Improving Domain-Specific Translation from English into Ukrainian with Retrieval-Augmented Generation

Anton Shpigunov
Taras Shevchenko National University of Kyiv
14 Taras Shevchenko Blvd., Kyiv 01601, Ukraine
shpigunov@knu.ua

Abstract

Large language models have demonstrated competence as language translators, including for lower-resourced languages like Ukrainian. However, in specialized or novel domains, translation quality can suffer without adequate lexical and stylistic reference material. We present a retrieval-augmented approach to English–Ukrainian machine translation in a narrow domain (a private legal/military bilingual corpus), where semantically similar translation units retrieved via vector embeddings are provided as in-context examples to the LLM. We evaluate three open-weight Gemma 3 models (4B, 12B, 27B) against Gemini 3 Flash as a baseline across five augmentation conditions ($k \in \{0, 3, 5, 10, 25\}$) on a 2,581-pair index/test set split into 2,323 indexed pairs and 258 test pairs. We find that context augmentation yields statistically significant improvements in both ChrF++ and COMET for all models, with the smallest model's COMET improving by $+0.076$ at $k = 3$. However, smaller models exhibit context saturation: the 4B model's performance peaks at $k = 10$ and degrades with additional context, losing 9.72 ChrF++ points and 0.007 COMET between $k = 10$ and $k = 25$, while larger models continue to benefit.

1 Introduction and Related Work

It has been established that large language models appear to be competent language translators (Ye et al., 2025), including for lower-resourced languages (Enis and Hopkins, 2024). However, their capabilities rely on static, parametric knowledge encoded during training. This can lead to poor adaptation to new domains and difficulties translating rare, specialized, or culturally specific terminology. Additionally, frontier models provided by leading labs over an API may not be optimal for translating text in multiple scenarios, including medical, government, defence, and other sensitive settings where there are concerns over privacy, sensitive topics, or national security priorities.

To address limitations of knowledge acquired during the fixed training phase, the text-generation capabilities of LLMs have been extended using a variety of approaches. As one instance, retrieval-augmented generation (RAG) (Lewis et al., 2020) has been integrated into translation tasks to create hybrid architectures that ground models in external, non-parametric knowledge bases.

In human translation and localization, assistance based on retrieval of previously translated segments (translation memory) has been industry standard for decades, although such retrieval is overwhelmingly based on fuzzy string matching. As neural machine translation (NMT) emerged, early efforts sought to integrate this existing TM technology directly into neural workflows; for example, Bulte and Tezcan (2019) demonstrated that augmenting NMT input with fuzzy TM matches significantly boosted translation performance.

Expanding on the value of retrieved examples, subsequent non-LLM approaches moved beyond surface-level fuzzy matching. Zhang et al. (2018) employed a search engine to retrieve previously seen translation examples and incorporate them into the NMT model's decoding process, while Khandelwal et al. (2020) introduced kNN-MT, which augmented NMT decoding with dense vector similarity search over a massive datastore of cached examples to improve domain adaptation at test time without additional training.

Building upon these dense retrieval methods in the era of large language models, Donthi et al. (2024) utilized vector-based cosine similarity to retrieve culturally specific terms such as idioms from external databases, successfully improving LLM translation quality in both high-resource (Chinese) and low-resource (Urdu) languages over direct-translation baselines. Li et al. (2023) used an external knowledge base (IdiomKB) and employed

Proceedings of the Fifth Ukrainian Natural Language Processing Conference (UNLP 2026), pages 1–11

retrieval augmentation to provide context (figurative meanings of idioms) to smaller LLMs, resulting in better translations according to their evaluation. These approaches demonstrate the value of adding non-parametric, flexible human knowledge to LLMs for translation tasks.

