The Impact of Example-Selection Metrics on LLM-based Machine Translation

Bálint Levente Mórász
memoQ
Budapest, Hungary
balint.morasz@memoq.com

László János Laki
memoQ
Budapest, Hungary
laszlo.laki@memoq.com

Abstract

We present a systematic evaluation of eight large language models (LLMs) on the KDE4 English→Hungarian translation benchmark, comparing zero-shot performance with two few-shot regimes defined by prompt format (*Chat* vs. *Instruct*) and example-selection metric (*fuzzy* vs. *vector*). Using both BLEU and COMET as evaluation metrics, we show that GPT-4o-mini remains the strongest zero-shot translator, while smaller or generic models such as Gemma-3-12b can match or surpass specialized translation models (EuroLLM-9b) under well-tuned few-shot settings. Surprisingly, simple character-level Levenshtein retrieval (*fuzzy*) often outperforms embedding-based (*vector*) selection, and the optimal prompt format is model-dependent: some LLMs benefit more from structured Chat-template prompts, whereas others excel with free-form instructions. Our study highlights the interaction between prompt engineering and retrieval strategy and offers practical guidelines for deploying LLMs in cost-effective machine translation.

1 Introduction

In recent years, large language models (LLMs) have emerged as state-of-the-art in natural language processing (NLP). Pre-trained on massive text corpora, these models acquire broad linguistic and factual knowledge, enabling them to handle tasks once considered out of reach — from question answering to summarization. Machine translation is no exception: modern LLMs can often rival or surpass dedicated neural translation systems, even in zero-shot settings where no parallel translation examples are provided at inference time.

An LLM's translation performance is shaped primarily by its parameter count and training data. Larger models tend to be more expressive but require more compute, while smaller LLMs run efficiently on modest hardware but may sacrifice translation quality. Just as crucial is multilingual coverage: models exposed to a language pair during pre-training generally produce more fluent translations.

A promising way to narrow the gap between zero-shot and fully supervised translation is through In-Context Learning (ICL), where a few example sentence pairs are provided at inference. Few-shot prompting can significantly boost performance, but not all models leverage examples equally well. Model architecture, training objectives, and prompt strategies all affect how well an LLM integrates demonstrations during decoding.

We systematically compare a range of open and proprietary LLMs across several benchmarks and language pairs. Here, we report results on a 1000-segment subset of the KDE4 English→Hungarian test set (Tiedemann, 2016), evaluating BLEU and COMET scores under zero-shot and two few-shot regimes: *fuzzy–Chat* (Levenshtein-based selection with a chat-style prompt) and *vector–Instruct* (embedding-based selection with an instruction-style prompt). We analyze how prompt structure and selection metrics affect translation quality, showing that optimal few-shot configurations vary by model.

This paper offers three main contributions: (1) a head-to-head evaluation of eight LLMs on a low-resource translation task; (2) an analysis of the trade-offs between model size, efficiency, and accuracy; and (3) insights into prompt and example selection strategies that maximize few-shot gains across diverse model families. Our findings offer practical guidance for deploying LLM-based MT systems.

2 Related works

Large language models have recently demonstrated strong translation capabilities through prompting. GPT-3 showed that LLMs can perform translation in few-shot settings without fine-tuning (Brown et al., 2020). Later studies confirmed that prompt-based methods, even on monolingual LLMs, can rival dedicated MT systems on standard benchmarks. Zero-shot outputs tend to be fluent and convey meaning well, though they may differ stylistically from reference translations. Accuracy improves with in-context examples and clear instructions — for instance, the 540B PaLM model performed well with optimized prompting, though it still lagged behind top supervised Transformer systems on high-resource pairs.

Comparative evaluations of LLMs and traditional MT systems have yielded mixed results. Some studies report that few-shot LLM translations can rival or surpass specialized NMT models (Bawden & Yvon, 2023; Moslem et al., 2023). Zhu et al. (2024) found that GPT-4 outperformed Meta's 54B NLLB model on 41% of directions in a multilingual benchmark, though it still trailed leading commercial MT systems overall.

Conversely, evaluations on WMT news tasks show that even advanced LLMs (e.g., ChatGPT) generally underperform compared to dedicated MT models (Hendy et al., 2023). Only the largest LLMs, such as GPT-4, approach the accuracy of state-of-the-art encoder–decoder models like NLLB on high-resource pairs, and still struggle in low-resource settings (Robinson et al., 2023).

Direct fine-tuning of LLMs for MT remains nascent but promising: early results show BLEU gains up to 8 points, with some fine-tuned mid-size LLMs outperforming GPT-4 on domain-specific tasks (Luo et al., 2025; Zhang et al., 2023). This points to a compelling path for combining broad LLM knowledge with the precision of supervised training.

