

Translation Analytics for Freelancers:

I. Introduction, Data Preparation, Baseline Evaluations

Yuri Balashov
University of Georgia
Athens, Georgia, USA
yuri@uga.edu

Alex Balashov
Evariste Systems, LLC
Athens, Georgia, USA
abalashov@evaristesys.com

Shiho Fukuda Koski
SFK Language Solutions
Rochester, New York, USA

Abstract

This is the first in a series of papers exploring the rapidly expanding new opportunities arising from recent progress in language technologies for individual translators and language service providers with modest resources. The advent of advanced neural machine translation systems, large language models, and their integration into workflows via computer-assisted translation tools and translation management systems have reshaped the translation landscape. These advancements enable not only translation but also quality evaluation, error spotting, glossary generation, and adaptation to domain-specific needs, creating new technical opportunities for freelancers. In this series, we aim to empower translators with actionable methods to harness these advancements. Our approach emphasizes Translation Analytics, a suite of evaluation techniques traditionally reserved for large-scale industry applications but now becoming increasingly available for smaller-scale users. This first paper introduces a practical framework for adapting automatic evaluation metrics—such as BLEU, chrF, TER, and COMET—to freelancers’ needs. We illustrate the potential of these metrics using a trilingual corpus derived from a real-world project in the medical domain and provide statistical analysis correlating human evaluations with automatic scores. Our findings emphasize the importance of proactive engagement with emerging technologies to not only adapt but thrive in the evolving professional environment.¹

1 Introduction

This is the first in a series of papers exploring the rapidly expanding new opportunities arising from

© 2025 The authors. This article is licensed under a Creative Commons 4.0 licence, no derivative works, attribution, CC-BY-ND.

¹Data: <https://github.com/YuriBalashov/reeve-corpus>. Code: <https://github.com/abalashov/llm-translation-testbed/>.

recent progress in language technologies for individual translators and language service providers (LSPs) with modest resources.

1.1 Background and related work

Many translators use MT output in their workflow. In fact, MTPE (machine translation post-editing) has become the default *modus operandi* in the industry (Pérez, 2024) and is seamlessly integrated into computer-assisted translation (CAT) tools and translation management systems (TMS). (For a recent review, see Moorkens et al., 2025, Ch. 8.) Most CAT tools can now send real-time queries over the Internet (widely referred to as “API calls”) to any number of generally available neural machine translation (NMT) engines or MT aggregators and present the retrieved translation suggestions to the users for their consideration, alongside translation memory (TM) matches.

The advent of large language models (LLM) made the work environment of a typical freelancer more complex because, among other things, LLMs can translate, demonstrating performance competitive with that of dedicated NMT engines for some language pairs and domains (Castilho et al., 2023; Fernandes et al., 2023; Garcia et al., 2023; Hendy et al., 2023; Peng et al., 2023; Wang et al., 2023; Zhang et al., 2023; Peters and Martins, 2024; Li et al., 2024b; Li et al., 2024a; Lyu et al., 2024; Zhu et al., 2024). Even more importantly, with the right prompting, they can perform increasingly more sophisticated and advanced operations including, but not limited to:

- Evaluating the quality of translation output, including their own (Kocmi and Federmann, 2023; Lu et al., 2024), with or without reference translations.
- Spotting and categorizing translation errors and suggesting corrections (Berger et al., 2024; Feng et al., 2024).

- Automatic post-editing of raw MT output, including their own (Raunak et al., 2023; Ki and Carpuat, 2024; Alves et al., 2024; Rei et al., 2024).
- Adapting translation output:
 - to the required terminology (Ghazvininejad et al., 2023; Rios, 2024);
 - to a given domain (e.g. medical, legal, IT, aerospace engineering, etc.) (Sia and Duh, 2023; Zheng et al., 2024);
 - to existing translation memories and other project-, client- or domain-specific instructions and reference materials, often outperforming in these respects more traditional approaches earlier implemented in NMT systems (Moslem et al., 2023; Moslem, 2024; Vieira et al., 2024).
- Generating mono- and bilingual glossaries of special terms from pairs of source and target documents (Ding et al., 2025; Halpern, 2025)
- Improving the quality of translation in low-resource directions (e.g. DE-HI) by following a COT-style (“chain-of-thought,” Wei et al., 2023) prompt which explicitly requires them to pivot (“Translate this sentence from DE to EN first; then translate the EN output to HI”); see, in particular, Jiao et al., 2023.
- Following, with benefit, a human translation workflow (Chen et al., 2024; He et al., 2024) by engaging LLMs in a multi-turn interaction involving pre-translation research, drafting, refining, and proofreading (Briakou et al., 2024).

The possibilities in this area are virtually unlimited. Tech giants, larger LSPs, and MT aggregators are losing no time experimenting with these and other approaches in the context of massive localization workflows, with the goal of reducing the role of the proverbial “human expert in the loop” to the very minimum (see, e.g., Intento, Inc., 2024; Zekpa and Peter, 2025; RWS Group, 2025). CAT and TMS developers are hurrying to incorporate the latest LLM-powered features into their systems (e.g. memoQ, 2025; Bureau Works, 2025). New dedicated LLM-based applications are being offered to human translators,² sometimes premised on the assumption that translation memory is a depreciating asset.

²E.g. CotranslatorAI.

1.2 Our goals in this series of papers

There is no doubt that these trends will continue to shape the future of translation, human and machine, and will introduce numerous new and unforeseen changes to the fundamental nature of our work. Freelance translators, like everyone else, are adapting to the ongoing changes brought about by the latest developments in AI to the best of their ability. While this adaptation is crucial to the future of the profession, we submit that to get ahead of the curve, a more proactive approach is required.

Linguistic expertise has always been a distinctive mark of excellence in human translation work. However, freelancers are asked to perform other tasks such as sentence alignment, TM clean-up or glossary creation. In our own experience as translators, these tasks are growing in demand, which is consistent with anecdotal evidence from our colleagues and recent industry reports which emphasize “an increasing need for human translators to occupy new roles” (Crangasu, 2025), such as “AI Content Strategy,” “Big Data Curation,” or “QA Automation” (Da Fieno Delucchi et al., 2025). See also Slator, 2024; Al-Batineh and Al Tenaijy, 2024.

Freelancers are also increasingly asked to offer their advice on the quality of project- or domain-specific linguistic resources such as TMs or termbases (TB). Use cases include “a company looking to improve its AI translations,” a task that requires “experienced translators to pour through large volumes of the translated text” (Crangasu, 2025). A request to compare the relative quality of several candidate TMs for a given project is another good example of a task that would benefit from a novel combination of linguistic and technical knowledge. In some cases, pairwise automatic scoring of one TM against another, used as a reference, may be a good first step in the process. We believe that developing new technical skills proactively would make us better prepared for the upcoming challenges. To put it in slogan form, this could make a difference between the “AI is taking our jobs” and “AI is creating new opportunities for us” standpoints pervading much of the current discourse about AI.

Needless to say, many translators already have sophisticated technical capabilities. We think, however, that *Translation Analytics*—an umbrella category we shall use to refer to a variety of methods for the evaluation of the quality of translation-related linguistic assets—have not been deployed by free-

lancers to its full capacity. In fact, for most of them, ‘Translation Analytics’ may be synonymous with pre-translation analysis performed by CAT tools to generate the statistics for fuzzy TM matches at the start of a new project—for pricing, time planning, and other business purposes. Translation Analytics, however, are much broader in scope. We think of them as including, but not limited to:

- Human evaluation methods ranging from linear scoring to Multidimensional Quality Metrics (Lommel et al., 2013; Freitag et al., 2021; Knowles and Lo, 2024; Lommel et al., 2024).
- Automatic evaluation metrics, such as BLEU (Papineni et al., 2002), chrF (Popović, 2015), TER (Snover et al., 2006), and COMET (Rei et al., 2020).
- Any number of *ad hoc* tools and methods for statistical analysis and quality estimation that may be developed for a given project and tailored to its specific demands.

Our main goal in this series of papers is to explore the full potential of Translation Analytics in the context of a typical freelancer workflow. We aim to empower fellow translators with new methods that would allow them to add value to their services at the time of big changes and to gain control of the processes that usually happen “under the hood.” We also hope this will stimulate developers of CAT/TM tools and TMS systems to incorporate some of the analytic methods we describe in this series of papers into their products.

In the end, freelancers should be able to implement many of the sophisticated operations mentioned in Section 1.1 above, in their local translation environment, with practical, theoretical, and strategic benefits. Instead of contributing the last, indispensable but increasingly small, bit of human expertise to the proverbial “loop” the translator can get back into the driver’s seat by learning a small number of new technical skills.