Taking these developments further with semantic similarity search based on vector embeddings and pairing it with open-weight language models, a localized RAG approach emerges as a viable means to improve translation quality and consistency, especially where pre-existing training data is insufficient or otherwise inadequate. In other words, we expect a beneficial in-context learning effect on translation quality (Brown et al., 2020). With respect to all of the above, we specifically hypothesize that:

> *smaller, open-weight models augmented with example sentences retrieved using vector similarity search can improve MT quality and consistency within a specific domain, approaching or even surpassing MT quality provided by larger models without such augmentation.*

We demonstrate and assess this RAG-MT approach, its effect on MT quality as assessed using two automated metrics, and discuss our findings and implications for further research. The code, prompt template, and de-identified evaluation outputs supporting this work are released at `https://github.com/shpigunov/unlp2026-rag-mt`.

2 Data

2.1 Dataset Selection

We have considered two principal sources for translation data:

- WMT-2025 `uk-en` corpus with its multiple sub-components;
- a private en–uk corpus of 2,864 sentences which constitutes a full translation of the US Unified Code of Military Justice into Ukrainian.

Ultimately, we chose the private dataset out of two main considerations. Firstly, there is no way to establish whether WMT data has been used in training of frontier models (i.e., that the LLM does not already contain this data in its training set). Secondly, the WMT dataset is `uk-en`, and reversing the direction would introduce translationese artifacts into the source sentences. Zhang and Toral (2019) demonstrated that translationese in test sets inflates MT evaluation scores and can alter system rankings; Graham et al. (2020) further show that reverse-created test data compounds this problem by reducing the statistical power of human evaluations, potentially masking real quality differences between systems. For these reasons, Graham et al. recommend against using reverse-created test data in MT evaluation.

2.2 "Train"–Test Split

The split between the reference set for populating the vector index (we shall refer to it further as the "train" set for convenience and convention, although strictly speaking, no model was trained on it) has been chosen at 90% train and 10% test. This somewhat skewed split is influenced by the relatively small size of the private dataset. To ensure statistical significance on a smaller test set, we employed significance testing protocol, as explained below.

2.3 Data Preparation

Our initial exploration has shown that the chosen dataset contains repetitive segments such as references to section names, boilerplate legal language, and similar formulaic text. On a first evaluation, this led to what we assessed as leakage of training data into the testing set.

To mitigate this, we implemented a rigorous two-stage deduplication pipeline:

1. **Exact Deduplication:** All translation units first underwent strict text normalization involving HTML unescaping, NFKC Unicode normalization, markup tag removal, downcasing, and aggressive number and punctuation masking to eliminate superficial character and casing variations. We then discarded exact duplicate pairs by comparing hashes of the concatenated, normalized source and reference segments. This step removed 203 sentence pairs from the set.

2. **Near Deduplication (Fuzzy Matching):** To identify and remove near-duplicates differing only by minor syntactic variations or formatting, we applied locality-sensitive hashing (LSH) using MinHash signatures. Operating exclusively on the normalized source text, we

evaluated character-level 5-grams over 64 permutations, discarding any sentence that exhibited a Jaccard similarity of 0.90 or higher against previously indexed segments. This step further removed 80 pairs from the set.

From the 2,581 de-duplicated sentence pairs, we created a strict 90:10 train–test split (2,323 train segments and 258 test segments). To prevent data leakage, identical normalized source segments were grouped together prior to splitting, ensuring that any remaining exact source-text variants were assigned exclusively to a single split partition.

The train subset was encoded using the multilingual embedding model `baai/bge-m3` (Chen et al., 2024), which produced 1024-dimensional dense vectors for the English source segments; these were indexed in `qdrant` for cosine-similarity retrieval.

3 Experiment

3.1 Setup

To validate our hypothesis stating that smaller models augmented with retrieval can approximate or surpass translations generated by frontier models without additional context, we implemented a custom retrieval-augmented machine-translation (RAG-MT) prototype and designed a controlled evaluation framework.

Our training phase consisted of embedding the source sentence from the train set and inserting the sentence pair into the vector index.

