Translation Analytics for Freelancers II: Benchmarking Local LLMs for Confidential Translation Workflows

Yuri Balashov*, **Rex VanHorn***, **Mingxi Xu, Austin Downes**
University of Georgia, Athens, Georgia, USA
{yuri,rex.vanhorn,mingxi.xu,austin.downes}@uga.edu

Abstract

Building on our previous work, this paper develops practical, low-barrier methods for freelance translators and smaller language service providers to evaluate translation technologies using rigorous yet accessible analytic methods. Here we address a high-stakes, specialized need: offline translation for confidentiality-sensitive domains in which privacy constraints preclude the use of cloud-based engines and commercial LLMs. We expand the Reeve Foundation Trilingual Corpus (RFTC) used in our previous work into a multilingual corpus (RFMC) by adding sentence-aligned German and Simplified Chinese reference translations. We then benchmark several locally runnable language models (via Ollama) across four language directions on 1000+ sentences selected from this corpus. We use consistent single-prompt calls without fine-tuning or domain adaptation, comparing local LLM outputs against commercial NMTs (DeepL, Baidu), a frontier LLM (GPT-5.2), and professional-grade local NMT systems (OPUS-CAT, NeuralDesktop, Promt). Automatic evaluation is conducted with MATEO. Results reveal substantial variation in local LLM performance across language directions and model sizes. The best local LLMs match or surpass local NMT systems and a frontier LLM, though they remain behind top commercial NMTs. These findings underscore the viability of carefully selected local LLM translation for privacy-constrained professionals and inform future research on model scaling and multilingual capability.

1 Introduction

Language Service Providers (LSPs), large and small, are increasingly using large language models (LLMs) in their localization workflows. Freelance translators do it too in sporadic and ad hoc ways (Penet et al., 2025). In an earlier paper (Balashov et al., 2025), we demonstrated the value of simple analytic tools—automatic metrics, structured human assessment, and lightweight statistical analysis—that were historically confined to MT research and large-scale industrial QA but are increasingly accessible to individual translators and small LSPs. Using the Christopher & Dana Reeve Foundation Trilingual Corpus (RFTC) derived from a real medical-domain translation project, we showed how freelancers can compute and interpret automatic evaluation metric scores (BLEU, chrF, TER, COMET) with minimal infrastructure and can correlate those scores with structured human judgments to determine which metrics track translator-perceived quality in a specific setting. A key takeaway was pragmatic: evaluation need not be "big tech only." Even modest, sample-based analyses can be informative and statistically stable for system comparison in a real workflow. Our findings emphasize the importance of proactive engagement with modern technologies and analytic methods to not only adapt but thrive in the rapidly evolving professional environment.

In our second paper, we apply this framework to a niche area of the translation services market. We focus on the base-level translation performance of smaller language models that can be installed and

* Equal contribution. Corresponding author.

run locally by users with modest resources and no special programming skills. This is important for several reasons outlined below.

1.1 Practical Considerations: Why Offline Translation May Be Required

In parallel with the spread of cloud MT and frontier LLMs, translation practice has also seen a countervailing demand: strict data-control workflows (Berger, 2024; Dogru and Moorkens, 2024; Lyu et al., 2024; Sandrini, 2025). Certain projects in defense, security, finance, and patent-related work require offline processing or air-gapped environments. In such scenarios, API-based commercial systems are not suitable, regardless of quality. This creates a concrete need for translators to evaluate and deploy locally runnable technologies—either local NMT toolchains[1] or newer open-weight LLMs that can be executed locally through end-user applications.

Local LLM execution has rapidly become more feasible for non-programmers due to desktop inference tools and model packaging ecosystems. The remaining question for freelancers is the empirical one we address below: whether base translation quality from small-to-mid-size local LLMs is good enough to support confidential professional workflows.

1.2 Theoretical Considerations: Scaling Laws and Interpretability

From a more theoretical perspective, studying the translation performance of smaller language models of different sizes, such as llama3.1-8b, gpt-oss-20b, or gemma3-27b, could cast additional light on "scaling laws." The earlier literature on neural language modeling shows robust scaling-law behavior for model performance as a function of parameters, data, and compute (Kaplan et al., 2020; Hoffmann et al., 2022). Translation provides a concrete capability domain where scaling may or may not behave monotonically across languages and writing systems (Ghorbani et al., 2021; Alves et al., 2024; Rei et al., 2024; Zhu et al., 2024; Richburg and Carpuat, 2024). Our multilingual corpus and a broad selection of local LLMs of different sizes and provenance give us an opportunity to study this behavior in a practical setting.

[1]For example, OPUS-CAT (Nieminen, 2021) built around OPUS-MT models (Tiedemann and Thottingal, 2020) and designed for offline professional use. Other locally installable NMT systems include NeuralDesktop and Promt (introduced as baselines in Section 3 below).

Additionally, compared with many proprietary systems, open-weight releases often provide more transparent model cards, training descriptions, and language coverage (Google DeepMind, 2025; Qwen Team, 2025a; Qwen Team, 2025b; Dang et al., 2024; Meta, 2025). Observed translation differences can therefore be discussed in relation to architecture and multilingual training choices with fewer unknowns.

1.3 Related Work

The application of LLMs to translation and other multilingual tasks began essentially as soon as LLMs themselves emerged. Brown et al. (2020) demonstrated that GPT-3 could perform zero- and few-shot translation, and subsequent work rapidly explored the boundaries of this capability. Major evaluations by Hendy et al. (2023), Jiao et al. (2023), and Vilar et al. (2023) suggested that frontier LLMs could achieve translation quality competitive with dedicated NMT systems for high-resource language pairs, though gaps remain for low-resource directions and specialized domains. A comprehensive recent survey by Ataman et al. (2025) traces the trajectory from early neural MT architectures through the current LLM era, highlighting both the promise and the open problems including domain adaptation, evaluation methodology, and the role of parallel versus monolingual training data. Lyu et al. (2024) argue that a paradigm shift is underway in which LLMs will increasingly subsume traditional NMT, particularly for document-level and stylized translation. Prompting, few-shot, and iterative-refinement strategies have continued to push LLM translation quality upward (Zhang et al., 2023; Peng et al., 2023; Garcia et al., 2023; Vilar et al., 2023; He et al., 2024; Briva-Iglesias et al., 2024; Briakou et al., 2024; Chen et al., 2024; Berger et al., 2024; Aldosari and Altuwairesh, 2025; Feng et al., 2025; Rajaee et al., 2026). Alongside general-purpose families, dedicated translation LLMs such as Tower (Alves et al., 2024; Rei et al., 2024) and, more recently, SalamandraTA (Gonzalez-Agirre et al., 2025) and Tower+ (Rei et al., 2025) have continued to improve open-weight translation quality at modest scale.

The more recent rollout and rapidly increasing accessibility of local language models, facilitated by inference platforms such as Ollama (Ollama, 2024) and LM Studio (LM Studio, 2024), have led MT researchers and translation scholars to explore

the translation capabilities of models that can run entirely on consumer hardware. Cui et al. (2025) systematically evaluate open LLMs with fewer than ten billion parameters on multilingual MT tasks, finding that models like Gemma2-9B exhibit impressive multilingual translation capabilities. They introduce the GemmaX2-28 model, which achieves competitive performance with Google Translate and GPT-4-turbo across 28 languages. Sandrini (2025) investigates the feasibility of locally deployable, free language models as alternatives to proprietary cloud-based solutions from the perspective of practicing translators, evaluating three open-source models (llama3-8b, mixtral-8x7b, and gemma2-27b) with three toolkits (GPT4ALL, Llamafile, and Ollama) to generate translations in tourism/marketing and legal domains from Italian to German. In Sandrini's setup, MATEO was used to score those translations against outputs from ChatGPT-4.0-mini and Gemini-1.5-flash, which served as reference translations.

Our work falls in this category but differs from both (Cui et al., 2025) and (Sandrini, 2025) in important respects. Unlike the former, our main audience comprises users, not developers, of translation technologies. Additionally, while Sandrini's (2025) user study is pioneering in its category, we take it to be subject to several limitations: the sample size is modest (fewer than twenty-five sentences); inference is performed on a CPU, typically relying on lower-precision quantization and suboptimal backends, which may introduce greater output variance than GPU-based higher-precision inference; and LLM outputs, rather than professional human translations, serve as reference translations.

We study a broader range of more recent local LLMs using a medical corpus originating from a real translation project. While our source language is English, our target languages (DE, RU, JA, ZH) comprise a typologically diverse set with different writing systems, thus allowing us to investigate interesting cross-linguistic phenomena. While we only use Ollama for all our translation calls, our sample is much larger (over a thousand sentences) and the translation behavior is more stable due to our use of a consumer-grade GPU. The larger sample makes our results statistically significant.

1.4 Our Contributions

This paper contributes:

- An expansion of the Reeve Foundation Trilingual Corpus (RFTC) to German and Simplified Chinese. The resulting Reeve Foundation Multilingual Corpus (RFMC) now includes approximately 3,500 source sentences in the medical domain aligned with their recent professional translations to four typologically different languages: EN→{DE, RU, JA, ZH}. With the client's permission, we make the RFMC available for noncommercial/academic use.
- A freelancer-oriented benchmark of 9 locally runnable LLMs against commercial NMT, a frontier LLM, and local NMT baselines, under "base translation only" constraints.
- A meta-analysis of prompting and decoding strategies.

What makes our approach distinctive is the combination of high-quality data from a recent professional translation project, systematic leverage of accessible small LLMs, and the translation experience of some team members, including an ATA-certified professional translator, which informs our evaluation methodology and interpretation of results. We frame this work explicitly as a *user*-oriented benchmark, written by users of translation technology, for tech-savvy freelance translators and smaller LSPs who have only modest resources and no or little programming skills. The methods and code are intended to be accessible to this audience.