1.3 Our goals in this first article

In the first article in this series we focus on adapting automatic evaluation metrics to the needs and work environment of individual translators and smaller LSPs who may want to take their technical capabilities to the next level.

Automatic evaluation of MT quality has been a prominent focus in the industry for years. Traditional metrics like BLEU (Papineni et al., 2002), chrF (Popović, 2015), and TER (Snover et al.,

2006) assess the output of MT systems by comparing it to reference translations, ideally created by skilled human translators. These comparisons rely on word or character-level string matching. Newer metrics such as COMET (Rei et al., 2020), BLEURT (Sellam et al., 2020), and BERTScore (Zhang et al., 2020) evaluate translations within the semantic space of neural networks. This approach is less reliant on specific word choices and instead prioritizes the underlying linguistic meaning.

The correlation between automatic metrics and human evaluation remains a topic of debates (for a recent overview of these debates, see Moorkens et al., 2025, Ch. 5), yet these metrics are essential in MT research and development. They enable developers to quickly compare model outputs after numerous adjustments to determine whether a particular change improves quality. Additionally, automatic metrics can monitor the training of NMT models by calculating, for instance, a BLEU score on a reserved test set after each iteration. Training can be stopped when no further improvement is observed.

Historically, automatic metrics were both technically complex and irrelevant to human translators, who depended on their linguistic expertise and manual analysis. However, with the seamless integration of MT engines into CAT tools, the vast availability of bilingual data at translators’ fingertips, and recent advancements in generative AI, the landscape is evolving rapidly. Many translators now incorporate MT into their workflows and often need to choose among multiple MT engines for specialized projects, sometimes spanning tens of thousands of words. Translators frequently possess valuable bilingual resources, such as TMs and TBs from similar projects, which allow them to evaluate MT engine outputs in minutes using automatic metrics. Free online tools designed for users without programming expertise facilitate this process.³ One such tool, MATEO (MACHINE Translation Evaluation Online) (Vanroy et al., 2023), is used in our work.

To illustrate the power and practical value of such methods for individual translators, we need high-quality data—parallel documents in two or more languages. While most of industrial-scale translation quality evaluation research is based

³And many other processes. Thanks to free online toolkits such as *adaptNMT* (Lankford et al., 2023), anyone can now build, train, fine-tune, and evaluate an NMT system more or less from scratch!

on the datasets made available on Workshops on Machine Translation (WMT) benchmarks (Kocmi et al., 2024) and other shared task repositories, we take our data from a recent real-life translation project completed in summer 2024 for a non-government organization, as described below.

Our main contributions detailed in this paper are as follows:

- We present, with the client’s permission, a **trilingual corpus of over 4.5K sentences in English, Russian, and Japanese in the medical domain (the Christopher & Dana Reeve Foundation Trilingual Corpus, RFTC)**, resulting from a recent human translation project completed by YB (EN-RU) and SFK (EN-JA) who are certified by the American Translators Association in their respective language pairs. We hope this corpus will be used for non-commercial research purposes by others and that it will grow both in coverage and language varieties.
- We use this corpus to develop and implement a relatively simple **approach to translation quality evaluation** which can be adapted by technically oriented translators and LSPs with modest resources to assess the quality of translation output from traditional NMT engines and LLMs in an informed way.
- We report the **BLEU**, **chrF2**, **TER**, and **COMET** scores for the translation outputs (EN-RU and EN-JA) for a slightly smaller but more challenging version of our corpus (about 3.5K English sentences) from **three popular NMT engines** (labeled MT1–MT3) and **three popular LLM models** (LLM1–LLM3), using our professional human translations for reference (see Appendix A for details).
- We adopt a simple linear 0.0–4.0 scale **modeled after academic grading** to perform preliminary **human evaluation** of **540** MT- and LLM-produced translations in each of our target languages (i.e. **1080** sentences overall).
- We report and discuss the results of our **preliminary statistical analysis** in order to determine:
 - whether the automatic scores computed for smaller non-overlapping parts of our source document (229, 1143, and 2183 sentences) correlate with each other;
 - whether sentence-level COMET scores for select segments for each output correlate with the human grades for them.

Two of the authors (YB and SFK) are ATA-certified professional translators with little or no programming experience or skills. Our perspective, therefore, fits the goals of this use case study. We should add that while we could, in principle, meet our coding needs by asking LLMs to write simple programs for our operations, coding with LLMs can be a haphazard process; the output can be very good and correctly focused on the problem, or can be mediocre and not especially applicable to what one is trying to accomplish. Stitching it all together for the purpose of a systematic study is still a task that benefits from a great deal of human expertise. Our experiments would be far from complete at this point without the tremendous help from a professional programmer on our team (AB) who took care of all the LLM-related operations, API call parallelization, streamlining, and more, as described in Section 3 below.

We adapt our discussion throughout the paper to the specific needs of individual translators and smaller LSPs. While the size of our corpus is small by MT industry standards, it is quite large for a single human translation project, and it generates statistically significant evaluation data. Furthermore, since our corpus is unlikely to have been seen and used for training or fine-tuning by generic NMT engines and popular LLMs at the time of conducting our baseline experiments, it adds new evidence for the ongoing debate about the quality and reliability of automatic quality metrics.

We believe that getting under the hood of translation quality evaluation is very important for freelancers and smaller LSPs at the time when traditional workflow models are being replaced by increasingly more sophisticated tasks requiring new technical expertise and willingness to learn more advanced methods. We submit that equipping individual translators with the additional technical capabilities described in this series of papers will help them adapt their toolkits to the rapidly changing work demands and new challenges brought about by the rocket speed development of language technologies.

The plan for the paper is as follows. Section 2 presents the Reeve Foundation Trilingual Corpus, complete with our reference translations. Section 3 describes how we obtained MT and LLM translation outputs for our source documents. In Section 4 we report and discuss the automatic metric scores for the entire corpus. In Section 5 we investigate pairwise correlations among the scores for three

smaller parts of the corpus. In Section 6 we develop our approach to human evaluation of select MT and LLM output, present its results, and discuss their statistical significance. In Section 7 we note the limitations of our study and outline plans for future work. Section 8 summarizes our findings and conclusions.

2 The Christopher & Dana Reeve Foundation Trilingual Corpus

We illustrate our Translation Analytics methods with the resources from a large translation-editing-proofreading project completed in summer 2024 for the Christopher & Dana Reeve Foundation.⁴ Specifically, the Foundation’s [Paralysis Resource Guide](#) is “a free comprehensive 392-page book designed to empower individuals living with and impacted by paralysis to lead healthy and fulfilling lives.” A shorter (80K words) international edition of the Guide was recently translated into several languages. The Guide (referred to below as ‘PRG’) is a coherent structured document divided into chapters and sections, complete with a descriptive glossary of about 200 technical terms. The translation project (EN-RU and EN-JA) came in the form of IDML (InDesign Markup Language) files for separate chapters. The PDF layouts of the EN, RU, and JA versions of PRG are included in our corpus for reference.

As the first step in data preparation, we took the versions of our TMs which preserve the order of source sentences in the original full document. We removed IDML and other tags from the TMs, discarded repetitions, and produced a spreadsheet that combined the source text (EN) and our reference translations (RU, JA). Next, we performed additional cleanup operations to remove:

- leading and trailing spaces;
- bullets and other special characters at the beginning of segments;
- segments with only or mostly numbers;
- segments with only or mostly URLs;
- segments with only or mostly address lines or phone numbers.

The resulting Excel file `1-10_en-ru-ja_long.xlsx` contains 4528 segments supplied with stable ID numbers (Column A), which are used in all our experiments.

⁴The authors thank the [Christopher & Dana Reeve Foundation](#) for a kind permission to use their linguistic resources in this work.

To make the translation task more challenging for MT engines and LLMs, we also decided to remove segments shorter than 6 source words from our set and generated “short” versions of the data (`1-10_en-ru-ja_short.xlsx`, etc.). The source sentence length (Len) is calculated in Column F.

An additional minor reduction was necessitated by the limitations MATEO imposes on the input file size ($\leq 1\text{MB}$) for evaluation (Section 4 below). To preserve the natural order of the segments, we met this requirement by removing the last two parts of PRG (“Glossary” and “Back Cover”), which brought the segment count down to 3555 (`1-8_en-ru-ja_short.xlsx`). The resulting Excel document was used to prepare tri- and monolingual Unicode text files for our experiments.

The materials referenced above comprise the Christopher & Dana Reeve Foundation Trilingual Corpus (alternatively, the Reeve Foundation Trilingual Corpus, RFTC), complete with the PDF layouts. Additional corpus details can be found in Appendix B. With the client’s permission, we make the corpus described here available for non-commercial/academic use.