For the testing phase, the prototype executed the following sequence for each test pair:

1. embed the source sentence with the same vector embedding model;
2. use the source-segment embeddings to perform a cosine-similarity search in the vector index populated during the training phase, and obtain k similar source-target pairs; no similarity computation over target text is performed;
3. template the source sentence and retrieved similar sentence pairs into the translation prompt, essentially creating a dynamic few-shot prompt for the LLM;
4. run LLM inference on the prompt and extract only the target text;
5. for each condition, apply ChrF++ and COMET metrics to the entire test set.

3.2 Embedding Model

For this experiment we chose `baai/bge-m3` (Chen et al., 2024). The model was chosen due to its published open weights enabling reproducibility, its strong performance on shorter multilingual segments, and the absence of a fine-tuning requirement. For the experiment it was served from Hugging Face's infrastructure, but in a local privacy-preserving setup, local inference is possible.

We initially ran the same experiment using a closed-source embedding model (`text-embedding-3-large` by OpenAI), and found that the overall metrics were slightly lower when using the open-weight embedder, but the overall dynamic described in the Results was the same. This implies that while the influence of the embedding model is not as significant as that of the LLM, the choice of embedder in a RAG-MT pipeline is more than a mere infrastructural consideration.

A conceivable way to strengthen the pipeline is to introduce a re-ranker model to increase the internal variety of retrieved examples and prevent near-duplication, a direction that can be explored in further work. Furthermore, using different retrieval methods altogether (Bouthors et al., 2024) would be a promising ablation to explore, but falls outside the scope of this investigation.

3.3 Large Language Models

Model Family. For the large language models, we selected Google's Gemini/Gemma model family. For the commercial baseline, we use `gemini-3-flash` due to its reasonable inference availability, speed, and cost.

For the target open-weight models, we use Gemma 3 family models (Team et al., 2025) in different sizes with 4B, 12B, and 27B parameters, deployable on consumer hardware. Gemma 3 is also convenient due to its proximity to our baseline in terms of similarities in tokenizer, attention mechanism, and training infrastructure. We do note differences, however, with Gemini models being sparse mixture-of-experts systems and not disclosing their exact number of parameters.

For inference infrastructure, we used OpenRouter while controlling for quantization (excluding quantized models and limiting to `bf16` only to maintain consistency and reproducibility).

To isolate the impact of retrieved context and measure optimal context-size limits, we evaluate

outputs of four models: `gemini-3-flash` as the commercial baseline, and then the open-weight Gemma 3 models `gemma3-4b`, `gemma3-12b`, and `gemma3-27b`.

Completion Parameters. Following the findings of Li et al. (2025), we select a lower completion temperature at `0.1` to increase repeatability and reduce probability of hallucination. The reasoning setting on `gemini-3-flash` was set to `minimal`, and the rest of the parameters were kept to provider defaults.

Each of these four models was evaluated with different context augmentation sizes: $k = 0$ to assess baseline model performance, and then $k \in \{3, 5, 10, 25\}$.

3.4 Choice of Evaluation Metrics

We rely on automated evaluation metrics that best reflect the rich morphology and flexible word order of Ukrainian as a target language (Shpigunov, 2025). Word-level n-gram metrics such as BLEU and edit-distance metrics such as TER over-penalize the morphological variants and word-order shifts that are natural in Ukrainian, treating different inflections of the same lemma as full mismatches and discounting permissible reorderings. Character-level matching with ChrF++ sidesteps the morphology problem by scoring partial overlap across word forms, while neural reference-based metrics such as COMET evaluate semantic equivalence in a multilingual contextual embedding space; both have been shown to correlate more strongly with human judgment than surface-level metrics for Ukrainian-target translation (Shpigunov, 2025). Pairwise differences between the conditions are evaluated segment by segment on the identical test set.

- **Lexical and Character-level:** We report ChrF++ (Popović, 2017) to capture character n-gram matches, which is particularly robust for morphologically rich target languages like Ukrainian.
- **Semantic:** We utilize neural reference-based evaluation via COMET (Rei et al., 2020) (variant used: `unbabel/wmt22-comet-da`) to measure semantic accuracy and fluency improvements over the baseline.