The remainder of the paper presents corpus expansion (Section 2), experimental design (Section 3), translation generation (Section 4), automatic evaluation (Section 5), linguistic observations (Section 6), and implications for practice-, scaling-, and cost-oriented research (Section 7).

2 Corpus Expansion

The Reeve Foundation Trilingual Corpus (RFTC), introduced in (Balashov et al., 2025), pairs English source sentences from the International Edition of the Christopher & Dana Reeve Foundation's Paralysis Resource Guide (PRG) with their recent professional translations to Russian and Japanese. To enable a broader, multilingual study, we expanded the RFTC by adding two additional target languages: German (DE) and Simplified Chinese (Mainland China; ZH).

The expansion required aligning the existing professional translations of PRG International in German and Chinese, published by the Reeve Foun-

dation as PDF documents, with the existing trilingual segments (EN–RU–JA) already contained in the RFTC. This entailed two main steps. First, we extracted text from the PDF files in DE and ZH, performing cleanup and downsizing analogous to those described in Section 2 of (Balashov et al., 2025). Second, we undertook corpus curation to address a version mismatch. The existing translations of the PRG International to RU and JA, already contained in the RFTC, derive from the most recent English edition (2023). However, the official translations of PRG International to DE and ZH were based on the earlier English edition (2022). As a preliminary step, we identified the segments common to both editions (approximately 70% of the corpus) and retained the existing DE and ZH translations for those segments. For the remaining segments (approximately 30%), which had been added or modified in the 2023 edition, we updated the translations using expert knowledge of DE and ZH.

The resulting Christopher & Dana Reeve Foundation Multilingual Corpus (RFMC) presents approximately 3,500 source sentences (EN) aligned with their original professional translations to RU and JA, as well as their partially curated reference translations to DE and ZH. The RFMC is made available on GitHub for noncommercial/academic use.[2]

3 Experimental Design

We selected a continuous portion of RFMC—the entire cleaned and curated Chapter 1 of PRG International-2023 containing over a thousand sentences—for our experiments to explore the translation performance of 9 local language models available on Ollama (Ollama, 2024):

- aya-expanse-32b (Dang et al., 2024)
- deepseek-r1-32b (DeepSeek-AI, 2025)
- gemma3-27b (Google DeepMind, 2025)
- gpt-oss-20b (OpenAI, 2025a)
- llama3.1-8b-instruct-fp16 (Llama Team, 2024)
- mistral-small3.2-24b (Mistral AI, 2025)
- qwen2.5-32b (Qwen Team, 2025a)
- qwen3-32b (Qwen Team, 2025b)
- translategemma-27b (Finkelstein et al., 2026)

and compared their performance with three important baselines:

- The best-performing commercial NMT system: DeepL for EN–DE, EN–RU, and EN–JA; Baidu for EN–ZH[3]
- The best-performing frontier commercial LLM: GPT-5.2 (OpenAI, 2025b)
- The best-performing generally available local NMT system: Promt-23 for EN–RU, EN–JA, and EN–ZH; NeuralDesktop for EN–DE[4]

This section describes our model and prompt selection process.

3.1 Model Selection

Our model selection was guided by several practical constraints reflecting the resources typically available to tech-savvy freelance translators and smaller LSPs. We sought open-weight models that (i) could run locally via Ollama on a consumer-grade GPU, (ii) covered a range of parameter counts from 8B to 32B to allow investigation of scaling effects, (iii) represented diverse model families and provenance (open-weight releases from Google, Meta, Mistral, Alibaba, Cohere, DeepSeek, and OpenAI), (iv) routinely or historically performed well on translation benchmarks, and (v) included both general-purpose and translation-specialized architectures.

Our inference hardware consisted of NVIDIA GeForce RTX 3090 GPUs (24 GB VRAM), a consumer-grade card widely available on the secondary market at the time of writing. Each translation run was executed on a single GPU, reflecting the setup a typical freelancer or a small LSP might reasonably assemble. Although our workstation housed multiple cards, they were used only to parallelize independent experiments; no model was distributed across GPUs, and all reported results reflect single-card performance.

The inference engine was Ollama v0.14.2, running on Ubuntu 24.04. Ollama provides a straightforward command-line interface for downloading, managing, and serving open-weight models, requiring no programming expertise beyond basic terminal familiarity. We note that comparable results would be expected from other llama.cpp-based inference frontends, such as LM Studio, which offers a graphical interface and may be more approachable for users less comfortable with the command line.

[2] https://github.com/YuriBalashov/reeve-mftc

[3] https://www.deepl.com/en/translator; https://www.baidu.com

[4] https://www.promt.com; https://neuraldesktop.com

Models at or above 24B parameters (aya-expanse-32b, deepseek-r1-32b, gemma3-27b, mistral-small3.2-24b, qwen2.5-32b, qwen3-32b, translategemma-27b) were served in Ollama's default 4-bit quantization (Q4_K_M), balancing inference speed and memory footprint against output quality. The smaller models (gpt-oss-20b, llama3.1-8b-instruct) fit comfortably within the 24 GB VRAM budget without quantization: gpt-oss-20b was run at its default precision, while llama3.1-8b-instruct was run at full fp16 to explore whether preserving numerical fidelity at a smaller scale could compensate for fewer parameters. As a general principle, we selected the lowest quantization level (i.e., highest precision) that could reasonably fit into 24 GB of VRAM. All models were loaded entirely into GPU memory for each run. Because our translation pipeline processes each sentence as an independent API call with no carried context (Section 4.1), the effective context window per request was short, and VRAM was not a binding constraint for any model in our selection.

We note that while by industry standards, these are modest resources, most freelancers, even technically-inclined ones, may not have immediate access to comparable hardware. However, the landscape is changing rapidly, with GPUs and bundled inference engines increasingly packaged into new PC and Mac consumer offerings, and cloud GPU rental services becoming more affordable. Further details on hardware specifications and runtime configurations are provided in Appendix B.

3.2 Prompt Selection

Recent work on using LLMs for translation suggests that smaller models' outputs tend to be more sensitive to prompting details than the outputs from larger models (Zhang et al., 2023; Peng et al., 2023; Zhu et al., 2024; Aldosari and Altuwairesh, 2025). Since our explicit goal is to study the translation performance of smaller language models, we began with pilot experiments on a smaller set of 118 test sentences from other parts of PRG International not overlapping our main source document. These sentences were manually selected to cover linguistic phenomena representative of the entire corpus: sentences containing URLs, named entities (organization, program, and product names; personal proper names), technical terms and acronyms; and longer clauses.

The models were then asked to translate these 118 sentences to a single language from our target set (DE) with 11 different candidate prompts. Those candidate prompts were compiled as follows: our 8 selected local language models and 3 frontier LLMs (Claude Opus 4.5, GPT-5.2, and Gemini Pro 3) were "meta-prompted" to generate their own preferred prompt for translation:

> We need a system prompt for a translation task.
> The requirements are:
> - Source language: English
> - Target language: German
> - Audience: expert/academic
> - Output: translation only, no explanations, annotations, or transliterations
>
> Write the prompt you would want to receive if you were performing this task. Output only the prompt text itself, nothing else.

Candidate prompts were elicited from each model: local models via the Ollama API used later for translation, frontier LLMs via context-free manual chat sessions. Two prompt suggestions, from llama3.1-8b-instruct-fp16 and aya-expanse-32b, were judged inadequate and discarded. The prompt from llama3.1-8b-instruct was discarded because it produced an improperly formatted response that would not have functioned as a system prompt, and we chose not to manually correct it to avoid introducing experimenter bias.

Interestingly, aya-expanse-32b attempted to create its preferred translation prompt in German instead of English. This intent is consistent with recent studies that explicitly prompt LLMs in languages other than English, sometimes resulting in improved performance on multilingual tasks (Nguyen et al., 2024; Mondshine et al., 2025; Gupta et al., 2025). We decided to let our models explore this opportunity and included a German version of a "standard" prompt from (Balashov et al., 2025) alongside its English version.

Each of the 11 candidate prompts (p1–p11; see Appendix C) compiled in this way was then used to query each of our eight models to translate 118 held-out test sentences from EN to DE. Thus, every model generated a translation from every candidate prompt. The COMET-22 scores for the resulting 88 translation outputs (Table 4) do not, by themselves, mark a clear winner among and do not appear to correlate with the "native" prompt choice: in many cases the prompt suggested by model X failed to maximize the COMET scores for X's own output.

Accordingly, we decided to look for a prompt that was best "on average." Two methods were used for this purpose: (i) maximizing the sum of

z-scored COMET improvements over the prompt mean across models; and (ii) rank aggregation across the models. Both methods selected p5 as "the best prompt on average":

> **p5:** Translate the following English text into {target_lang}. The target audience is academic/expert. Provide *only* the {target_lang} translation, without any explanations, annotations, or transliterations.

During this pilot phase (mid-January 2026), TranslateGemma became available (Finkelstein et al., 2026). We decided to add it to our 8 initial LLMs but used a different prompt (p0), which was explicitly recommended by its developers:

> **p0:** You are a professional {source_lang} ({src_lang_code}) to {target_lang} ({tgt_lang_code}) translator. Your goal is to accurately convey the meaning and nuances of the original {source_lang} text while adhering to {target_lang} grammar, vocabulary, and cultural sensitivities. Produce only the {target_lang} translation, without any additional explanations or commentary. Please translate the following {source_lang} text into {target_lang}: \n \n \n {TEXT}

We used p0 only for TranslateGemma. Although this introduces an additional variable, TranslateGemma was the sole translation-specialized model in our panel, and it was trained with p0 (*ibid.*, p. 5). We judged that respecting the developer-recommended prompt for a single model gave a fairer picture of its base translation capability than forcing a generic prompt onto it.