3 Translation Outputs

In this section we describe how we obtained MT and LLM translation outputs for our corpus.

3.1 Technical notes on MT output

To preserve data confidentiality, we used the “Pro” versions of three popular NMT engines (labeled MT1, MT2, and MT3) to translate the entire `1-10_en_short.txt` document (3896 segments, one per line). The process was implemented as “pre-translation” in memoQ for MT1 and MT3, and was performed directly for MT2. We tracked the run-times for these operations (Table 5 below).

3.2 Technical notes on LLM output

Translation with LLMs was more complicated. API calls over the Internet must be used to interact with the major LLMs because the latter offer both the use of the model, and the significant parallel computing resources required to run it, as an integrated, metered “cloud” service. We used the Python programming language and the Python SDKs provided by major LLM vendors. We used paid subscription accounts for all LLM calls, with maximum data security/privacy settings allowed for these accounts.

3.3 Bulk processing and LLMs

There are, in principle, a number of ways to feed a large list of sentences to major LLMs. Some of them, for example, offer API constructs for batch processing, specifically intended for non-time-critical bulk tasks. In this approach, large data sets are uploaded for the LLM provider’s backend to churn through on a best-effort basis. To limit scope creep and eliminate variation in how we used different LLMs, we did not explore this option. It is also possible to submit multiple sentences with every request; this we did try, but we found the formatting characteristics of the resulting output to be too inconsistent for automatic evaluation. Therefore, the only method we evaluated was sentence-by-sentence, with one sentence per request.

It is worth taking a moment to reflect on the fact that this sentence-by-sentence approach is relatively naive, in a sense, even if it also eliminates some confounding factors. In contrast to the contextual environment of an ongoing ChatGPT conversation, in which the model keeps a running context window where prior prompts and responses reside, every one of our API requests instantiated a *de novo* context that was not informed by prior state. We did not attempt to evaluate the impact of context windows upon translation quality for two reasons: (1) the additional variables introduced would be unwieldy for the modest ambitions of this paper, and (2) some *ad hoc* experimentation did not suggest that there was much, if anything, to be gained in translation *quality* this way, and therefore it did not seem a propitious avenue for our specific aims. Still, this may be worth exploring in future research.

3.4 Prompt specificity

We found that brief and broad requests are not rewarded with as much consistency as long and specific ones. For example, when commanded to “translate the following sentence to Russian: _____” major LLMs would, for the most part, return the translated sentence and nothing else. However, every once in a while, the resulting sentence would contain additional verbiage: “Here is the following sentence in Russian: _____”.

With a more laborious prompt, which spelled out some examples of extraneous contributions unrelated to the translation of the sentence, this effect could be mostly, but not entirely obviated:

“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 sentence to Russian (Japanese): _____”

Rarely, extraneous output would still appear, although the prompt was highly effective at reducing the incidence of it. (We did not specifically attempt to measure the incidence.) This is a salient consideration for any endeavor that relies on low-touch bulk translation by LLMs.

3.5 Temperature and determinism

It is well known that LLMs’ output is not 100% deterministic. All LLM providers offer an API call parameter called “temperature” (T) which regulates the degree of acceptable stochastic variance in responses; higher temperatures allow more randomness, and lower values less. We set $T = 0.0$ in all of our requests across the board, but occasional variation in responses to identical prompts, while rare, was still present.

3.6 “Buggy” prompts

Upon completing all the operations with LLM and collecting all the outputs, we discovered that our optimized prompting routine concatenated the language name (i.e., ‘Russian’ or ‘Japanese’) to the prompt prefix twice:

“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 sentence to Russian (Japanese): Russian (Japanese): _____”

Given the length of the prompt, we hypothesized that this did not have a significant impact on the output. But we decided to perform a safety check comparing the LLM outputs for a shorter part of PRG (5.en.short, 229 segments) with the above prompt (which we used in our experiments) as well as with the corrected prompt:

“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 sentence to Russian (Japanese): _____”

We generated two outputs with the “bug-free” prompt to see if the differences between them due to the usual sampling (even with $T = 0.0$) in LLMs are significantly smaller than the differences between each of them and the output for the “buggy”

prompt. The results reported in Appendix C suggest that the answer is No. In terms of automatic scores, the differences among the three outputs are marginal and statistically insignificant, both for EN-RU and EN-JA. Interestingly, the “buggy” prompt actually did slightly better!

Tempting as it was to call it a feature not a bug, our safety check leads us to categorize it as insignificant and discardable statistical noise. We add this to the growing list of observations of rather unpredictable sensitivity of LLMs’ output to the fine details of the prompts in some cases, and their surprising robustness to prompt changes in other cases. We further hypothesize that LLMs’ insensitivity to the potentially misleading second occurrence of ‘Russian’ (or ‘Japanese’) in the “buggy” prompt may have to do with (i) their default preference for English; and/or (ii) their ability to identify the language of the string that actually follows ‘:’; and/or (iii) the fact that transformer-based neural networks, unlike the older LSTM- and GRU-based architectures, compute the attention scores between all pairs of tokens in the entire input directly and in parallel, rather than consecutively, so the fact that the second occurrence of the language name (‘Russian’ or ‘Japanese’) immediately precedes the source sentence does not make the former more important than the other preceding tokens.

We release all translation outputs from the systems we tested in the form of a single Excel file named `1-10_en-ru-ja_short_MT-LLM-outputs.xlsx`, where Column A contains the segment IDs, Column F the source segment length (in words), and the other columns are labeled with the target language and the system which generated the output.

4 Automatic Quality Evaluation

In this section we report and discuss the automatic metric scores for the entire corpus.

As already noted in Section 2 above, we had to reduce the length of our corpus by about 9% to 3555 segments to meet the file size requirements of MATEO (Vanroy et al., 2023), the tool we utilized to calculate the BLEU, chrF2, TER, and COMET scores for our outputs. We provide additional details in Appendix D.

The evaluation scores for 1-8_en are represented in Table 1 and Figure 1 below. Consistently with other reports, the string-based scores for EN-JA are lower than for EN-RU. We note, however, that the

COMET scores are neck-to-neck; in fact, slightly higher for EN-JA for all LLM outputs and MT3. All the score differences are statistically significant.

Since the distinction between the linguistic concepts of character and word is blurred in Japanese, questions may be raised about the separate significance of chrF for translation directions involving this language. We do not have a considered view on this. But we calculated pairwise Pearson correlation values for BLEU-chrF2, BLEU-TER, and BLEU-COMET between the scores for our six systems for both language pairs (Table 10). We note high correlations between BLEU and chrF2, and between BLEU and TER for both language pairs, a somewhat lower but still solid correlation between BLEU and COMET for EN-RU, and the lack of correlation between BLEU and COMET for EN-JA. Along with COMET’s neck-to-neck results for both language pairs, this underscores the importance of neural-based metrics.

In our experiments, performed on a 13th Gen Intel(R) Core(TM) i9-13900KF 3.00 GHz 64.0 GB PC, MATEO took roughly 20 minutes to compute the four scores for a single output against a reference; for JA it took slightly longer than for RU. Of these 20 minutes, roughly 16 minutes go into computing the COMET scores, 2 minutes into TER, and 2 minutes into bootstrap resampling at the very end. The calculation of BLEU and chrF2 is very fast. In light of the above-noted considerations, the time spent on computing the COMET scores is the time well spent. Users should be aware of this.

5 What Sample Size is Needed for Reliable Automatic Quality Evaluation?

Another important question that may arise for freelancers inclined to use automatic evaluation of MT/LLM outputs in choosing the best system for a new project is the minimal size of a sample required to make a reliable decision. A freelancer may have a good TM from a previous project in the same domain or for the same client that could be used for reference. Alternatively, a freelancer may complete a representative part of a new project and decide to add the best-performing MT and LLM-based system to their workflow going forward. One can imagine similar scenarios. Such deliberations should, of course, take into account typological differences between target languages which may affect the automatic scores for string- and neural-based metrics differently. In all cases of this sort,

the size of the sample to be used for MT/LLM translation quality evaluation must be statistically significant. What is the minimal size that meets this requirement?

To approach this question empirically we generated additional sets of automatic MT/LLM evaluation scores for the outputs from three distinct parts of our corpus, `229_en` (229 segments, identical to `5_en_short`), `1143_en` (1143 segments, identical to `3_en_short`), and `2183_en` (2183 segments comprising the rest of `1-8_en_short`) to see how well they correlate with each other. Tables 11 and 12 in Appendix E feature the four sets of scores, including those for `1-8_en_short` ($3555 = 229 + 1143 + 2183$ segments).