3.5 Tests of Statistical Significance

To evaluate the statistical significance of improvements over the unaugmented baseline ($k = 0$), we employ paired bootstrap resampling on both ChrF++ and COMET scores. To control the family-wise error rate across multiple comparisons, we apply the Holm–Bonferroni correction to the raw p-values.

3.6 Results

The performance of the selected models with the set context conditions on ChrF++ and COMET can be seen in Tables 1 and 2, respectively.

Detailed paired-bootstrap significance results are reported in Table 3 in Appendix A.

4 Discussion

By translating the same 258-sentence test set with retrieval from a 2,323-sentence set on a grid of four models by five conditions ($k \in \{0, 3, 5, 10, 25\}$) and keeping other variables such as prompts and completion temperature unchanged (temperature was kept fixed at `0.1` for all models tested), we have shown in a controlled experiment that context augmentation using retrieved similar sentences leads to a pronounced increase in both character-based and neural metrics (ChrF++ and COMET-da, respectively), with the extent of this increase being dependent on model capacity and the size of the provided context. The aggregate trend is shown in Figure 1.

4.1 Baseline Performance ($k = 0$)

In the zero-shot baseline, larger models have expectedly outperformed smaller ones, with even the smaller commercial cloud-only Gemini variant, `gemini-3-flash`, scoring higher (COMET 0.896 and ChrF++ 62.47) than the largest open-weight Gemma 3 variant, `gemma3-27b` (COMET 0.875 and ChrF++ 60.73).

Within the Gemma 3 family, the largest 27B variant scored the highest at COMET 0.875 and ChrF++ 60.73, followed by 12B (COMET 0.864, ChrF++ 54.85), and finally 4B with a markedly lower performance (COMET 0.812, ChrF++ 46.87), which reiterates the known correlation between model size and machine-translation performance (cf. Tables 1 and 2).

4.2 Impact of Context Augmentation

Augmenting the translation prompt with semantically similar translated sentences from the training portion of the same-domain corpus ($k \in \{3, 5, 10, 25\}$) has consistently and significantly ($p < 0.01$ for all paired-bootstrap configurations)

Model	$k = 0$	$k = 3$	$k = 5$	$k = 10$	$k = 25$
gemini-3-flash	62.47	77.49	77.64	78.62	78.76
gemma3-27b	60.73	77.65	79.14	79.82	79.61
gemma3-12b	54.85	71.19	73.34	74.94	75.23
gemma3-4b	46.87	68.26	68.28	68.81	59.09

Table 1: Summary of evaluation metrics. ChrF++ by model, by condition.

Model	$k = 0$	$k = 3$	$k = 5$	$k = 10$	$k = 25$
gemini-3-flash	0.896	0.931	0.934	0.935	0.936
gemma3-27b	0.875	0.920	0.928	0.927	0.931
gemma3-12b	0.864	0.906	0.914	0.911	0.916
gemma3-4b	0.812	0.888	0.892	0.892	0.885

Table 2: Summary of evaluation metrics. COMET by model, by condition.

improved assessed MT quality for all models considered.

Small Model, Big Gains. The most drastic relative improvements in translation quality were observed in the smallest model assessed, gemma3-4b. Augmenting the prompt with $k = 3$ examples led to a strong increase in both metrics, increasing COMET by +0.076 (from 0.812 to 0.888) and ChrF++ by +21.39 (from 46.87 to 68.26) over the zero-shot baseline, bringing the performance of the smallest model above the $k = 0$ baseline for gemma3-27b by COMET and above gemini-3-flash by ChrF++.

Reaching for the Cloud. Context augmentation has allowed smaller, open-weight models to surpass the MT quality of unaugmented frontier models. Even the smallest gemma3-4b, which can comfortably run on consumer hardware, when augmented with $k = 3$, is able to closely trail the baseline performance of gemini-3-flash. On the high end, gemma3-27b, which may require higher-end consumer hardware to run but can still be used offline, with $k \geq 5$ augmentation was able to surpass gemini-3-flash by ChrF++ and closely approach it by COMET scores. This allows translators to get a meaningful MT assist from offline open-weight models while translating specialized documents, without the need for potentially sensitive information to go to a cloud provider.