3.3 Temperature Settings

We conducted temperature sensitivity experiments for a subset of our models across the range $T = 0.0$ to $T = 1.0$ in increments of 0.05. Consistent with prior findings on LLM-based translation (Peng et al., 2023; Balashov et al., 2025), translation quality as measured by COMET did not vary substantially across temperature settings. However, analysis of pairwise Levenshtein ratios across temperature conditions revealed a clear pattern: outputs generated at $T = 0.0$ were maximally similar to outputs at all other temperatures, effectively serving as the centroid of the output distribution. As temperature increased, outputs diverged not only from the low-temperature baseline but increasingly from one another, with the lowest pairwise similarity observed between adjacent high-temperature settings (e.g., $T = 0.95$ vs. $T = 1.0$). These results suggest that for translation, temperature primarily affects output consistency rather than quality. We therefore set $T = 0.0$ for our main experiments with "non-reasoning" models, to maximize reproducibility. A sample Levenshtein ratio matrix for the temperature sweep is provided in Appendix D and Figure 3.

However, we found that running three "reasoning" models with zero decoding temperature caused them to "loop" and time out, which is consistent with other recent studies (Pipis et al., 2025). To minimize the negative impact of infinite "reasoning loops" we decided to use $T = 1.0$ for gpt-oss-20b and $T = 0.6$ for qwen3-32b (model card overrides), and $T = 0.8$ for deepseek-r1-32b (Ollama default). We provide additional details on this in Section 4.1 and Appendix E.

3.4 Target Language Specification

Prompting language models for translation requires explicit specification of a target language. In our case, the prompt recommendation for TranslateGemma also required specifying an ISO code for the target language. This was straightforward for German (DE), Russian (RU), and Japanese (JA). The initial options for Chinese included 'Simplified Chinese', 'Mandarin', 'Mandarin Chinese', 'Chinese (Mandarin)', 'Simplified Chinese (Mandarin)', and 'Simplified Chinese (Mainland China)'. Upon consultation with a native Chinese speaker familiar with the subject matter and with ChatGPT-5.2, the set of options was reduced to 'Simplified Chinese (Mainland China)', 'Simplified Chinese', and 'Simplified Chinese (Mandarin)'. A quick translation of 24 English sentences selected from our corpus with three versions of the p5 prompt incorporating these options did not result in significant COMET score differences for any of our models. We chose to use 'Simplified Chinese (Mainland China)' and 'zh-CN' in our main experiments.

4 Generation of Translation Outputs

As noted earlier (Section 3), we selected a continuous part of RFMC—the entire curated Chapter 1 of PRG International-2023 containing 1,143 sentences—for our main experiments.

4.1 Translation Outputs from Local Language Models

We generated translations of the selected document to {DE, RU, JA, ZH}, using p5 (our "best prompt an average," Section 3.2) to interact with all the selected models except TranslateGemma, for which we used p0, as recommended by its developers.

Outputs were stored in a normalized, line-aligned format to support automated scoring and later manual review.[5]

Each source sentence was translated through an independent API call to the local Ollama server, with the selected prompt (Section 3.2) provided as the system message and the source sentence as the user message. No conversational context or translation memory was carried between calls. This sentence-by-sentence approach, while less sophisticated than document-level translation methods that leverage surrounding context (Hu et al., 2025), eliminates confounding factors related to context window management and ensures straightforward reproducibility. It also mirrors the segment-by-segment workflow common in CAT tool environments, where MT suggestions are typically generated for individual translation segments. We note that providing additional context—for example, by constructing a lightweight translation memory from previously translated segments—could plausibly improve output quality for local LLMs, but investigating such strategies is outside the scope of the present study. Our results should therefore be understood as a lower bound on achievable translation quality with these models.

As already noted, six "non-reasoning" models were run at $T = 0.0$, consistent with the findings reported in Section 3.3. Output post-processing was minimal: where a model returned extraneous formatting (e.g., markdown fences, list prefixes, or wrapping quotation marks), the output was automatically stripped to the translation content only, though such artifacts were infrequent with the models in our final selection.

The three reasoning-oriented models in our set—gpt-oss-20b, deepseek-r1-32b, and qwen3-32b—were the only models susceptible to generation failures due to their internal chain-of-reasoning processing. These failures occurred when the model entered repetitive reasoning loops, consuming the token budget without producing a final translation. gpt-oss-20b was by far the most affected, producing 99 failed lines across the four language directions when the decoding temperature was set to zero. Changing the temperature to 1.0 dramatically reduced the number of failed calls to 7, but also appears to have a negative effect on the output quality. Appendix E provides further details.

[5]Code available at `https://github.com/asdownes/translation-analytics-llm-benchmark`

All remaining models completed the full corpus without failure with $T = 0.0$. Wall-clock translation times per language direction on a single RTX 3090 ranged from roughly 19–40 minutes for non-reasoning models to 2.9–5.5 hours for reasoning-oriented models (Table 3, Appendix B); a fuller per-model runtime breakdown will accompany the code release.

4.2 Additional Translation Outputs

For each of our translation directions (EN→{DE, RU, JA, ZH}), we generated additional outputs from generally-available commercial NMT systems (Google Translate, DeepL, Baidu), a frontier LLM (GPT-5.2), and generally-available local NMT systems (Promt, NeuralDesktop, OPUS-CAT). These systems represent the spectrum of tools currently accessible to professional translators: cloud-based engines offering state-of-the-art quality, API-accessible frontier LLMs, and offline NMT solutions designed for confidential work. See Section 3 for references to these systems.

As detailed below (Section 5), based on the document-level COMET-22 scores of our outputs, we selected three systems for each language pair: the best-performing commercial NMT system, the best-performing frontier LLM, and the best-performing generally available local NMT system. These served as important baselines for our evaluations.

5 Automatic Evaluation

5.1 Metrics and Tools

Since our target users are tech-savvy freelance translators and smaller language service providers (LSPs) who have only modest resources and little or no programming skills, we used MATEO (Vanroy et al., 2023)—a free, user-friendly web application—to generate the BLEU (Papineni et al., 2001), chrF (Popović, 2015), TER (Snover et al., 2006), and COMET (Rei et al., 2020; Rei et al., 2022) scores for all 12 translation outputs (9 local LLMs + 3 baselines) for each of our language pairs. MATEO accepts parallel text files (source, reference, and up to four system outputs) and computes multiple automatic metrics through an intuitive web interface, making it an ideal tool for our target audience.

The evaluation results are reported in Section 5.2 below and Appendix F, as well as in the companion materials in our repository. We note that most of

the pairwise score differences are statistically significant at the $p < 0.05$ level, lending confidence to the system rankings we derive.

We observe strong pairwise Pearson correlation between COMET and two string-based metrics, BLEU and chrF2, across all available systems for EN→DE and EN→RU: $r = 0.73$–0.92; and moderate correlation for EN→JA and EN→ZH: $r = 0.56$–0.68. The weakest correlation, for EN→ZH, is likely attributable to the character-level properties of Chinese writing, which can introduce discrepancies between string-matching and neural evaluation approaches. Following common practice in MT evaluation, we set the string metrics aside at this point and used document- and sentence-level COMET scores in all subsequent evaluations.

5.2 Summary of Document-Level Results

Table 1 presents document-level COMET scores for all the available translation outputs, including two different temperature settings for three "reasoning models." The colored bars in Figure 2 display the best results for each of the 9 local LLMS set against the available baselines (gray bars).

Based on these scores, the best-performing local LLM for EN→{RU, ZH, JA} is translategemma-27b, while mistral-small3.2-24b demonstrates the best performance for EN→DE. We note that these local LLMs are ahead of GPT-5.2 for EN→{DE, ZH, JA} and slightly behind it for EN→RU. However, all the LLM outputs, including our local models and GPT-5.2, score significantly lower than the best commercial NMT systems: DeepL for EN→{DE, RU, JA} and Baidu for EN→ZH.

Classical scaling laws predict that model performance improves as a smooth function of parameter count (Kaplan et al., 2020; Hoffmann et al., 2022). Our document-level COMET scores (Table 1 and Figure 2) offer only partial support for this prediction in the translation domain. The clearest scaling signal appears at the lower end of the parameter range: llama3.1-8b-instruct-fp16 is one of the weakest local LLM across all four language directions, with COMET scores 1.5–2.5 points below the best-performing models. This suggests that 8B parameters, even at full fp16 precision, remain insufficient for consistently competitive medical-domain translation.

Beyond this threshold, however, scaling behavior becomes decidedly non-monotonic. Among models in the 20B–32B range, parameter count alone is a poor predictor of translation quality. The 20B gpt-oss matches or exceeds several 32B models for EN→DE and EN→JA, while the translation-specialized translategemma-27b outperforms all 32B general-purpose models for EN→RU, EN→JA, and EN→ZH. Conversely, two 32B models, deepseek-r1 and qwen2.5, rank near the bottom for EN→DE and EN→RU despite their size advantage. These reversals suggest that at the parameter scales we tested, multilingual training data composition and task-specific optimization exert a stronger influence on translation quality than raw model size.

A language-specific pattern is also visible. Performance variance across models is wider for EN→DE and EN→RU than for EN→ZH, where the scores cluster more tightly. This may reflect asymmetries in the multilingual training corpora of different model families. Notably, the relatively strong EN→ZH performance of Qwen and DeepSeek models, both originating from Chinese organizations, hints at training-data effects that interact with scaling in complex ways.