The lack of overlap among `229_en`, `1143_en`, and `2183_en` (cumulatively comprising the entire `1-8_en_short` document), which is evidenced in our memoQ analysis (Table 3) makes them suitable for correlation analysis, as does their thematic coherence: all three originate in a single narrow-domain document. Tables 13 and 14 (Appendix F) represent the Pearson correlation values r along with their p -values for three pairs of evaluation scores sets corresponding to `229_en`, `1143_en`, and `2183_en`.

We observe that the correlations are very strong in all cases, across all the metrics. We are thus led to conclude that computing automatic scores for a small part of our document ($229/3555 = 6.4\%$) would give us a good sense of the relative performance of several MT/LLM systems. However, this approach has its limitations. See Appendix F where we also provide additional details regarding the use of statistical methods for freelancers and discuss the prospects for future work.

6 Manual Evaluation of Select Translation Outputs

We selected 180 and 360 MT- and LLM-translated sentences from the outputs for `5_en_short` and `3_en_short` respectively for each of our language pairs (i.e. 1080 segments overall) to perform manual evaluation of their quality with a simple linear scale in order to estimate whether the outputs' sentence-level COMET scores correlate with our "human grades" for them. Below we outline our selection process, the evaluation scale, and the results.

6.1 Segment selection

We ranked MT- and LLM-generated translations by their sentence-level COMET scores and selected 10 highest-scoring segments, 10 intermediate-scoring segments, based on their median ranks, and 10 lowest-scoring segments from the outputs for `5_en_short`. We doubled these numbers (20-20-20) for `3_en_short`.

6.2 Human grading

To assign "human grades" to the selected translations we adopted a linear 0.0–4.0 scale modeled after academic grading (Table 15 in Appendix G). Two of the co-authors who have extensive academic teaching experience found this approach intuitive and efficient: it is easy for them to imagine they are grading student work. Along with the letter/numeric grades, we supplied brief notes for each graded translation highlighting 1–2 most serious issues from the following list: Accuracy; Clarity; Consistency; Fluency; Grammar (including spelling, typography, and syntax); Register; Style; Terminology; Tone. To minimize our bias in grading, we sorted these segments by their ID numbers rather than by their COMET scores.

While we are fully aware of the multiple limitations of this approach, our primary goal in this first round of baseline evaluations was to develop and offer to fellow translators a potentially fruitful method that would allow them to see whether automatic scores correlate with their human judgment in their particular use case.

6.3 Are automatic evaluation scores correlated with human grades?

We calculated Pearson (r) and Spearman (ρ) correlation coefficients between sentence-level COMET scores and our numeric grades for the 10-10-10 and 20-20-20 selections from each translation output (Tables 16 and 17 in Appendix H) for each language pair. Most of the r and ρ values suggest moderate to strong correlation; but the variation is rather wide, between MT/LLM outputs, language pairs, and sample sizes. Some of the variation may be an artifact of our somewhat impressionistic and non-rigorous grading and/or the sampling method. These may be adjusted depending on the available human resources. But calculating sentence-level correlations is a very natural and easy strategy to pursue in all cases where "human grades" of select outputs are available.

	COMET	BLEU	chrF2	TER
MT1	88.1	41.1	64.4	43.1
MT2	90.8	57.2	74.2	31.1
MT3	90.2	45.4	67.4	40.0
LLM1	88.8	38.4	63.5	45.4
LLM2	89.3	37.1	63.0	46.2
LLM3	88.6	33.2	60.1	50.1

English-Russian

	COMET	BLEU	chrF2	TER
MT1	88.1	31.1	39.5	55.3
MT2	89.7	38.6	46.0	47.5
MT3	90.6	36.8	44.1	49.7
LLM1	89.5	31.9	38.6	53.0
LLM2	90.1	30.2	37.6	53.9
LLM3	89.5	28.9	36.3	55.2

English-Japanese

Table 1: Evaluation metric scores for MT and LLM models for English-Russian and English-Japanese translations for 1-8_en_short.

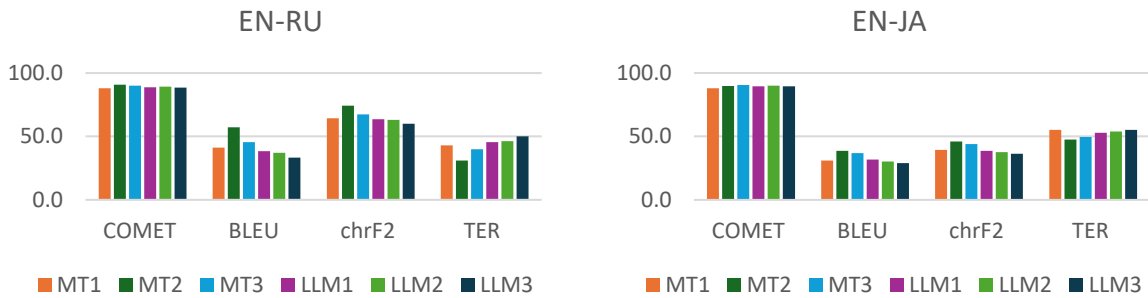


Figure 1: Visualization of MATEO-generated metric scores for EN-RU and EN-JA translations, broken down by MT engine and LLM, for 1-8_en_short.

We release all the selected sentences along with their COMET scores and our grades and comments in the form of two Excel files: 3_en-ru-ja_short_comet_grades.xlsx and 5_en-ru-ja_short_comet_grades.xlsx.

7 Limitations of Our Study and Future Plans

Translation directions. We had an opportunity to experiment with two interestingly different language pairs because we ourselves produced the translations for them and the client gave us permission to use them. We hope that other language directions and document families will be added to our corpus in the future.

Automatic metrics. We limited our choice of them to BLEU, chrF2, TER, and COMET, to maximize efficiency and ease of use. More advanced users should consider other metrics and consult the current best practices (Kocmi et al., 2024).

Correlation experiments with sample sizes reported in Section 5 need to be complemented with power analysis to determine the minimal size of the statistically significant sample. Ideally, future

experiments should also include other contrasting pairs—from different domains, registers etc.

Human evaluation requires independent raters and a uniform blinding and randomization protocol. While extensive, our reported results must be taken with a grain of salt. We do believe they serve as a proof of concept.

In future experiments with our corpus (RFTC) we want to explore the potential of various dedicated systems and LLMs for (i) extracting a bilingual glossary from a set of parallel sentences, and (ii) using a glossary thus obtained to improve the quality of translation in the context of the freelancer’s workflow.

Other ideas are briefly described in Section 1.1 above. We may pursue some of them and invite fellow translators and other interested parties to join us in this effort.

8 Conclusions

This study demonstrates the potential of Translation Analytics to help freelance translators and smaller language service providers (LSPs) thrive in a rapidly evolving industry. By adapting evaluation

metrics such as BLEU, chrF, TER, and COMET to individual workflows, we provide methods for assessing MT and LLM outputs with rigor and precision. The findings underscore several critical insights:

Utility of automatic evaluation metrics. Automatic metrics, particularly COMET, consistently align with human assessments, reinforcing their value as robust tools for translation quality evaluation. Translators can confidently leverage these metrics to make informed decisions about incorporating MT and LLM systems into their workflows.

Efficiency of sample-based evaluation. Even small, strategically selected samples of documents can yield statistically reliable insights into the relative performance of different translation systems. This approach enables resource-efficient evaluation for freelancers working on large-scale projects.

Integration of human judgment. While automatic metrics are helpful, the integration of human evaluation, anchored in linguistic expertise, remains critical. Our experiments validate the complementary roles of human judgment and automated tools in achieving nuanced and accurate quality assessments.

Empowering freelancers. By demystifying technical methods and tools, we equip translators with the confidence and skills to engage proactively with advanced language technologies. We hope this will help them move beyond being mere participants in the workflow to assuming leadership in optimizing and innovating translation practices. We offer one concrete entry point, with examples of expanded capabilities, in Appendix I.

Future work will focus on expanding the corpus to include additional language pairs, domains, and registers to further validate and refine our methods. Moreover, exploring advanced techniques such as glossary extraction, domain-specific adaptation, reference-free quality estimation, automatic post-editing, and more sophisticated multi-step operations using LLMs represents promising avenues for enhancing translation quality and efficiency.

As the landscape of translation continues to evolve, it is imperative for freelance translators and smaller LSPs to embrace new tools and methodologies. By doing so, they can not only adapt to the changes but also seize the opportunities presented by advancements in language technology. This proactive approach will ensure that translators remain at the forefront of a profession that is as dynamic as it is indispensable.