Diminishing Returns, Context Saturation and Context Rot. At the same time, we observed a marked divergence in how models of different sizes handle larger context windows ($k \geq 10$). Larger models like gemini-3-flash, gemma3-27b, and gemma3-12b demonstrate gains in both metrics with an increasing number of similar segments, but these gains exhibit sharply diminishing returns. However, the smallest model, gemma3-4b, demonstrates regression in both metrics when given larger contexts. This suggests that smaller models are more susceptible to confusion or hallucination when overwhelmed with excessive reference examples, a phenomenon referred to as context saturation or context rot (Bianchi et al., 2025; Hong et al., 2025).

Another observation with respect to the smallest model, gemma3-4b: descriptive statistics such as the minimum, maximum, mean, and standard deviations on both metrics (see Appendix B, Figures 2 and 3) point to similar conclusions. Overburdened with context, the smallest model appears to produce more inconsistent and low-quality translations, while larger models become more consistent and produce higher-assessed output more often. For larger models, adding more examples appears to increase consistency in both metrics.

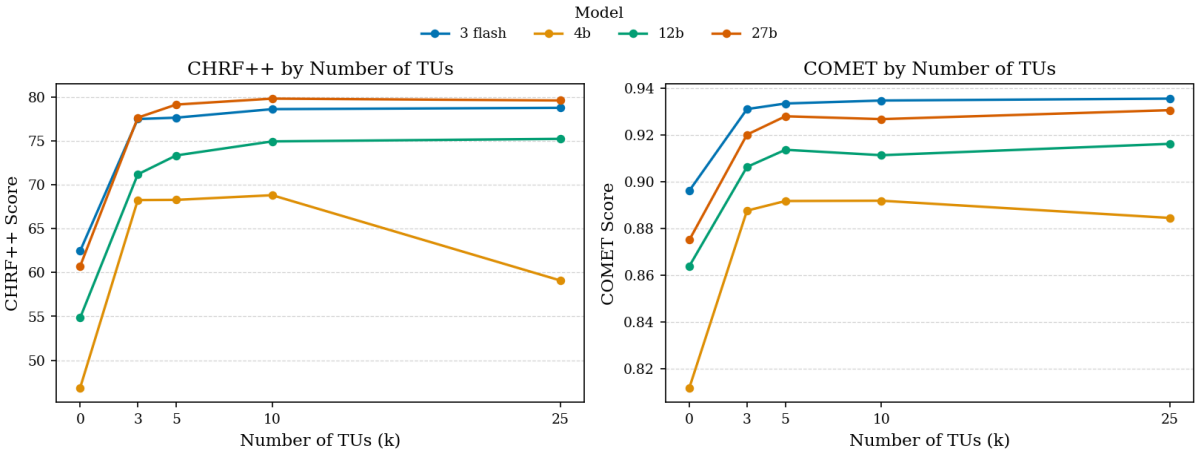

Figure 1: Average ChrF++ and COMET scores by model, by k.

Limitations

The principal limitations of this study mostly concern the data used.

First, at fewer than 3,000 sentence pairs pre-deduplication and 2,581 pairs post-deduplication, the set is rather small. On the other hand, it represents a cohesive legal document translated in the course of a real-world translation project. This translation has not been published previously and is thus guaranteed to be excluded from the training data of the LLMs assessed, as opposed to public bilingual corpora. With this said, **we do not claim or expect the observed improvements to scale to translation of unrelated documents from a different domain**; the improved MT quality can nevertheless be noticeable and helpful to translators working on large legal, technical, or other projects with consistent terminology and stylistic requirements.