These results caution against extrapolating general scaling laws to translation: model size alone is an unreliable proxy for translation quality, and empirical benchmarking on relevant language pairs remains essential.

We also tested three locally installable NMT systems used by some translators for offline workflows—Promt, OPUS-CAT, and NeuralDesktop—on the supported language pairs (Table 1). Their document-level COMET scores vary widely. This variability may be due to the uneven quality of the base NMT models they use for different language directions (Tiedemann and Thottingal, 2020) as well as the varying degrees of multilingual coverage and domain support.

We emphasize that all outputs reported here come from base models with no domain adaptation. Many systems, including commercial engines, local NMT, and local LLMs, can be fine-tuned or adapted to terminology and style constraints. Understanding adaptation regimes with modest resources is a natural next step and is the focus of ongoing research (Moslem et al., 2023; Moslem, 2024; Vieira et al., 2024; Zheng et al., 2024; Rios, 2025), with obvious practical implications. We plan to investigate various adaptation regimes, using our corpus, in the future.

We also acknowledge a limitation of our context-

MT/LLM category	System	EN–DE	EN–RU	EN–JA	EN–ZH
NMT: Frontier	deepl-auto	**90.07**	**90.72**	n/a	89.31
	deepl-polite	n/a	n/a	**91.01**	n/a
	baidu-gen	n/a	n/a	n/a	**91.04**
	baidu-pe	n/a	n/a	n/a	90.95
	googletranslate	89.51	n/a	n/a	n/a
LLM: Frontier	gpt-5.2	88.75	89.91	90.11	89.53
LLM: Local	aya-expanse-32b	88.18	89.14	89.84	89.19
	deepseek-r1-32b-t0.0	85.61	86.15	87.77	89.29
	deepseek-r1-32b-t0.8	84.30	84.63	86.89	88.67
	gemma3-27b	88.05	89.13	89.67	89.34
	gpt-oss-20b-t0.0	86.90	86.23	89.00	88.85
	gpt-oss-20b-t1.0	88.01	88.45	89.27	88.62
	llama3.1-8b-instruct-fp16	86.23	87.01	87.36	87.11
	mistral-small3.2-24b	**88.91**	89.08	89.63	88.96
	qwen2.5-32b	85.75	86.68	88.60	89.48
	qwen3-32b-t0.0	87.23	88.30	88.98	89.55
	qwen3-32b-t0.6	87.31	88.07	89.02	89.36
	translategemma-27b	88.31	**89.55**	**90.34**	**89.88**
NMT: Local	promt-23	87.94	**88.87**	**85.33**	**86.51**
	neuraldesktop.2.1.0-auto	n/a	88.26	n/a	n/a
	neuraldesktop.2.1.0-medical	**88.68**	88.66	n/a	n/a
	opus-cat_2020-02-11	86.28	83.87	n/a	n/a
	opus-cat_bt-2021-04-14	n/a	84.98	n/a	n/a

Table 1: COMET-22 scores for all translation outputs. Highest values in each category are boldfaced.

free, sentence-by-sentence evaluation approach: both the automatic and manual evaluation scores we report reflect segment-level quality and may not fully capture document-level coherence, consistency, or discourse-related translation phenomena (Castilho et al., 2023; Hu et al., 2025). This is a deliberate constraint: many MT/LLM pipelines still operate sentence by sentence, and local inference costs make long-context prompting expensive and slow. At the same time, lack of discourse context can hurt pronoun resolution, lexical consistency, and terminology coherence. We discuss these limitations and plans for future work in Section 6 below.

6 Discussion, Limitations, and Future Work

Human evaluation has long been regarded as the gold standard in machine translation research. A companion paper to the present study therefore places human evaluation at the center of its analysis.

Human evaluation in the sequel. A large-scale, blind, randomized human evaluation of all twelve outputs (nine local LLMs plus three baselines) plus reference translations on 100+ strategically selected sentences for EN→DE, RU, ZH, using a novel two-step direct-assessment protocol and a purpose-built Streamlit interface (VanHorn, 2026), is the focus of a companion paper currently in preparation. Two preliminary findings are worth flagging here: (i) while COMET is a strong system-level proxy for human judgment, it is a much weaker one at the sentence level, which is consistent with other recent findings; (ii) a non-trivial number of reference translations were judged by human experts to be inferior to the best system outputs. This has potential implications for reference-free quality estimation methods and related approaches (Lavie et al., 2025), which could be particularly valuable for freelancers and smaller LSPs who may not always have access to high-quality reference translations.

Prompt and configuration effects as a workflow variable. Unlike conventional NMT engines, local LLM MT makes prompt design part of the system. Our pilot experiments illustrate that prompt choices can materially shift quality and error profiles, and that "native" prompts suggested by a model do not necessarily optimize that model's outputs. For small teams, this implies that prompts should be versioned and treated like configuration code: changes should be evaluated on a fixed internal benchmark set.

Decoding temperature. For reasoning models, temperature settings appear to be a double-edged sword: We find that $T = 0.0$ is associated with a substantial number of translation-call failures in

reasoning models. Changing the temperature to the model card or inference engine default ($T = 0.6 - 1.0$) reduces the number of failed calls, but also appears to degrade the overall quality of the translation outputs.

When local LLM translation is viable. Translation with local LLMs is most attractive when confidentiality constraints disallow cloud services, when marginal API costs are prohibitive, or when translators need predictable offline availability. Our results indicate that the best open models can provide high-quality drafts. However, for high-risk deliverables, such drafts should be used within a workflow that includes systematic validation rather than as a fully autonomous system.

Toward a fuller panel. The selection of nine local LLMs reflected a snapshot of late-2025 availability; specialized recent releases such as SalamandraTA and Tower+ are obvious targets for the future.

7 Concluding Remarks

Using a real-world medical-domain corpus expanded to four target languages, we compared locally runnable open models against commercial MT, a frontier LLM baseline, and locally installable NMT systems.

Two conclusions stand out. First, translation-specialized open models and those trained on balanced multilingual data substantially narrow the translation quality gap between local models and commercial engines, making translation with local LLMs increasingly viable under confidentiality and cost constraints. Second, prompt and decoding temperature choices can materially affect outcomes, so reproducibility and re-validation are essential.

Author Contributions

YB conceived the study, led the corpus expansion and curation, conducted automatic evaluation of all the outputs with MATEO, and drafted the manuscript. RVH performed the German reference translation curation and contributed to the research strategy and evaluation design. MX performed the Chinese reference translation curation and contributed linguistic observations. AD implemented all local LLM inference operations, conducted prompt selection experiments, managed the technical infrastructure, and authored the description of these processes in Sections 3 and 4. All authors reviewed and approved the final manuscript.

Acknowledgments

YB's work was supported by NSF Grant No. SES-2336713. We reiterate our thanks to the Christopher & Dana Reeve Foundation for permission to use their linguistic resources in our experiments. We are indebted to Julian Hennemann for reviewing the EN–DE outputs and contributing valuable linguistic observations. We thank Chris Schertler for sharing the TMX and XLIFF documents for the German translation of PRG International. We are grateful to Sheila Castilho and to the participants in the graduate seminar on translation technologies taught at UGA in Fall 2025 for their input and advice. Finally, we sincerely thank the reviewers for their comments, most of which we have incorporated into this revised version.

References

Aldosari, Lama Abdullah and Nasrin Altuwairesh. 2025. Assessing the effects of translation prompts on the translation quality of GPT-4 Turbo using automated and human evaluation metrics: a case study. *Perspectives*, pages 1–25, May. https://doi.org/10.1080/0907676X.2025.2464120 https://www.tandfonline.com/doi/full/10.1080/0907676X.2025.2464120

Alves, Duarte M., José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An Open Multilingual Large Language Model for Translation-Related Tasks. Version Number: 1. https://doi.org/10.48550/ARXIV.2402.17733 https://arxiv.org/abs/2402.17733

Ataman, Duygu, Alexandra Birch, Nizar Habash, Marcello Federico, Philipp Koehn, and Kyunghyun Cho. 2025. Machine Translation in the Era of Large Language Models:A Survey of Historical and Emerging Problems. *Information*, 16(9):723, August. https://doi.org/10.3390/info16090723 https://www.mdpi.com/2078-2489/16/9/723

Balashov, Yuri, Alex Balashov, and Shiho Fukuda Koski. 2025. Translation Analytics for Freelancers: I. Introduction, Data Preparation, Baseline Evaluations. In Bouillon, Pierrette, Johanna Gerlach, Sabrina Girletti, Lise Volkart, Raphael Rubino, Rico Sennrich, Ana C. Farinha, Marco Gaido, Joke Daems, Dorothy Kenny, Helena Moniz, and Sara Szoc, editors, *Proceedings of Machine Translation Summit XX: Volume 1*,

pages 538–565, Geneva, Switzerland, June. European Association for Machine Translation. https://aclanthology.org/2025.mtsummit-1.42/

Berger, Nathaniel, Stefan Riezler, Miriam Exel, and Matthias Huck. 2024. Prompting Large Language Models with Human Error Markings for Self-Correcting Machine Translation. In Scarton, Carolina, Charlotte Prescott, Chris Bayliss, Chris Oakley, Joanna Wright, Stuart Wrigley, Xingyi Song, Edward Gow-Smith, Rachel Bawden, Víctor M Sánchez-Cartagena, Patrick Cadwell, Ekaterina Lapshinova-Koltunski, Vera Cabarrão, Konstantinos Chatzitheodorou, Mary Nurminen, Diptesh Kanojia, and Helena Moniz, editors, *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 636–646, Sheffield, UK, June. European Association for Machine Translation (EAMT). https://aclanthology.org/2024.eamt-1.54/

