated translations as exemplars, emphasizing the importance of diversity among in-context exemplars.

V.CONCLUSION

In this paper, we evaluated the multilingual translation capabilities of various large language models (LLMs), including ChatGPT, across 102 languages and 202 English-centric translation directions. Our findings highlight both the strengths and limitations of LLMs in multilingual machine translation (MMT). Notably, even the top-performing LLM, ChatGPT, falls short of the robust multilingual supervised baseline, NLLB, in 83.33% of translation directions.

Our analysis reveals new operational patterns in LLMs when applied to machine translation. For instance, we found that prompt semantics can often be disregarded during in-context learning, and cross-lingual exemplars can offer better task guidance for low-resource translations. Additionally, we noted that BLOOMZ exhibited overestimated performance on public datasets, underscoring the need for caution when evaluating LLMs.

VI.ACKNOWLEDGEMENT

We would like to express our gratitude to Fei Yuan and Zhenyu Wu for their support on this project. Shujian Huang is the corresponding author.

VII.REFERENCES

- Agrawal, S., Zhou, C., Lewis, M., Zettlemoyer, L., & Ghazvininejad, M. (2022). In-context examples selection for machine translation. arXiv preprint arXiv:2212.02437.
- Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., et al. (2023). A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. arXiv preprint arXiv:2302.04023.
- Bengio, Y., Ducharme, R., & Vincent, P. (2000). A neural probabilistic language model. Advances in Neural Information Processing Systems (NeurIPS).
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.

(2020). Language models are few-shot learners. Advances in Neural Information Processing Systems (NeurIPS).

- Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. (2022). Palm: Scaling language modeling with pathways. arXiv preprint arXiv:2204.02311.
- Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., et al. (2022). No language left behind: Scaling human-centered machine translation. arXiv preprint arXiv:2207.04672.
- Dong, Q., Li, L., Dai, D., Zheng, C., Wu, Z., Chang, B., Sun, X., Xu, J., Li, L., & Sui, Z. (2022). A survey for in-context learning. arXiv preprint arXiv:2301.00234.
- Elangovan, A., He, J., & Verspoor, K. (2021). Memorization vs. generalization: Quantifying data leakage in NLP performance evaluation. In Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL).
- Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., Baines, M., Celebi, O., Wenzek, G., Chaudhary, V., et al. (2021). Beyond English-centric multilingual machine translation. The Journal of Machine Learning Research (JMLR).
- Garcia, X., Bansal, Y., Cherry, C., Foster, G., Krikun, M., Feng, F., Johnson, M., & Firat, O. (2023). The unreasonable effectiveness of few-shot learning for machine translation. arXiv preprint arXiv:2302.01398.
- Goyal, N., Gao, C., Chaudhary, V., Chen, P.-J., Wenzek, G., Ju, D., Krishnan, S., Ranzato, M. A., Guzmán, F., & Fan, A. (2022). The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. Transactions of the Association for Computational Linguistics (TACL).
- Guerreiro, N. M., Alves, D., Waldendorf, J., Haddow, B., Birch, A., Colombo, P., & Martins, A. F. T. (2023). Hallucinations in large multilingual translation models. arXiv preprint arXiv:2303.16104.
- Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., & Dean, J. (2017). Google’s multilingual neural machine translation system: Enabling zero-shot translation. Transactions of the Association for Computational Linguistics (TACL).
- Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling laws for neural language models. arXiv preprint arXiv:2001.08361.
- Khandelwal, U., Levy, O., Jurafsky, D., Zettlemoyer, L., & Lewis, M. (2020). Generalization through memorization: Nearest neighbor language models. In International Conference on Learning Representations (ICLR).
- Kim, J., Kim, H. J., Cho, H., Jo, H., Lee, S. W., Lee, S.-G., Yoo, K. M., & Kim, T. (2022). Ground-truth labels matter: A deeper look into input-label demonstrations. arXiv preprint arXiv:2205.12685.
- Kudo, T., & Richardson, J. (2018). SentencePiece: A simple and language-independent subword tokenizer and detokenizer for neural text processing. arXiv preprint arXiv:1808.06226.
- Li, M., Gong, S., Feng, J., Xu, Y., Zhang, J., Wu, Z., & Kong, L. (2023). In-context learning with many demonstration examples. arXiv preprint arXiv:2302.04931.