Second, the human translations originated from three human translators, with an overwhelming contribution from one of them. This implies that we are evaluating how close an LLM can replicate the style and choice of terminology of an individual translator, not necessarily how universally good the translation is.

Finally, the dataset pertains to the legal/military domain where the terms and definitions are clear and unambiguous. We would not expect the demonstrated gains to scale to domains like literature, art, or colloquial speech. Studying the effects of RAG-MT across domains could be a subject of further investigation.

We also note two further ablations left for future work. First, we have not separated the contribution of semantic relevance of retrieved examples from the lift of merely having additional in-context examples; a controlled comparison against a random-pair retrieval baseline drawn uniformly from the training pool would isolate this effect and is a natural next experiment. Second, comparison against a generic-domain test set would help characterize the extent to which the observed gains are domain-specific rather than a general property of retrieval-augmented translation.

Ethical Considerations

The private testing dataset has been compiled during a volunteer translation project undertaken at the Theory and Practice of Translation from English department of the Institute of Philology, National Taras Shevchenko University of Kyiv. Permission to use translated texts has been obtained from the department and the translators involved in the project.

The source text constitutes a US legal act which is a part of US Code and is published by multiple open sources, thus no permission to use it was necessary.

Use of AI. The author used a frontier AI assistant for preliminary review feedback on drafts, verification of numerical consistency between tables, and language editing. All research design, experimental work, data analysis, and scientific conclusions are entirely the author's own. The author takes full responsibility for the content of this paper.

References

Owen Bianchi, Mathew J. Koretsky, Maya Willey, Chelsea X. Alvarado, Tanay Nayak, Adi Asija, Nicole Kuznetsov, Mike A. Nalls, Faraz Faghri, and Daniel Khashabi. 2025. Hidden in the Haystack: Smaller Needles are More Difficult for LLMs to Find. *arXiv preprint*. ArXiv:2505.18148 [cs].

Maxime Bouthors, Josep Crego, and François Yvon. 2024. Retrieving Examples from Memory for Retrieval Augmented Neural Machine Translation: A Systematic Comparison. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3022–3039, Mexico City, Mexico. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language Models are Few-Shot Learners. *arXiv preprint*. ArXiv:2005.14165 [cs].

Bram Bulte and Arda Tezcan. 2019. Neural Fuzzy Repair: Integrating Fuzzy Matches into Neural Machine Translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1800–1809, Florence, Italy. Association for Computational Linguistics.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. BGE M3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Sundesh Donthi, Maximilian Spencer, Om Patel, Joon Doh, Eid Rodan, Kevin Zhu, and Sean O'Brien. 2024. Improving LLM Abilities in Idiomatic Translation. *arXiv preprint*. ArXiv:2407.03518 [cs].

Maxim Enis and Mark Hopkins. 2024. From LLM to NMT: Advancing Low-Resource Machine Translation with Claude. *arXiv preprint*. ArXiv:2404.13813.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2020. Statistical Power and Translationese in Machine Translation Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 72–81, Online. Association for Computational Linguistics.

K. Hong, A. Troynikov, and J. Huber. 2025. Context rot: How increasing input tokens impacts LLM performance. Chroma Research.

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Nearest Neighbor Machine Translation. *arXiv preprint*. ArXiv:2010.00710 [cs].

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Lujun Li, Lama Sleem, Niccolo' Gentile, Geoffrey Nichil, and Radu State. 2025. Exploring the Impact of Temperature on Large Language Models: Hot or Cold? *arXiv preprint*. ArXiv:2506.07295 [cs].

S. Li, J. Chen, S. Yuan, X. Wu, H. Yang, S. Tao, and Y. Xiao. 2023. Translate meanings, not just words: IdiomKB's role in optimizing idiomatic translation with language models. ArXiv:2308.13961.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

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.

Anton Shpigunov. 2025. Using automated quality metrics to improve machine translation into Ukrainian. *Humanities Science Current Issues*, 2:282–290.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 Technical Report. *arXiv preprint*. ArXiv:2503.19786 [cs].