Berger, Carola. 2024. Machine Translation Post-Editing (MTPE, PEMT): Facts, not Fiction. https://www.cfbtranslations.com/wp-content/uploads/2024/06/LEO-MTPE-Berger.pdf

Briakou, Eleftheria, Jiaming Luo, Colin Cherry, and Markus Freitag. 2024. Translating Step-by-Step: Decomposing the Translation Process for Improved Translation Quality of Long-Form Texts. https://arxiv.org/abs/2409.06790

Briva-Iglesias, Vicent, Gokhan Dogru, and João Lucas Cavalheiro Camargo. 2024. Large language models "ad referendum": How good are they at machine translation in the legal domain? *MonTI. Monografías de Traducción e Interpretación*, (16):75–107, May. https://doi.org/10.6035/MonTI.2024.16.02 https://www.e-revistes.uji.es/index.php/monti/article/view/7514

Brown, Tom B., 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, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. https://arxiv.org/abs/2005.14165

Castilho, Sheila, Clodagh Quinn Mallon, Rahel Meister, and Shengya Yue. 2023. Do online Machine Translation Systems Care for Context? What About a GPT Model? In Nurminen, Mary, Judith Brenner, Maarit Koponen, Sirkku Latomaa, Mikhail Mikhailov, Frederike Schierl, Tharindu Ranasinghe, Eva Vanmassenhove, Sergi Alvarez Vidal, Nora Aranberri, Mara Nunziatini, Carla Parra Escartín, Mikel Forcada, Maja Popovic, Carolina Scarton, and Helena Moniz, editors, *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 393–417, Tampere, Finland, June. European Association for Machine Translation. https://aclanthology.org/2023.eamt-1.39/

Chen, Pinzhen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2024. Iterative Translation Refinement with Large Language Models. In Scarton, Carolina, Charlotte Prescott, Chris Bayliss, Chris Oakley, Joanna Wright, Stuart Wrigley, Xingyi Song, Edward Gow-Smith, Rachel Bawden, Víctor M Sánchez-Cartagena, Patrick Cadwell, Ekaterina Lapshinova-Koltunski, Vera Cabarrão, Konstantinos Chatzitheodorou, Mary Nurminen, Diptesh Kanojia, and Helena Moniz, editors, *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 181–190, Sheffield, UK, June. European Association for Machine Translation (EAMT). https://aclanthology.org/2024.eamt-1.17/

Cui, Menglong, Pengzhi Gao, Wei Liu, Jian Luan, and Bin Wang. 2025. Multilingual Machine Translation with Open Large Language Models at Practical Scale: An Empirical Study. In Chiruzzo, Luis, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5420–5443, Albuquerque, New Mexico, April. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.naacl-long.280 https://aclanthology.org/2025.naacl-long.280/

Dang, John, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. 2024. Aya Expanse: Combining Research Breakthroughs for a New Multilingual Frontier. https://arxiv.org/abs/2412.04261

DeepSeek-AI. 2025. DeepSeek-R1. Technical report. https://github.com/deepseek-ai/DeepSeek-R1

Dogru, Gokhan and Joss Moorkens. 2024. Data Augmentation with Translation Memories for Desktop Machine Translation Fine-tuning in 3 Language Pairs. *The Journal of Specialised Translation*, (41):149–178, January. https://doi.org/10.26034/cm.jostrans.2024.4716 https://www.jostrans.org/article/view/4716

Feng, Zhaopeng, Yan Zhang, Hao Li, Bei Wu, Jiayu Liao, Wenqiang Liu, Jun Lang, Yang Feng, Jian Wu, and Zuozhu Liu. 2025. TEaR: Improving LLM-based Machine Translation with Systematic Self-Refinement. In Chiruzzo, Luis, Alan Ritter, and Lu Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3922–3938, Albuquerque, New Mexico, April. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.findings-naacl.218 https://aclanthology.org/2025.findings-naacl.218/

Finkelstein, Mara, Isaac Caswell, Tobias Domhan, Jan-Thorsten Peter, Juraj Juraska, Parker Riley, Daniel Deutsch, Geza Kovacs, Cole Dilanni, Colin Cherry, Eleftheria Briakou, Elizabeth Nielsen, Jiaming Luo, Kat Black, Ryan Mullins, Sweta Agrawal, Wenda Xu, Erin Kats, Stephane Jaskiewicz, Markus Freitag, and David Vilar. 2026. TranslateGemma Technical Report. Technical report. https://arxiv.org/abs/2601.09012

Garcia, Xavier, Yamini Bansal, Colin Cherry, George Foster, Maxim Krikun, Fangxiaoyu Feng, Melvin Johnson, and Orhan Firat. 2023. The unreasonable effectiveness of few-shot learning for machine translation. https://doi.org/10.48550/ARXIV.2302.01398 https://arxiv.org/abs/2302.01398

Ghorbani, Behrooz, Orhan Firat, Markus Freitag, Ankur Bapna, Maxim Krikun, Xavier Garcia, Ciprian Chelba, and Colin Cherry. 2021. Scaling Laws for Neural Machine Translation. _eprint: 2109.07740. https://arxiv.org/abs/2109.07740

Gonzalez-Agirre, Aitor, Marc Pàmies, Joan Llop, Irene Baucells, Severino Da Dalt, Daniel Tamayo, José Javier Saiz, Ferran Espuña, Jaume Prats, Javier Aula-Blasco, Mario Mina, Iñigo Pikabea, Adrián Rubio, Alexander Shvets, Anna Sallés, Iñaki Lacunza, Jorge Palomar, Júlia Falcão, Lucía Tormo, Luis Vasquez-Reina, Montserrat Marimon, Oriol Pareras, Valle Ruiz-Fernández, and Marta Villegas. 2025. Salamandra technical report. https://arxiv.org/abs/2502.08489

Google DeepMind. 2025. Gemma 3 Technical Report. Technical report. https://arxiv.org/abs/2503.19786

Gupta, Aman, Yingying Zhuang, Zhou Yu, Ziji Zhang, and Anurag Beniwal. 2025. How and Where to Translate? The Impact of Translation Strategies in Cross-lingual LLM Prompting. _eprint: 2507.22923. https://arxiv.org/abs/2507.22923

He, Zhiwei, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2024. Exploring Human-Like Translation Strategy with Large Language Models. *Transactions of the Association for Computational Linguistics*, 12:229–246. https://doi.org/10.1162/tacl_a_00642 https://aclanthology.org/2024.tacl-1.13/

Hendy, Amr, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation. https://doi.org/10.48550/ARXIV.2302.09210 https://arxiv.org/abs/2302.09210

Hoffmann, Jordan, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack W. Rae, and Laurent Sifre. 2022. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Hu, Hanxu, Jannis Vamvas, and Rico Sennrich. 2025. Source-primed Multi-turn Conversation Helps Large Language Models Translate Documents. https://arxiv.org/abs/2503.10494

Jiao, Wenxiang, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is ChatGPT A Good Translator? Yes With GPT-4 As The Engine. https://doi.org/10.48550/arXiv.2301.08745

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models. https://arxiv.org/abs/2001.08361

Lavie, Alon, Greg Hanneman, Sweta Agrawal, Diptesh Kanojia, Chi-Kiu Lo, Vilém Zouhar, Frederic Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Deoghare, Archchana Sindhujan, Jiayi Wang, David Ifeoluwa Adelani, Brian Thompson, Tom Kocmi, Markus Freitag, and Daniel Deutsch. 2025. Findings of the WMT25 Shared Task on Automated Translation Evaluation Systems: Linguistic Diversity is Challenging and References Still Help. In Haddow, Barry, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Tenth Conference on Machine Translation*, pages 436–483, Suzhou, China, November. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.wmt-1.24 https://aclanthology.org/2025.wmt-1.24/

Llama Team. 2024. The Llama3 Herd of Models. https://ai.meta.com/research/publications/the-llama-3-herd-of-models/

LM Studio. 2024. LM Studio: Discover, download, and run local LLMs. https://lmstudio.ai/

Lyu, Chenyang, Zefeng Du, Jitao Xu, Yitao Duan, Minghao Wu, Teresa Lynn, Alham Fikri Aji, Derek F. Wong, and Longyue Wang. 2024. A Paradigm Shift: The Future of Machine Translation Lies with Large

Language Models. In Calzolari, Nicoletta, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1339–1352, Torino, Italia, May. ELRA and ICCL. https://aclanthology.org/2024.lrec-main.120/

Meta. 2025. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation. https://ai.meta.com/blog/llama-4-multimodal-intelligence/

Mistral AI. 2025. Mistral Small 3.1. Technical report. https://mistral.ai/news/mistral-small-3-1

Mondshine, Itai, Tzuf Paz-Argaman, and Reut Tsarfaty. 2025. Beyond English: The Impact of Prompt Translation Strategies across Languages and Tasks in Multilingual LLMs. In Chiruzzo, Luis, Alan Ritter, and Lu Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1331–1354, Albuquerque, New Mexico, April. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.findings-naacl.73 https://aclanthology.org/2025.findings-naacl.73/

Moslem, Yasmin, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive Machine Translation with Large Language Models. In Nurminen, Mary, Judith Brenner, Maarit Koponen, Sirkku Latomaa, Mikhail Mikhailov, Frederike Schierl, Tharindu Ranasinghe, Eva Vanmassenhove, Sergi Alvarez Vidal, Nora Aranberri, Mara Nunziatini, Carla Parra Escartín, Mikel Forcada, Maja Popovic, Carolina Scarton, and Helena Moniz, editors, *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere, Finland, June. European Association for Machine Translation. https://aclanthology.org/2023.eamt-1.22/