Author Contributions

YB developed the initial plan, prepared the data, ran the evaluation experiments, and wrote most of the content including literature review and bibliography, but excluding Sections 3.2–3.5 and Appendix I, which were contributed by AB, who took care of all our programming needs. YB and SFK are ATA-certified translators who worked with their partners in summer 2024 on translating the International Edition of the Reeve Foundation’s Paralysis Resource Guide to Russian and Japanese. They performed manual evaluation of the 1080 selected MT- and LLM-generated segment translations as described in Section 6. They also provided additional notes on the RFTC corpus in Appendix B. SFK curated the Japanese portion of the data.

Sustainability Statement

Our experiments performed on personal computers did not involve training of neural models. Computing COMET scores and querying LLMs for translation were the longest operations. We report the runtimes for them in Section 4 and Table 5.

We used one of the recommended algorithms⁵ to estimate a carbon impact of our computations according to (Lanelongue et al., 2021). A brief report is included in Appendix J.

Acknowledgments

YB’s work is supported by the NSF grant No. SES-233671. We are grateful to Bran Vanroy for helpful comments and clarifications on MATEO. We thank the reviewers for their very helpful comments. We reiterate our thanks to the Christopher & Dana Reeve Foundation for the permission to use their linguistic resources in our experiments.

References

- Mohammed Al-Batineh and Moza Al Tenaijy. 2024. *Adapting to technological change: An investigation of translator training and the translation market in the Arab world*. *Heliyon*, 10(7). Publisher: Elsevier.
- Duarte M. Alves, 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*. *Preprint*, arXiv:2402.17733.

⁵<https://calculator.green-algorithms.org/>

- Nathaniel Berger, Stefan Riezler, Miriam Exel, and Matthias Huck. 2024. [Prompting large language models with human error markings for self-correcting machine translation](#). In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 636–646, Sheffield, UK. European Association for Machine Translation (EAMT).
- Eleftheria Briakou, Jiaming Luo, Colin Cherry, and Markus Freitag. 2024. [Translating Step-by-Step: Decomposing the Translation Process for Improved Translation Quality of Long-Form Texts](#).
- Bureau Works. 2025. [Augmented translation actions: Enhancing translation efficiency with ai](#). Accessed: 2025-01-20.
- Sheila Castilho, Clodagh Quinn Mallon, Rahel Meister, and Shengya Yue. 2023. [Do online machine translation systems care for context? what about a GPT model?](#) In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 393–417, Tampere, Finland. European Association for Machine Translation.
- Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2024. [Iterative translation refinement with large language models](#). In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 181–190, Sheffield, UK. European Association for Machine Translation (EAMT).
- Adria Crangasu. 2025. [How is Artificial Intelligence Changing the Translation Services Industry?](#)
- Agustín Da Fieno Delucchi, Alfredo de Almeida, and Jorge Russo dos Santos. 2025. [How Language Industry Jobs Are “Shifting Left”](#). *Multilingual*, (January).
- Qiuyu Ding, Hailong Cao, Zihao Feng, Muyun Yang, and Tiejun Zhao. 2025. [Enhancing bilingual lexicon induction via harnessing polysemous words](#). *Neurocomputing*, 611:128682.
- Zhaopeng Feng, Yan Zhang, Hao Li, Bei Wu, Jiayu Liao, Wenqiang Liu, Jun Lang, Yang Feng, Jian Wu, and Zuozhu Liu. 2024. [Tear: Improving llm-based machine translation with systematic self-refinement](#). *Preprint*, arXiv:2402.16379.
- Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André Martins, Graham Neubig, Ankush Garg, Jonathan Clark, Markus Freitag, and Orhan Firat. 2023. [The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 1066–1083, Singapore. Association for Computational Linguistics.
- Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. [Experts, errors, and context: A large-scale study of human evaluation for machine translation](#). *Transactions of the Association for Computational Linguistics*, 9:1460–1474.
- Xavier Garcia, 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](#). *Preprint*, arXiv:2302.01398.
- Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. [Dictionary-based phrase-level prompting of large language models for machine translation](#). *Preprint*, arXiv:2302.07856.
- Jack Halpern. 2025. [Language resource action guide](#). Accessed: 2025-01-20.
- Zhiwei He, 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.
- Amr Hendy, 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](#). *Preprint*, arXiv:2302.09210.
- Intento, Inc. 2024. [Machine translation report 2024](#). Accessed: 2025-01-20.
- Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. [Is chatgpt a good translator? yes with gpt-4 as the engine](#). *Preprint*, arXiv:2301.08745.
- Dayeon Ki and Marine Carpuat. 2024. [Guiding large language models to post-edit machine translation with error annotations](#). *arXiv preprint arXiv:2404.07851*.
- Rebecca Knowles and Chi-kiu Lo. 2024. [Calibration and context in human evaluation of machine translation](#). *Natural Language Processing*, pages 1–25.
- Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinhórfur Steingrímsson, and Vilém Zouhar. 2024. [Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.
- Tom Kocmi and Christian Federmann. 2023. [Large language models are state-of-the-art evaluators of translation quality](#). In *Proceedings of the 24th Annual*

- Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.
- Muhammed Yusuf Kocyigit, Eleftheria Briakou, Daniel Deutsch, Jiaming Luo, Colin Cherry, and Markus Freitag. 2025. [Overestimation in LLM Evaluation: A Controlled Large-Scale Study on Data Contamination’s Impact on Machine Translation](#). *Preprint*: 2501.18771.
- Séamus Lankford, Haithem Afli, and Andy Way. 2023. [adaptNMT: an open-source, language-agnostic development environment for neural machine translation](#). *Language Resources and Evaluation*, 57(4):1671–1696.
- Loïc Lanelongue, Jason Grealey, and Michael Inouye. 2021. [Green Algorithms: Quantifying the Carbon Footprint of Computation](#). *Advanced Science*, 8(12):2100707.
- Chen Li, Meishan Zhang, Xuebo Liu, Zhaocong Li, Derek Wong, and Min Zhang. 2024a. [Towards demonstration-aware large language models for machine translation](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13868–13881, Bangkok, Thailand. Association for Computational Linguistics.
- Jiahuan Li, Hao Zhou, Shujian Huang, Shanbo Cheng, and Jiajun Chen. 2024b. [Eliciting the translation ability of large language models via multilingual finetuning with translation instructions](#). *Transactions of the Association for Computational Linguistics*, 12:576–592.
- Arle Lommel, Serge Gladkoff, Alan Melby, Sue Ellen Wright, Ingemar Strandvik, Katerina Gasova, Angelika Vaasa, Andy Benzo, Romina Marazzato Sparano, Monica Foresi, Johani Innis, Lifeng Han, and Goran Nenadic. 2024. [The multi-range theory of translation quality measurement: MQM scoring models and statistical quality control](#). In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 2: Presentations)*, pages 75–94, Chicago, USA. Association for Machine Translation in the Americas.
- Arle Richard Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2013. [Multidimensional quality metrics: a flexible system for assessing translation quality](#). In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.
- Qingyu Lu, Baopu Qiu, Liang Ding, Kanjian Zhang, Tom Kocmi, and Dacheng Tao. 2024. [Error analysis prompting enables human-like translation evaluation in large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8801–8816, Bangkok, Thailand. Association for Computational Linguistics.
- Chenyang Lyu, 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 *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1339–1352, Torino, Italia. ELRA and ICCL.
- memoQ. 2025. [memoq agt: Advanced machine translation technology](#). Accessed: 2025-01-20.
- Joss Moorkens, Andy Way, and Séamus Lankford. 2025. *Automating translation*. Routledge introductions to translation and interpreting. Routledge, Abingdon, Oxon ; New York, NY.
- Yasmin Moslem. 2024. [Language modelling approaches to adaptive machine translation](#). *Preprint*, arXiv:2401.14559.
- Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. [Adaptive machine translation with large language models](#). In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere, Finland. European Association for Machine Translation.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Keqin Peng, 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](#). *Preprint*, arXiv:2303.13780.
- Ben Peters and André F. T. Martins. 2024. [Did translation models get more robust without anyone even noticing?](#) *Preprint*, arXiv:2403.03923.
- Maja Popović. 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. Association for Computational Linguistics.
- Celia Rico Pérez. 2024. [Re-thinking machine translation post-editing guidelines](#). *The Journal of Specialised Translation*, 41:10–29.
- Vikas Raunak, Amr Sharaf, Yiren Wang, Hany Awadalla, and Arul Menezes. 2023. [Leveraging GPT-4 for automatic translation post-editing](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12009–12024, Singapore. Association for Computational Linguistics.
- Ricardo Rei, 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 *Proceedings of the Ninth Conference on Machine Translation*, pages 185–204, Miami, Florida, USA. 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.
- Miguel Rios. 2024. [Instruction-tuned large language models for machine translation in the medical domain](#). *Preprint*, arXiv:2408.16440.
- RWS Group. 2025. [Beyond words: Exploring the future of translation and localization](#). Accessed: 2025-01-20.
- Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. [BLEURT: Learning robust metrics for text generation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.
- Suzanna Sia and Kevin Duh. 2023. [In-context learning as maintaining coherency: A study of on-the-fly machine translation using large language models](#). In *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 173–185, Macau SAR, China. Asia-Pacific Association for Machine Translation.
- Slator. 2024. [How AI Impacts Jobs, Skills, and Tools for Localization Professionals](#). Accessed: 2025-04-03.
- Matthew Snover, 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. Association for Machine Translation in the Americas.
- Bram Vanroy, Arda Tezcan, and Lieve Macken. 2023. [MATEO: MACHine translation evaluation online](#). In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 499–500, Tampere, Finland. European Association for Machine Translation.
- Inacio Vieira, 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 *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 236–249, Chicago, USA. Association for Machine Translation in the Americas.
- Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. [Document-level machine translation with large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore. Association for Computational Linguistics.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. [Chain-of-thought prompting elicits reasoning in large language models](#). *Preprint*, arXiv:2201.11903.
- Narcisse Zekpa and Ajeeb Peter. 2025. [Evaluate large language models for your machine translation tasks on AWS](#). Accessed: 2025-01-20.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with bert](#). In *International Conference on Learning Representations*.
- Xuan Zhang, Navid Rajabi, Kevin Duh, and Philipp Koehn. 2023. [Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with QLoRA](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 468–481, Singapore. Association for Computational Linguistics.
- Jiawei Zheng, Hanghai Hong, Feiyan Liu, Xiaoli Wang, Jingsong Su, Yonggui Liang, and Shikai Wu. 2024. [Fine-tuning large language models for domain-specific machine translation](#). *Preprint*, arXiv:2402.15061.
- Haotian Zhu, Denise Mak, Jesse Gioannini, and Fei Xia. 2020. [NLPStatTest: A toolkit for comparing NLP system performance](#). In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 40–46, Suzhou, China. Association for Computational Linguistics.
- Wenhao Zhu, 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 *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