3 Experiments

In our experimental setup, we systematically varied two dimensions of prompt engineering – prompt format and example-selection metric – to quantify their impact on translation quality across eight LLMs. First, we compared a structured, Chat-template–based prompt (*Chat*) against a simpler, free-form string instruction prompt (*Instruct*). Second, for selecting in-context translation pairs, we evaluated a vector-embedding–based similarity metric (*vector*) against a character-level Levenshtein distance metric (*fuzzy*).

To construct the evaluation set from KDE4, we first applied a character-length filter to exclude segments shorter than 20 characters. Next, we employed a RapidFuzz-based similarity filter, retaining only source segments that had at least five similar counterparts (each with a minimum similarity score of 75) in the training corpus. From the resulting filtered pool, 1000 segments were randomly sampled to form the final test set.

All comparisons were conducted on the KDE4 English→Hungarian task using GPT-4o-mini as our zero-shot baseline and seven locally hosted models (EuroLLM-9b, Phi-4, Gemma-3-12b, Granite-3.2-8b, Aya-Expanse-8b, Meta-LLaMA-3.1-8b, and Qwen2.5-14b). Each model was prompted under both the *Chat* and *Instruct* templates, with 4 shots selected by each similarity metric. This factorial design allowed us to isolate the effects of prompt structure and retrieval metric on both BLEU and COMET scores, revealing how different LLMs leverage in-context examples.

4 Results

Overall, GPT-4o-mini retains its lead in zero-shot translation (BLEU 22.65, COMET 79.10), but few-shot prompting substantially narrows the gap for smaller, locally hosted models. Under fuzzy–Chat, EuroLLM-9b achieves a BLEU of 50.59 (only 8% below GPT-4o-mini's 54.64) and a COMET of 86.52 (within 0.7% of GPT's 87.08). Meanwhile, Gemma-3-12b under *fuzzy–Instruct* scores BLEU 59.92 (just 4.4% under GPT's 64.29) and COMET 89.35 (virtually on par with GPT's 89.62), demonstrating that a well-tuned few-shot setup can bring smaller or generic LLMs to parity with state-of-the-art models.

Prompt format exerts a model-specific influence: GPT-4o-mini's BLEU rises by 17.7% and COMET by 2.9% when switching from *fuzzy–Chat* to *fuzzy–Instruct*, whereas EuroLLM-9b experiences a 15.1% BLEU drop and 4.1% COMET decrease. Models such as Meta-LLaMA-3.1-8b show marginal prompt

Model	Zero-Shot	Few-Shot			
		Chat		Instruct	
		fuzzy	vector	fuzzy	vector
gpt4o-mini	22.65	54.64	51.97	64.29	64.06
eurollm-9b	17.46	**50.59**	**50.10**	42.94	36.04
phi-4	16.14	49.57	48.80	47.74	47.27
gemma-3-12b	**18.74**	49.18	48.08	**59.92**	**57.23**
granite-3.2-8b	11.31	50.11	47.75	43.51	42.22
aya-expanse-8b	13.60	49.13	47.87	47.38	46.00
meta-llama-3.1-8b	15.13	40.75	41.74	41.12	39.82
qwen2.5-14b	17.86	49.83	49.54	57.17	56.21

Table 1: BLEU scores (KDE4 EN→HU) under zero-shot and few-shot settings.

Model	Zero-Shot	Few-Shot			
		Chat		Instruct	
		fuzzy	vector	fuzzy	vector
gpt4o-mini	79.10	87.08	87.01	89.62	89.70
eurollm-9b	**77.65**	86.52	86.03	83.00	79.78
phi-4	75.50	86.48	86.58	86.48	86.26
gemma-3-12b	76.47	**87.24**	**86.76**	**89.35**	**88.81**
granite-3.2-8b	57.36	82.41	82.95	80.13	81.00
aya-expanse-8b	61.25	82.91	82.69	82.61	83.07
meta-llama-3.1-8b	73.79	83.32	84.39	83.71	83.65
qwen2.5-14b	71.50	83.83	84.42	85.90	86.66

Table 2: COMET scores (KDE4 EN→HU) under zero-shot and few-shot settings.

sensitivity (COMET $+0.8\%$ for *vector–Chat* over *vector–Instruct*), while Gemma-3-12b and Qwen2.5-14b favor the Instruct template – Gemma's BLEU jumps 21.8% under *fuzzy* and Qwen's COMET gains 2.6% under *vector*. These patterns suggest that structured Chat template prompts benefit models trained on more formal or schema-oriented objectives, while conversational or dialogue-tuned LLMs exploit free-form instructions more effectively.