Moslem, Yasmin. 2024. Language Modelling Approaches to Adaptive Machine Translation. https://arxiv.org/abs/2401.14559

Nguyen, Xuan-Phi, Mahani Aljunied, Shafiq Joty, and Lidong Bing. 2024. Democratizing LLMs for Low-Resource Languages by Leveraging their English Dominant Abilities with Linguistically-Diverse Prompts. In Ku, Lun-Wei, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3501–3516, Bangkok, Thailand, August. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.192 https://aclanthology.org/2024.acl-long.192/

Nieminen, Tommi. 2021. OPUS-CAT: Desktop NMT with CAT integration and local fine-tuning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 288–294, Online, April. Association for Computational Linguistics. https://www.aclweb.org/anthology/2021.eacl-demos.34

Ollama. 2024. Ollama: Get up and running with large language models locally. https://ollama.com/

OpenAI. 2025a. Introducing gpt-oss, August. https://openai.com/index/introducing-gpt-oss/

OpenAI. 2025b. Introducing GPT-5.2. Technical report. https://openai.com/index/introducing-gpt-5-2/

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, page 311, Philadelphia, Pennsylvania. Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135 http://portal.acm.org/citation.cfm?doid=1073083.1073135

Penet, JC, Joss Moorkens, and Yamada Masaru. 2025. Teaching translation in the age of generative AI: New paradigm, new learning?, November. https://doi.org/10.5281/ZENODO.17580856 https://zenodo.org/doi/10.5281/zenodo.17580856

Peng, Keqin, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards Making the Most of ChatGPT for Machine Translation. In Bouamor, Houda, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5622–5633, Singapore, December. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.373 https://aclanthology.org/2023.findings-emnlp.373/

Pipis, Charilaos, Shivam Garg, Vasilis Kontonis, Vaishnavi Shrivastava, Akshay Krishnamurthy, and Dimitris Papailiopoulos. 2025. Wait, wait, wait... why do reasoning models loop? https://arxiv.org/abs/2512.12895

Popović, Maja. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September. Association for Computational Linguistics. https://doi.org/10.18653/v1/W15-3049 https://aclanthology.org/W15-3049

Qwen Team. 2025a. Qwen2.5 Technical Report. Technical report. https://arxiv.org/abs/2412.15115

Qwen Team. 2025b. Qwen3 Technical Report. Technical report. https://arxiv.org/abs/2505.09388

Rajaee, Sara, Sebastian Vincent, Alexandre Berard, Marzieh Fadaee, Kelly Marchisio, and Tom Kocmi. 2026. Unlocking reasoning capability on machine translation in large language models. https://arxiv.org/abs/2602.14763

Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. In Webber, Bonnie, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online, November. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.213 https://aclanthology.org/2020.emnlp-main.213/

Rei, Ricardo, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 Submission for the Metrics Shared Task. In Koehn, Philipp, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid), December. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.wmt-1.52 https://aclanthology.org/2022.wmt-1.52/

Rei, Ricardo, Jose Pombal, Nuno M. Guerreiro, João Alves, Pedro Henrique Martins, Patrick Fernandes, Helena Wu, Tania Vaz, Duarte Alves, Amin Farajian, Sweta Agrawal, Antonio Farinhas, José G. C. De Souza, and André Martins. 2024. Tower v2: Unbabel-IST 2024 Submission for the General MT Shared Task. In Haddow, Barry, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Ninth Conference on Machine Translation*, pages 185–204, Miami, Florida, USA, November. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.wmt-1.12 https://aclanthology.org/2024.wmt-1.12/

Rei, Ricardo, Nuno M. Guerreiro, José Pombal, João Alves, Pedro Teixeirinha, Amin Farajian, and André F. T. Martins. 2025. Tower+: Bridging generality and translation specialization in multilingual llms. https://arxiv.org/abs/2506.17080

Richburg, Aquia and Marine Carpuat. 2024. How Multilingual Are Large Language Models Fine-Tuned for Translation? https://arxiv.org/abs/2405.20512

Rios, Miguel. 2025. Instruction-tuned Large Language Models for Machine Translation in the Medical Domain. In Bouillon, Pierrette, Johanna Gerlach, Sabrina Girletti, Lise Volkart, Raphael Rubino, Rico Sennrich, Ana C. Farinha, Marco Gaido, Joke Daems, Dorothy Kenny, Helena Moniz, and Sara Szoc, editors, *Proceedings of Machine Translation Summit XX: Volume 1*, pages 162–172, Geneva, Switzerland, June. European Association for Machine Translation. https://aclanthology.org/2025.mtsummit-1.13/

Sandrini, Peter. 2025. Beyond the Cloud: Assessing the Benefits and Drawbacks of Local LLM Deployment for Translators. Version Number: 1. https://doi.org/10.48550/ARXIV.2507.23399 https://arxiv.org/abs/2507.23399

Snover, Matthew, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA, August. Association for Machine Translation in the Americas. https://aclanthology.org/2006.amta-papers.25/

Tiedemann, Jörg and Santhosh Thottingal. 2020. OPUS-MT – Building open translation services for the World. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal, November. European Association for Machine Translation. https://aclanthology.org/2020.eamt-1.61

VanHorn, Rex. 2026. A streamlit interface for two-step direct assessment of translation quality. Forthcoming.

Vanroy, Bram, Arda Tezcan, and Lieve Macken. 2023. MATEO: MAchine translation evaluation online. In Nurminen, Mary, Judith Brenner, Maarit Koponen, Sirkku Latomaa, Mikhail Mikhailov, Frederike Schierl, Tharindu Ranasinghe, Eva Vanmassenhove, Sergi Alvarez Vidal, Nora Aranberri, Mara Nunziatini, Carla Parra Escartín, Mikel Forcada, Maja Popovic, Carolina Scarton, and Helena Moniz, editors, *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 499–500, Tampere, Finland, June. European Association for Machine Translation. https://aclanthology.org/2023.eamt-1.52/

Vieira, Inacio, Will Allred, Séamus Lankford, Sheila Castilho, and Andy Way. 2024. How Much Data is Enough Data? Fine-Tuning Large Language Models for In-House Translation: Performance Evaluation Across Multiple Dataset Sizes. In Knowles, Rebecca, Akiko Eriguchi, and Shivali Goel, editors, *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 236–249, Chicago, USA, September. Association for Machine Translation in the Americas. https://aclanthology.org/2024.amta-research.20/

Vilar, David, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting PaLM for Translation: Assessing Strategies and Performance. In Rogers, Anna, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427, Toronto, Canada, July. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.859 https://aclanthology.org/2023.acl-long.859/

Zhang, Biao, Barry Haddow, and Alexandra Birch. 2023. Prompting Large Language Model for Machine Translation: A Case Study. https://doi.org/10.48550/ARXIV.2301.07069 https://arxiv.org/abs/2301.07069

Zheng, Jiawei, Hanghai Hong, Feiyan Liu, Xiaoli Wang, Jingsong Su, Yonggui Liang, and Shikai Wu. 2024. Fine-tuning Large Language Models for Domain-specific Machine Translation. https://arxiv.org/abs/2402.15061

Zhu, Wenhao, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis. In Duh, Kevin, Helena Gomez, and Steven Bethard, editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico, June. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-naacl.176 https://aclanthology.org/2024.findings-naacl.176/

A The Reeve Foundation Multilingual Corpus

Our corpus comprises 3,500 English source sentences aligned with their original professional translations into Russian and Japanese, as well as partially curated reference translations into German and Chinese. With the client's permission, we **release** it for noncommercial/academic use.

B Local Inference Setup

All local LLM experiments were conducted on a workstation running Ubuntu 24.04.4 LTS, equipped with NVIDIA GeForce RTX 3090 GPUs (24 GB VRAM), NVIDIA driver version 580.126.09, and CUDA 13.0. Each translation run used a single GPU. The inference engine was Ollama v0.14.2.

Table 2 reports the quantization settings for each model. Models at or above 24B parameters were served in Ollama's default 4-bit quantization (Q4_K_M). The two smaller models, gpt-oss-20b and llama3.1-8b-instruct, were run at their default (unquantized) precision, as both fit within the 24 GB VRAM budget without compression. Llama3.1-8b-instruct was run at full fp16.

Model	Params	Quantization
aya-expanse-32b	32B	Q4_K_M
deepseek-r1-32b	32B	Q4_K_M
gemma3-27b	27B	Q4_K_M
gpt-oss-20b	20B	Default (unquant.)
llama3.1-8b-instruct	8B	fp16
mistral-small3.2-24b	24B	Q4_K_M
qwen2.5-32b	32B	Q4_K_M
qwen3-32b	32B	Q4_K_M
translategemma-27b	27B	Q4_K_M

Table 2: Model quantization settings.

Table 3 reports wall-clock times for translating the full 1,143-sentence corpus per language direction on a single RTX 3090. Non-reasoning models completed each direction in approximately 19 to 40 minutes. The three reasoning-oriented models were substantially slower: qwen3-32b required 4.4 to 5.5 hours per direction, deepseek-r1-32b approximately 2.9 to 3.8 hours, and gpt-oss-20b approximately 43 to 64 minutes. This overhead reflects the models' internal multi-step processing, which generates extended reasoning traces before producing the final translation.

Model	DE	RU	JA	ZH
aya-expanse	28m	27m	26m	23m
deepseek-r1	3h47	3h44	3h28	2h51
gemma3	27m	28m	25m	23m
gpt-oss	58m	64m	61m	43m
llama3.1-8b	23m	22m	22m	19m
mistral-small	19m	22m	23m	20m
qwen2.5	31m	39m	29m	19m
qwen3	5h32	5h22	5h29	4h21
translategemma	28m	29m	26m	24m

Table 3: Wall-clock time per language direction (1,143 sentences, single RTX 3090).