Yanfang Ye, Zheyuan Zhang, Tianyi Ma, Zehong Wang, Yiyang Li, Shifu Hou, Weixiang Sun, Kaiwen Shi, Yijun Ma, Wei Song, Ahmed Abbasi, Ying Cheng, Jane Cleland-Huang, Steven Corcelli, Robert Goulding, Ming Hu, Ting Hua, John Lalor, Fang Liu, and 10 others. 2025. LLMs4All: A Review of Large Language Models Across Academic Disciplines. *arXiv preprint*. ArXiv:2509.19580 [cs].

Jingyi Zhang, Masao Utiyama, Eiichro Sumita, Graham Neubig, and Satoshi Nakamura. 2018. Guiding Neural Machine Translation with Retrieved Translation Pieces. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1325–1335, New Orleans, Louisiana. Association for Computational Linguistics.

Mike Zhang and Antonio Toral. 2019. The Effect of Translationese in Machine Translation Test Sets. In

Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers), pages 73–81, Florence, Italy. Association for Computational Linguistics.

A Statistical Significance Results

Model	Candidate	Δ ChrF++	p-val (ChrF)	Δ COMET	p-val (COMET)
gemini-3-flash	$k = 3$	15.03	< 0.01	0.035	< 0.01
gemini-3-flash	$k = 5$	15.18	< 0.01	0.037	< 0.01
gemini-3-flash	$k = 10$	16.15	< 0.01	0.039	< 0.01
gemini-3-flash	$k = 25$	16.30	< 0.01	0.039	< 0.01
gemma3-27b	$k = 3$	16.92	< 0.01	0.045	< 0.01
gemma3-27b	$k = 5$	18.41	< 0.01	0.053	< 0.01
gemma3-27b	$k = 10$	19.09	< 0.01	0.052	< 0.01
gemma3-27b	$k = 25$	18.88	< 0.01	0.056	< 0.01
gemma3-12b	$k = 3$	16.34	< 0.01	0.042	< 0.01
gemma3-12b	$k = 5$	18.49	< 0.01	0.050	< 0.01
gemma3-12b	$k = 10$	20.09	< 0.01	0.047	< 0.01
gemma3-12b	$k = 25$	20.38	< 0.01	0.052	< 0.01
gemma3-4b	$k = 3$	21.38	< 0.01	0.076	< 0.01
gemma3-4b	$k = 5$	21.41	< 0.01	0.080	< 0.01
gemma3-4b	$k = 10$	21.94	< 0.01	0.080	< 0.01
gemma3-4b	$k = 25$	12.22	< 0.01	0.073	< 0.01

Table 3: Results of statistical significance testing (paired bootstrap, Holm-corrected vs. $k = 0$).