A NMT Engines and LLMs Used in our Experiments

- MT1 = ModernMT Professional
<https://www.modernmt.com/translate>
- MT2 = DeepL Translator Pro
<https://www.deepl.com/en/translator>
- MT3 = Google MT (Cloud Basic)

<https://translate.google.com>

LLM1 = GPT-4o

<https://platform.openai.com/docs/models/gpt-4o>

LLM2 = Claude 3.5 Sonnet

<https://console.anthropic.com>

LLM3 = Gemini 1.5 Pro

<https://ai.google.dev/gemini-ap>

We selected these engines and models at the time of conducting our baseline experiments (November 2024 – January 2025) based on a balance of the following considerations:

- their popularity among freelance translators and LSPs with limited resources;
- their subscription and per-token costs;
- their existing integration into CAT tools.

There are numerous other options available, including new NMT systems and the latest LLMs, and we plan to explore some of them in the future. We also believe that at the time of our initial experiments reported here, popular LLMs and NMT systems have not seen our trilingual data and, hence, could not have used it for re-training or fine-tuning. Now that this data is available, it might be of some interest to see if our chosen models' performance has changed (Kocuyigit et al., 2025).

B The Reeve Foundation Trilingual Corpus: Additional Details

The source document statistics for our corpus are compiled in Table 2 below. Table 3 presents memoQ analyses of both inputs, “long” and “short.” Table 4 provides further details of the corpus.

Although translations of the Reeve Foundation International Edition of the Paralysis Resource Guide (PRG) are intended to be generally available, their main target is the US population for whom English is a second language.

In the Russian translation of PRG, organization and program names and most of their acronyms were translated on their first occurrence followed by the English original and acronym in parentheses. In subsequent occurrences in the same section of the document, only translations or translated acronyms (where available) were used. Exceptions include acronyms such as ‘FDC’ and brand names of companies and their products, such as ‘Pfizer’ and ‘Tobii Dynavox’, which are kept in English. The brand medication names were translated or transliterated followed by their original English names on their first appearance. Only translations

were used on subsequent occurrences. Number notation generally follows Russian conventions, i.e. ‘33,000’ → ‘33 000’; ‘6.79’ → ‘6,79’; etc.

The Japanese translation of PRG generally adheres to the notation guidelines outlined in *JTF Style Guide for Translators Working into Japanese*. A polite and neutral style using the *desu/masu* form was applied, and the honorific suffix *san* was added after the names of individuals outside the Reeve Foundation. All personal names were transliterated. For medical terms, the original English term and its Japanese translation were juxtaposed in the headings of each section, separated by a slash, while only the Japanese versions were used in the body of the text. In the resource sections, organization names are presented in Japanese first, followed by the original English in parentheses. In the main body text, however, they are only in Japanese. When the source text includes abbreviations or acronyms that may be unfamiliar to Japanese readers, the full form is translated into Japanese. Physical and email addresses, URLs, and phone numbers are left in their original English form.

For typographic conventions, half-width characters are used for Arabic numerals, the percentage sign, slashes for fractions and acronyms, and colons (where unavoidable). Full-width characters are used for exclamation marks, question marks, Japanese middle dots, slashes (except in the cases mentioned above), ampersands, and parentheses. The UTF-8 encoding used in our experiments preserves all the relevant features of Japanese grammar and notation.

C Runtime Details

C.1 MT and LLM translation runtimes and costs

The available runtime and cost details for our translation operations are provided in Table 5 below.

C.2 “Bug-free” vs. “buggy” prompts

As noted in Section 3.6, we generated two sets of LLM outputs for 5_en_short (229 segments) with the “bug-free” prompt to compare them with the outputs for the “buggy” prompt and computed their automatic scores with MATEO. See Tables 6 and 7 below.

D Automatic Quality Evaluation Details

As noted in Section 1.3, evaluating the quality of MT output is a central concern in research and

development. Automatic MT quality metrics are tools that help measure how good an MT-translated sentence is, typically by comparing it to one or more human reference translations. While the field is rapidly evolving, several widely used metrics include BLEU, chrF, TER, and COMET. These can be broadly categorized into *string-based* and *neural-based* metrics.

String-based metrics evaluate translations by comparing the surface forms—words or characters—of MT output and reference human translations. BLEU (BiLingual Evaluation Understudy, Papineni et al., 2002) is one of the earliest and most well-known metrics. BLEU calculates how many n -grams (word sequences) in the MT output match those in the reference. While useful, it can be overly strict, penalizing valid translations that use synonyms or different phrasing. chrF (Character F-score, Popović, 2015), on the other hand, operates at the character level, making it more sensitive to morphologically rich languages and spelling. It computes F -scores based on overlapping character n -grams, which helps in capturing partial matches more effectively. TER (Translation Edit Rate, Snover et al., 2006) metric measures the number of edits (insertions, deletions, substitutions, and shifts) needed to change the MT output into the reference translation. A lower TER indicates better translation quality. It gives a more intuitive sense of the editing effort required.

Neural-based metrics leverage LLMs and machine learning techniques to evaluate translations more like humans do. These models can understand meaning beyond surface similarity. One such metric, used in this paper, is COMET (Crosslingual Optimized Metric for Evaluation of Translation, Rei et al., 2020). Built on pre-trained neural models and fine-tuned on human quality assessments, it can capture semantic similarity and fluency better than traditional metrics, even when there is little word overlap. It has been shown to correlate better with human judgments. Our results reported in Section 6.3 are consistent with this claim.

Machine translation evaluation is a fast-moving area of research, with new methods and tools emerging regularly. A great place to stay updated is <http://www.machinetranslate.org>, which offers accessible summaries of research, tools, and best practices in the field.

As noted in Section 1.3, one exciting development for practitioners is that freelance translators and non-specialists can now use web-based

tools to evaluate MT output themselves. MATEO (Vanroy et al., 2023), employed in our work, is a user-friendly Streamlit-based platform that allows anyone to calculate multiple MT quality metrics—including BLEU, chrF, TER, and COMET—without needing technical knowledge.

There is thus no mystery to MT quality evaluation. These metrics, whether simple or complex, are just tools to help us understand how well an NMT engine or an LLM has translated a piece of text. As the tools become more accessible and sophisticated, translators and content creators are empowered to make informed decisions about using and improving MT output.