Finally, the choice of example-selection metric also matters: fuzzy retrieval consistently matches or outperforms vector-based retrieval in both BLEU and COMET. For example, under Chat prompts EuroLLM-9b's fuzzy BLEU exceeds vector by 0.98% and COMET by 0.49%, and Gemma-3-12b shows a 2.28% BLEU boost and 0.55% COMET gain. Even with *Instruct* prompts, fuzzy selection elevates Gemma's BLEU by 4.7% and COMET by 0.61%. Although vector retrieval occasionally prevails for specific model–prompt combinations (e.g. Granite-3.2-8b under *vector–Instruct*), these results highlight the surprising strength and efficiency of simple edit-distance–based example retrieval in few-shot machine translation.

5 Conclusion

In this work, we have carried out a comprehensive, head-to-head comparison of eight competitive LLMs on a challenging, low-resource translation task, dissecting the effects of prompt template (*Chat* vs. *Instruct*) and retrieval metric (*fuzzy* vs. *vector*) under few-shot learning. We confirm that GPT-4o-mini leads in zero-shot translation, but demonstrate that smaller or generic models like Gemma-3-12b can close the gap—and even outperform specialized translation models—when provided with high-quality in-context examples. The consistent superiority of the *fuzzy* metric across both BLEU and COMET suggests that lightweight, edit-distance-based retrieval is a robust and efficient strategy for example selection.

Our findings underscore the importance of tailoring prompt formats to a model's training paradigm:

structured Chat prompts excel for schema-oriented architectures, while conversational or instruction-tuned LLMs leverage free-form templates more effectively. These insights translate into actionable best practices for practitioners: choose *fuzzy* retrieval as a first resort, experiment with both Chat and natural-language prompts, and consider smaller, general-purpose LLMs in few-shot scenarios to save on compute without sacrificing translation quality.

References

Bawden, R., & Yvon, F. (2023, June). Investigating the translation performance of a large multilingual language model: The case of BLOOM. In M. Nurminen, J. Brenner, M. Koponen, S. Latomaa, M. Mikhailov, F. Schierl, T. Ranasinghe, E. Vanmassenhove, S. A. Vidal, N. Aranberri, M. Nunziatini, C. P. Escartín, M. Forcada, M. Popovic, C. Scarton, & H. Moniz (Eds.), *Proceedings of the 24th annual conference of the european association for machine translation* (pp. 157–170). European Association for Machine Translation. https://aclanthology.org/2023.eamt-1.16/

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *Proceedings of the 34th International Conference on Neural Information Processing Systems*.

Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y. J., Afify, M., & Awadalla, H. H. (2023). How good are gpt models at machine translation? a comprehensive evaluation. https://arxiv.org/abs/2302.09210

Luo, Z., Xu, C., Zhao, P., Sun, Q., Geng, X., Hu, W., Tao, C., Ma, J., Lin, Q., & Jiang, D. (2025). Wizardcoder: Empowering code large language models with evol-instruct. https://arxiv.org/abs/2306.08568

Moslem, Y., Haque, R., Kelleher, J. D., & Way, A. (2023, June). Adaptive machine translation with large language models. In M. Nurminen, J. Brenner, M. Koponen, S. Latomaa, M. Mikhailov, F. Schierl, T. Ranasinghe, E. Vanmassenhove, S. A. Vidal, N. Aranberri, M. Nunziatini, C. P. Escartín, M. Forcada, M. Popovic, C. Scarton, & H. Moniz (Eds.), *Proceedings of the 24th annual conference of the european association for machine translation* (pp. 227–237). European Association for Machine Translation. https://aclanthology.org/2023.eamt-1.22/

Robinson, N., Ogayo, P., Mortensen, D. R., & Neubig, G. (2023, December). ChatGPT MT: Competitive for high- (but not low-) resource languages. In P. Koehn, B. Haddow, T. Kocmi, & C. Monz (Eds.), *Proceedings of the eighth conference on machine translation* (pp. 392–418). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wmt-1.40

Tiedemann, J. (2016). OPUS – parallel corpora for everyone. *Proceedings of the 19th Annual Conference of the European Association for Machine Translation: Projects/Products*. https://aclanthology.org/2016.eamt-2.8

Zhang, S., Fang, Q., Zhang, Z., Ma, Z., Zhou, Y., Huang, L., Bu, M., Gui, S., Chen, Y., Chen, X., & Feng, Y. (2023). Bayling: Bridging cross-lingual alignment and instruction following through interactive translation for large language models. https://arxiv.org/abs/2306.10968

Zhu, W., Liu, H., Dong, Q., Xu, J., Huang, S., Kong, L., Chen, J., & Li, L. (2024, June). Multilingual machine translation with large language models: Empirical results and analysis. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Findings of the association for computational linguistics: Naacl 2024* (pp. 2765–2781). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-naacl.176