C Prompt Portfolio

This appendix lists the full text of all candidate prompts used in the prompt selection pilot (Section 3.2). Prompts p1–p4 and p6–p9 were elicited by meta-prompting the indicated model to generate its own preferred translation prompt. Prompts p10 and p11 are "standard" prompts from (Balashov et al., 2025), in German and English respectively. Prompt p5 was generated by gemma3-27b and was selected as the best-performing prompt on average across all models (Section 3.2 and Table 4). Prompt

p0 is the developer-recommended prompt for TranslateGemma (Finkelstein et al., 2026) and uses a different template format with explicit language code placeholders.

All prompts shown below are in their EN→DE pilot form. For the main experiments, p5 was adapted to each target language by substituting the appropriate language name.

p0 (TranslateGemma developer-recommended): You are a professional {SOURCE_LANG} ({SOURCE_CODE}) to {TARGET_LANG} ({TARGET_CODE}) translator. Your goal is to accurately convey the meaning and nuances of the original {SOURCE_LANG} text while adhering to {TARGET_LANG} grammar, vocabulary, and cultural sensitivities. Produce only the {TARGET_LANG} translation, without any additional explanations or commentary. Please translate the following {SOURCE_LANG} text into {TARGET_LANG}: \n \n \n {TEXT}

p1 (Claude Opus 4.5): Translate the following English sentence to German. Output only the translation, nothing else.

p2 (GPT-5.2): You are a professional translator. Task: Translate the following sentence from English to German. Requirements: Output ONLY the German translation (no explanations, no quotes, no alternatives). Preserve meaning exactly; do not add or omit information. Keep names, numbers, units, and formatting unchanged. Use natural, idiomatic German. Use neutral formal register (Sie) unless the English clearly implies informal (du). Keep the same tense and voice; preserve emphasis if possible.

p3 (Gemini Pro 3): You are a professional translator proficient in English and German. Your task is to translate the provided English text into German. Constraints: Maintain the original tone and meaning. Output only the German translation. Do not include preambles, explanations, or enclosing quotes.

p4 (deepseek-r1-32b): You are a professional translator specializing in academic texts. Translate the following English text into German, maintaining its formal tone and precise terminology. Provide only the translation without any explanations, annotations, or transliterations.

p5 (gemma3-27b, selected for main experiments): Translate the following {SOURCE_LANG} text into {TARGET_LANG}. The target audience is academic/expert. Provide only the {TARGET_LANG} translation, without any explanations, annotations, or transliterations.

p6 (gpt-oss-20b): You are a professional translator. Translate the following English text into German for an expert/academic audience. Output only the translated text, with no explanations, annotations, or transliterations.

p7 (mistral-small3.2-24b): Translate the following English text into German, maintaining an expert/academic level of language and terminology. Provide only the translation, without any explanations, annotations, or transliterations.

p8 (qwen2.5-32b): Translate the following English text into German for an expert/academic audience. Provide only the translation without any explanations, annotations, or transliterations.

p9 (qwen3-32b): Translate the following English text into German. The translation should be formal, precise, and tailored for an expert/academic audience. Provide only the translated text, with no additional explanations, annotations, or transliterations. Ensure technical terminology and nuanced meanings are accurately preserved.

p10 (standard prompt, German, from (Balashov et al., 2025)): Sie sind ein erfahrener Übersetzer und übersetzen für ein Fachpublikum. Bitte fügen Sie Ihrer Übersetzung keine Anmerkungen, Erklärungen oder Transliterationen hinzu. Bitte übersetzen Sie den folgenden Satz ins Deutsche:

p11 (standard prompt, English, from (Balashov et al., 2025)): You are an expert translator, translating for an expert audience. Please do not provide any annotations, explanations or transliterations in your translation. Please translate the following English sentence to German:

Table 4 presents the COMET scores for the pilot experiments with 118 held-out test sentences (EN→DE).

D Levenshtein Ratio Matrices for Different Temperature Settings

To assess the effect of decoding temperature on output consistency, we ran a sweep across 21 temperature values from $T = 0.0$ to $T = 1.0$ in increments of 0.05, using the 118-sentence EN→DE pilot set (Section 3.2). We generated one full translation of

	aya	ds-r1	gemma3	gpt-oss	llama3.1	mistral	qwen2.5	qwen3
p1	86.30	83.01	86.93	85.71	83.27	85.96	83.10	85.19
p2	**86.55**	83.32	86.94	86.48	83.00	86.03	84.66	85.73
p3	85.78	83.17	**87.05**	86.24	83.76	86.31	84.19	85.36
p4	85.75	83.00	86.75	85.86	83.90	86.41	84.39	84.86
p5	85.87	83.86	86.75	86.50	**84.56**	86.56	84.16	85.53
p6	85.90	82.77	86.32	**86.91**	83.89	**86.63**	**84.84**	**85.74**
p7	86.00	83.43	85.65	86.42	84.10	86.42	84.28	85.34
p8	85.97	82.70	86.69	86.14	83.96	86.51	83.82	84.97
p9	86.06	83.25	85.29	86.49	83.96	86.04	83.79	85.43
p10	85.64	**83.99**	86.22	86.54	83.45	86.49	84.27	84.81
p11	86.47	83.72	86.51	86.38	83.23	86.26	83.70	85.00

Table 4: COMET-22 scores for pilot experiments with 118 held-out test sentences (EN→DE). The highest score for each model is boldfaced; the heat map of z-scored COMET improvements over the prompt mean across models is available in the companion materials.

the pilot set at each temperature setting, then computed the mean pairwise Levenshtein ratio between all 21 x 21 combination of outputs. Each cell in the resulting matrix reports the mean Levenshtein ratio (computed over character sequences) across all 118 sentences for the corresponding temperature pair.

Figure 3 shows the matrix for aya-expanse-32b. The row and column corresponding to $T = 0.0$ show the highest similarity values throughout, confirming that zero-temperatures outputs are maximally close to outputs at all other temperatures. Similarity declines as the distance between compared temperatures increases, with the lowest values concentrated in the bottom-right corner of the matrix (high-temperature pairs).

E Translation Failures From Infinite "Reasoning Loops"

Table 5 presents the number of failed translation calls resulting from prohibitively long "thinking loops" produced by three "reasoning models" on two temperature settings.

The scores reflected in Table 1 and Figure 2 reflect these failures: i.e. the models are penalized for the empty output lines. We think this is fair when reporting the actual performance of the models. To see how temperature settings affect the sentence-by-sentence quality of the outputs in cases of successful translation generation, we eliminated all outputs with at least one missing translation for each language pair and calculated the average document-level COMET score gains or losses denoted by '$T_0 \rightarrow T_1$'. We also calculated the document-level average of the absolute values of such gains and losses: '$|T_0 \rightarrow T_1|$'. The former gives a rough measure of the overall change in translation performance associated with the temperature change, while the latter represents the variation of such changes (Table 6).

We emphasize that these numbers represent relative differences in the automatic scores and do not by themselves tell us anything about the actual quality of the outputs. In fact, the outputs from qwen3-32b and deepseek-r1-32b appears to be substandard for all language pairs except EN–ZH (Section 5.2).

A typical histogram of the sentence-by-sentence COMET score differences representing a fine-grained breakdown of '$|T_0 \rightarrow T_1|$' from Figure 1 is centered in the negative area thus indicating the overall loss of translation quality. We have reason to believe the long tails of the distributions of such differences may have contributed the most to the average losses reflected in Table 6 and that these tails are associated with notable linguistic issues. We plan to review them in a sequel to this paper which will include human evaluation of a substantial number of our translation outputs.

F Automatic Evaluation: Charts

Figure 2 displays COMET-22 scores for translation outputs (from those listed in Table 1). Colored bars represent the best outputs from each of the 9 local LLMs, gray bars the outputs from the baselines for each translation direction.

Model	Temp	DE	RU	JA	ZH	Total
gpt-oss-20b	0.0	34	46	13	6	99
	1.0	2	1	-	-	3
deepseek-r1-32b	0.0	1	-	-	-	1
	0.8	1	3	-	-	4
qwen3-32b	0.0	6	-	3	-	9
	0.6	4	3	-	1	8

Table 5: Number of failed translation calls due to "thinking loops" generated by three "reasoning models."

Model	Temp	DE	RU	JA	ZH
gpt-oss-20b	$T_0 \rightarrow T_{1.0}$	-0.32	-0.14	-0.39	-0.54
	$\lvert T_0 \rightarrow T_{1.0}\rvert$	1.82	2.16	1.87	2.17
deepseek-r1-32b	$T_0 \rightarrow T_{0.8}$	-1.26	-1.46	-0.87	-0.61
	$\lvert T_0 \rightarrow T_{0.8}\rvert$	4.35	5.42	3.83	2.45
qwen3-32b	$T_0 \rightarrow T_{0.6}$	-0.00	-0.10	-0.09	-0.13
	$\lvert T_0 \rightarrow T_{0.6}\rvert$	2.29	2.26	2.16	1.89

Table 6: Document-level averages of COMET score gains or losses ('$T_0 \rightarrow T_1$') and of their absolute values ('$|T_0 \rightarrow T_1|$' after excision of outputs with missing translations.