A point of caution: when utilizing MATEO, or any other toolkit, for MT evaluation, it is crucial to select the appropriate *metric configurations* to ensure accurate and meaningful results. While the default settings in MATEO are designed to be practical for a wide range of target languages, evaluating translations into morphologically rich languages, such as Japanese or Korean, requires special attention. These languages exhibit complex word forms and inflections that standard metric configurations might not fully capture.

MATEO allows one to make the necessary changes in the “Metric selection” section of the application. In our case, evaluation of the outputs in Japanese required changing the default tokenization setting in BLEU to `ja-mecab`, and enabling `asian-support` in TER. We also found it useful to enable the normalized mode in TER, which is set to ‘False’ by default.

We report the configurations we used for the automatic evaluation metrics for both our target languages in Tables 8 and 9.

E Evaluation Scores for Sub-Documents of Different Sizes

See Tables 11 and 12 below, which represent the four sets of automatic metric scores for three non-overlapping parts of 1–8 en short along with the whole: $3555 = 228 + 1143 + 2183$ segments.

F Pearson Correlations for Three Pairs of Score Value Sets Across Translation Systems

Pearson’s correlation coefficient (r) is a statistical measure that quantifies the strength and direction of

the linear relationship between two variables, helping to determine whether changes in one variable are associated with changes in another.

The formula for calculating Pearson’s correlation is:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

where x_i and y_i are individual data points, \bar{x} and \bar{y} are their means, and n is the number of data pairs. In essence, r measures how much two variables change together relative to how much they change individually. The numerator represents the covariance between the variables, while the denominator normalizes this value using the standard deviations of both variables.

The value of r always lies between -1.0 (corresponding to perfect negative correlation) and 1.0 (perfect positive correlation). Typical guidelines for interpreting the values of r are as follows:

r	Correlation
0.9 to 1.0 or -0.9 to -1.0	Very strong
0.7 to 0.9 or -0.7 to -0.9	Strong
0.5 to 0.7 or -0.5 to -0.7	Moderate
0.3 to 0.5 or -0.3 to -0.5	Weak
0.0 to 0.3 or 0.0 to -0.3	Very weak or none

The p -value associated with Pearson’s correlation r estimates the statistical significance of the observed correlation. A p -value is the probability that the actual distribution of the data points would occur by random chance. A low p -value (typically < 0.05) suggests that the result is statistically significant.

In our case, the variables in questions are pairwise metric-specific scores which are reflected in the rows of Tables 11 and 12. For example, the first two rows in Table 11 (for EN-RU) show the COMET scores for 229_en and 1143_en across our six MT/LLM systems. Accordingly: $x_1 = 87.7, x_2 = 91.1, x_3 = 90.0, x_4 = 88.7, x_5 = 89.6, x_6 = 88.9, y_1 = 89.4, y_2 = 91.4, y_3 = 90.8, y_4 = 89.7, y_5 = 90.1, y_6 = 89.1$, yielding $r = 0.891; p = 0.0077$, reflected in the last column of Table 13.

Thus Tables 13 and 14 below display the correlation coefficients and their p -values for three

pairs of score value sets for the outputs from our range of six translation systems (i.e. MT1–MT3 and LLM1–LLM3), in both language directions.

The plots in Figure 2 provide the additional details of the distribution of our “data points” across MT1–3 and LLM 1–3.

As we noted in Section 5, all the pairwise Pearson correlations for our three non-overlapping sub-documents are very strong and statistically significant thus highlighting the consistency and stability of the rankings of our MT/LLM outputs across sub-documents of different sizes. If we wanted to select one or two best performing systems based on the automatic evaluation scores for our project, we could simply pick out the shortest chapter of PRG (i.e. Chapter 5 = 229_en) and treat it as a good representative of the entire document.

Even this shortest sample has over 4,000 source words, which exceeds the average daily output of a typical translator. It would be interesting to trim down the sample size even more to determine the point at which the correlation is lost and the scores become unreliable. The best way to do this is to perform a power analysis using one of the available toolkits (e.g. Zhu et al., 2020). It would also be desirable to include contrasting pairs of data points from different translation domains and registers. We leave it for further work.

Translators interested in implementing correlation or more advanced statistical analyses can use any number of generally available tools, from Excel to Python or R libraries. In our experience, LLMs can generate simple standalone Python scripts for such purposes, in response to sufficiently detailed prompts.

G Manual Grading Scale

Our manual scale modeled after academic grading is displayed in Table 15.

H Correlation Between Sentence-Level COMET Scores and Numeric Human Grades

Tables 16 and 17 represent Pearson and Spearman correlation between sentence-level COMET scores and our numeric human grades.

As noted above (Appendix F), Pearson correlation measures the strength and direction of a *linear* relationship between two variables. It assumes that both variables are normally distributed, and that the relationship is linear. This approximation

was adequate for six pairs of data points representing the scores for the outputs of our MT/LLM systems. But the number of our chosen sentence-level COMET scores and the corresponding human grades is larger: 30 or 60. In such cases Pearson correlation may be insufficient, especially if the relationship between variables is non-linear or if the data contains outliers. In such cases, adding Spearman correlation (ρ) can provide a more accurate picture of the association by focusing on the rank-order rather than precise values. Spearman correlation is a non-parametric measure that assesses how well the relationship between two variables can be described by a *monotonic* function. It uses the ranked values of the data, not the raw values, so it doesn't assume normality or linearity.

I APIs and Applied Technical Avenues for Freelancers

We acknowledge that most freelance translators are not programmers. However, as discussed elsewhere in this paper, we believe the future of translation work demands skills that are more conducive to the building blocks of machine intelligence and automation.

As a practical matter, the major LLM providers expose use of their models in two ways: a human-friendly way, via an interactive “chatbot” interface, and a machine-friendly way, via REST (REpresentational State Transfer) APIs, or Application Programming Interfaces. Despite the imposing weight of these acronyms to non-technical readers, the chasm between these modes of interaction is not, in fact, so vast. REST APIs use HTTP, the building-block protocol of the World Wide Web, as a transport, and a series of HTTP chatbot “verbs” whose meaning is not especially obscure: GET, POST, DELETE, and so on.

Contemporary REST APIs customarily encode information in a lightweight, human-readable encapsulation structure known as **JSON** (JavaScript Object Notation). The primary purpose of JSON is to define a hierarchical relational structure—for instance, to distinguish an object from its attributes.

A simple JSON structure might look like this:

```
{
  "people": [
    "Alex": {
      "org": "Evariste_Systems",
      "phd": false
    },
    "Yuri": {
```

```
      "org": "University_of_Georgia",
      "phd": true
    }
  ],
  "paper_type": "edifying",
  "lucky_numbers": [1, 7, 10]
}
```

The SDKs (Software Development Kits) of major LLM providers abstract away lower-level programmatic REST API interactions, which is more ergonomic for the software engineers using them. However, the LLM APIs can be directly queried, with the help of user-friendly tools such as [Postman](#). The interactive “chatbot” clients to which many readers will be well-accustomed are little more than simplified front-ends to these REST APIs.

Perusing the content of JSON responses from the major LLM providers’ REST APIs can open one’s mind to new possibilities. For example, one of the authors used Postman to prompt OpenAI’s GPT-4o model thus:

“You are a highly competent Russian to English translator. How would you explain the Russian concept of a ‘matryoshka’ in English? Please be brief.”

On the surface, the reply received was unremarkable:

```
{
  "id":
    "chatcmpl-BLGsJZWEWEfIBcKs46cXpNyyTfnY",
  "object": "chat.completion",
  "created": 1744409455,
  "model": "gpt-4o-2024-08-06",
  "choices": [
    {
      "index": 0,
      "message": {
        "role": "assistant",
        "content": "A_matryoshka,_also_known_as_a_Russian_nesting_doll,_is_a_set_of_wooden_dolls_of_decreasing_size_placed_one_inside_another._Each_doll_splits_in_half_at_the_middle_to_reveal_a_smaller_doll_inside,_symbolizing_themes_of_motherhood,_family,_and_continuity.",
        "refusal": null,
        "annotations": []
      },
      [...]
    }
  ]
}
```

However, after perusing [OpenAI chat API reference](#), the author learned that it is possible to supply the JSON attributes:

```
"logprobs": true,
"top_logprobs": 3
```

to the request, which tells OpenAI to share two other alternative probabilistic paths not taken for every generated token.