B Score Distributions

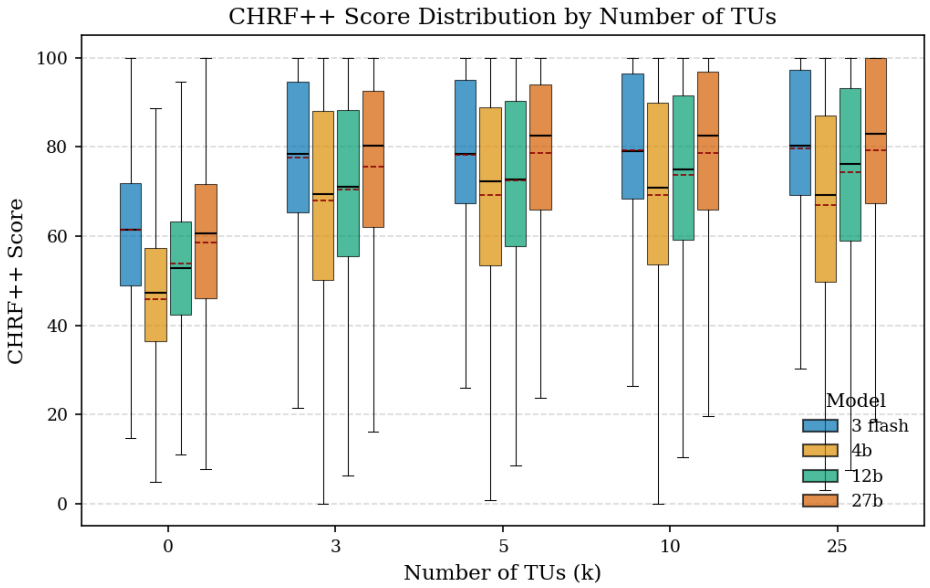

Figure 2: Distribution of ChrF++ scores across context sizes.

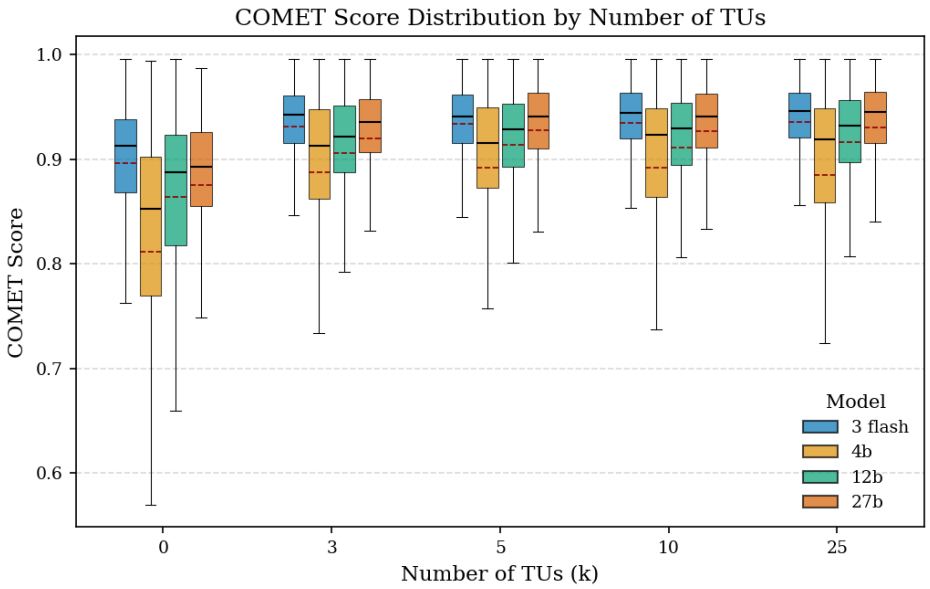

Figure 3: Distribution of COMET scores across context sizes.

C Translation Prompt

```
<Role>You are an expert language translator.</
Role>

<Instructions>

<Instruction>Please translate a text from {{
source_lang }} to {{ target_lang }} with utmost
accuracy and fluency.</Instruction>

<Instruction>

The source text may contain "tags" or "
placeables", such as `<tg123>...</tg123>` or `<
tg456>...</tg456>`. These must be rendered
verbatim, without any changes or alterations. If
 there are any tag attributes, they must not be
translated. If there is an opening tag and no
closing tags, omit the opening tag.

</Instruction>

<Instruction>Ensure that the translation
maintains original formatting, punctuation, and
special characters exactly as in the source text,
 including whitespace and especially soft-return
/soft-break characters.</Instruction>

</Instructions>

<References>

<SimilarTranslations>

<ReferenceDescription>These phrases and
sentences are semantically similar to the text
to be translated. Please refer to them for
vocabulary, grammar, style, and context.</
ReferenceDescription>

{% for tu in tus %}

<TranslationUnit>

<SourceSegment>{{ tu.source_segment }}</
SourceSegment>

<TargetSegment>{{ tu.ref_segment }}</
TargetSegment>

</TranslationUnit>

{% endfor %}

</SimilarTranslations>

</References>

<SourceText>{{ source_text }}</SourceText>

<FinalInstruction>Return ONLY the translation,
no other commentary or additional text:</
FinalInstruction>
```