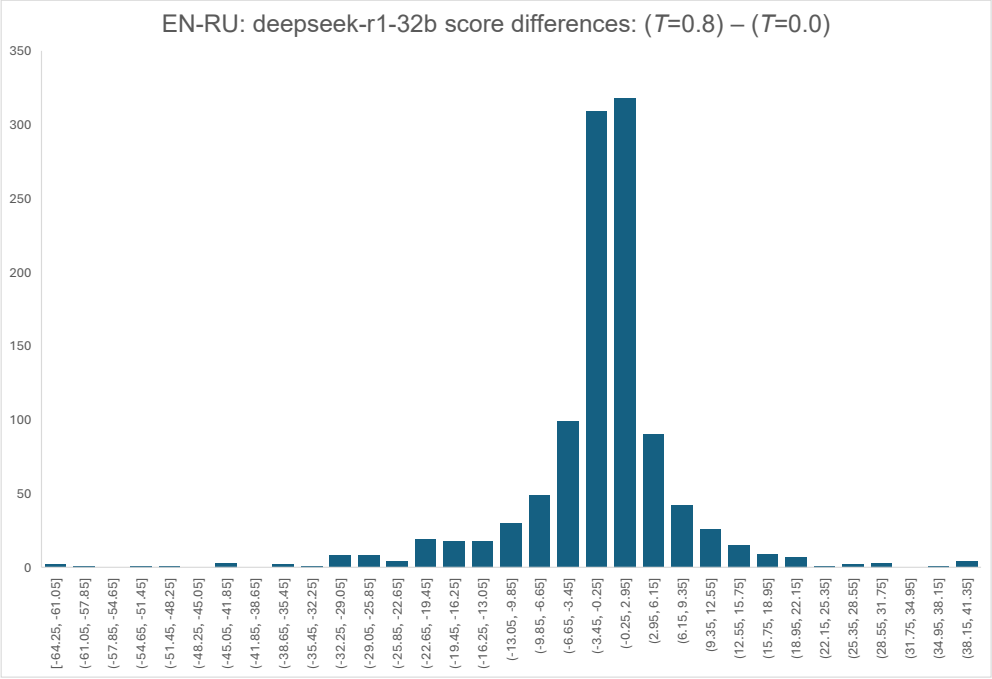

Figure 1: Histogram of sentence-level COMET score differences between the EN–RU outputs from deepseek-r1-32b for $T = 0.8$ and $T = 0.0$. Three empty lines were removed.

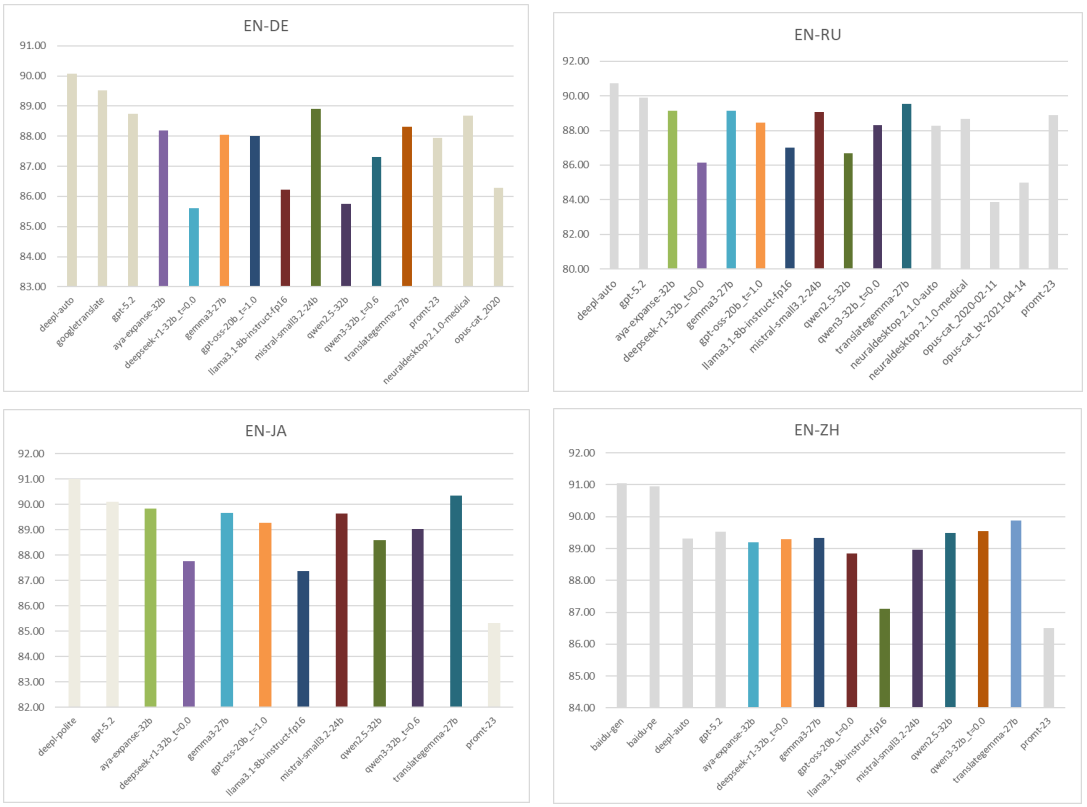

Figure 2: COMET-22 scores for translation outputs. Colored bars: best outputs from each of the nine local LLMs. Gray bars: available baselines.

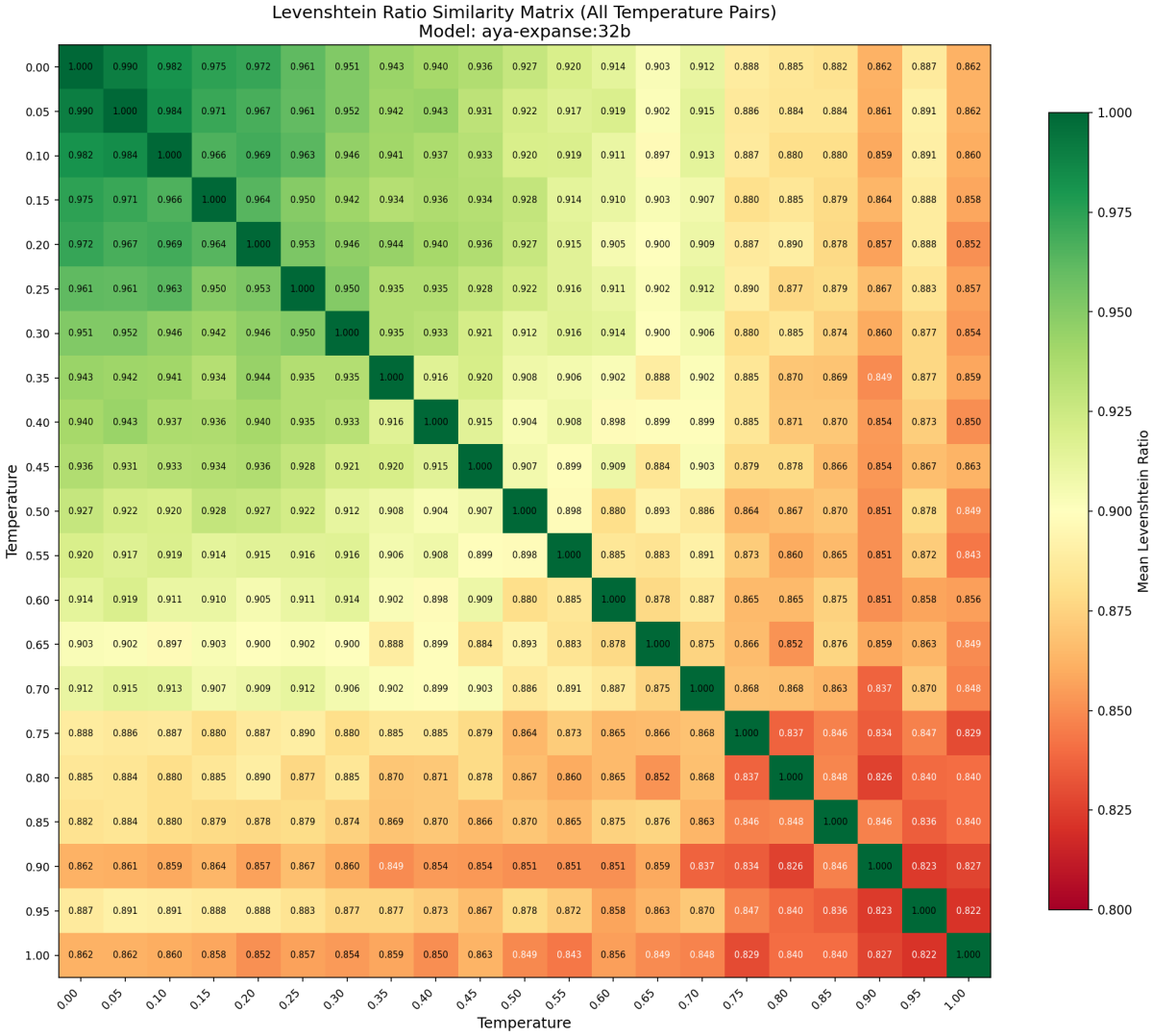

Figure 3: Levenshtein ratio similarity matrix for aya-expanse-32b across 21 temperature settings ($T = 0.0$ to $T = 1.0$, step 0.05), computed on the 118-sentence EN→DE pilot set. Each cell reports the mean Levenshtein ratio (computed over character sequences) between outputs generated at the two corresponding temperatures. Values range from 0 (no overlap) to 1 (identical outputs); the diagonal is trivially 1.0. The $T = 0.0$ row and column show the highest off-diagonal values, indicating that zero-temperature output is the centroid of the output distribution. Again these results suggest that temperature primarily affects output consistency rather than quality in translation; we therefore recommend users set temperature to 0.0.