Thus, although GPT-4o began this generated response with the article ‘A’, it considered alternatives:

```

"logprobs": {
  "content": [
    {
      "token": "A",
      "logprob": -0.011159946210682392,
      "bytes": [
        65
      ],
    },
    {
      "token": "The",
      "logprob": -4.511159896850586,
      "bytes": [
        84,
        104,
        101
      ],
    },
    {
      "token": "In",
      "logprob": -9.636159896850586,
      "bytes": [
        73,
        110
      ],
    }
  ],
}

```

The author fed this probability output for the first few tokens into Anthropic’s Claude Sonnet model and asked it to generate a flowchart, using the following prompt:

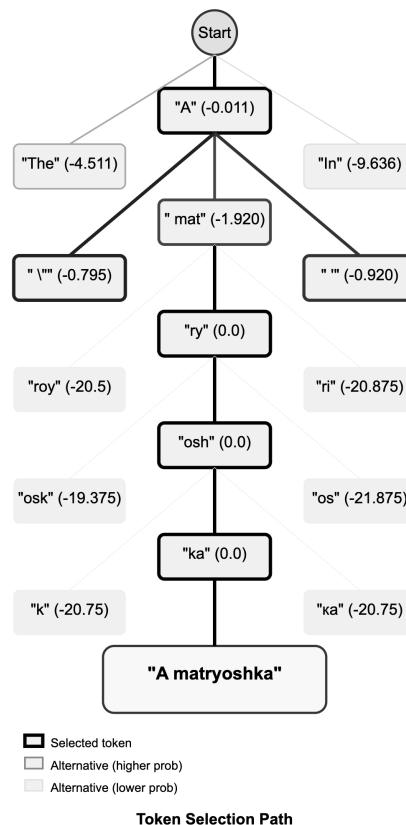
“This JSON file contains a ‘logprobs’ element from OpenAI’s API, which shows two other probabilistic responses considered by the model before returning the one with the lowest absolute value. Could you please draw a flowchart for the first five ‘logprobs’ elements which illustrates the traversal path taken? Please encapsulate every one of the five generated tokens in a rectangle, and use a darker or solid line to indicate the path actually taken based on the lowest absolute value of the ‘logprob’ entry, while

using lighter lines, their lightness in proportion to the relative absolute value of the ‘logprob’ value, to show alternatives not taken.”

The author then further prompted Claude to flatten the graphic for ease of inclusion here:

“Could you refine this flowchart to be more vertical, so that it is easier to incorporate into a two-column document without overflow beyond the margins?”

This was the result:



A key idea here is that this interesting foray would have never occurred to the author without digging into the OpenAI APIs and interacting with them directly. The commonplace interactive chatbot interfaces do not surface these possibilities to the end-user. No code was written for this exercise, just some slight tweaks to minimal, easy-to-read JSON data structures.

First and foremost, we believe that becoming conversant with the API surface of the major LLM providers can empower freelancers to make more specific and technically articulate LLM integration demands of the vendors of their preferred translation software tools. Second, we can reasonably speculate that the creation or enhancement of tools

reliant on API integrations may drift further away from the exclusive province of professional programmers, and become more reachable for technically minded end-users. This trend can be extrapolated from an ongoing trajectory to which veteran software engineers are privy: service APIs offered over the Internet have become far less arcane and easier to decipher over time, between the simplified vocabulary of REST and the human-readable wire format of JSON, for example.

J Carbon Imprint

We used the Green Algorithm developed in ([Lanlongue et al., 2021](#)) to estimate the carbon imprint of our computations performed on two computers (Figure 3 and Figure 4 below).

	1-10_en_long	1-10_en_short	1-8_en_short
Segments	4528	3896	3555
Words tokens (no punc)	76,553	74,667	68,989
Word types (no punc)	10,689	10,325	9,821
Characters (w/o \r\n)	500,347	485,830	448,652
Type/token ratio	0.14	0.14	0.14
MTLD	100.18	101.44	100.77
Average segment length (words)	16.91	19.17	19.41
Average word length (characters)	5.42	5.38	5.38

Table 2: Source document statistics.

Type	Segments.	Source words	Source chars	Source tags	Percent
1-10_en_short					
All	3896	74683	415016	0	100
X-translated / double context	0	0	0	0	0
Repetition	0	0	0	0	0
101%	0	0	0	0	0
100%	0	0	0	0	0
95%–99%	3	31	227	0	0.04
85%–94%	6	118	696	0	0.16
75%–84%	17	251	1703	0	0.34
50%–74%	201	3080	18615	0	4.12
No match	3669	71203	393775	0	95.34
1-8_en_short					
All	3555	68999	383178	0	100
X-translated / double context	0	0	0	0	0
Repetition	0	0	0	0	0
101%	0	0	0	0	0
100%	0	0	0	0	0
95%–99%	3	31	227	0	0.04
85%–94%	6	118	696	0	0.17
75%–84%	16	239	1636	0	0.35
50%–74%	172	2700	16465	0	3.91
No match	3358	65911	364154	0	95.52

Table 3: MemoQ analysis of the “long” and “short” PRG inputs. Fuzzy matches result from the *Homogeneity* feature of the memoQ analysis which measures internal similarities within a set of documents by adding each segment to a temporary TM and using it for lookup for every subsequently processed segment. For details, see [here](#).

Document Name	Excel/PDF	Notes
PARALYSIS RESOURCE GUIDE (PRG): International Edition	1-10_en-ru-ja_long.xlsx 1-10_en-ru-ja_short.xlsx PRG-IntEd_en.pdf PRG-IntEd_ru.pdf PRG-IntEd_ja.pdf	4528 segments (ID: 3–4687) 3896 segments (ID: 13–4687)
PRG: Chapters 1-6	1-8_en-ru-ja_short.xlsx	3555 segments (ID: 13–4299)
PRG: Front Cover		No segments
PRG: Introduction	2_en-ru-ja_short.xlsx	21 segments (ID: 13–45)
PRG: Chapter 1	3_en-ru-ja_short.xlsx	1143 segments (ID: 108–1430)
PRG: Chapter 2	4_en-ru-ja_short.xlsx	1237 segments (ID: 1437–2897)
PRG: Chapter 3	5_en-ru-ja_short.xlsx	229 segments (ID: 2903–3178)
PRG: Chapter 4	6_en-ru-ja_short.xlsx	228 segments (ID: 3186–3467)
PRG: Chapter 5	7_en-ru-ja_short.xlsx	583 segments (ID: 3473–4156)
PRG: Chapter 6	8_en-ru-ja_short.xlsx	114 segments (ID: 4164–4299)
PRG: Glossary	9_en-ru-ja_short.xlsx	335 segments (ID: 4306–4673)
PRG: Back Cover	10_en-ru-ja_short.xlsx	6 segments (ID: 4675–4687)

Table 4: Document details for the Paralysis Resource Guide (PRG).

		Runtime	Total cost	Notes
MT1	EN-RU	00:02:17	n/a	
	EN-JA	00:02:32	n/a	
MT2	EN-RU	00:01:28	n/a	
	EN-JA	00:01:05	n/a	
MT3	EN-RU	00:08:36	n/a	
	EN-JA	00:08:03	n/a	
LLM1	EN-RU	00:15:38	USD 7.01	Combined EN-RU and EN-JA
	EN-JA	00:15:28		
LLM2	EN-RU	02:29:43	USD 23.22	Combined EN-RU and EN-JA
	EN-JA	02:38:38		
LLM3	EN-RU	00:11:25	n/a	
	EN-JA	00:11:25		

Table 5: MT and LLM translation runtimes and costs for 1-10_en_short (3896 segments).

	COMET	BLEU	chrF2	TER
“Bug Free-1”	89.1 ± 0.4	33.6 ± 1.7	60.7 ± 1.2	49.6 ± 1.5
“Bug Free-2”	89.0 ± 0.5	33.7 ± 1.7 ($p = 0.23$)*	60.7 ± 1.1 ($p = 0.33$)*	49.6 ± 1.5 ($p = 0.33$)
“Buggy”	89.1 ± 0.5 ($p = 0.42$)	34.1 ± 1.7 ($p = 0.08$)	61.0 ± 1.2 ($p = 0.04$)*	49.3 ± 1.5 ($p = 0.09$)

Table 6: LLM1–3 combined outputs for 5_en_short: English-Russian.

	COMET	BLEU	chrF2	TER
“Bug Free-1”	89.7 ± 0.4	29.8 ± 1.3	36.8 ± 1.3	52.0 ± 1.3
“Bug Free-2”	89.7 ± 0.4 ($p = 0.34$)	29.6 ± 1.3 ($p = 0.15$)	36.4 ± 1.2 ($p = 0.05$)*	52.2 ± 1.3 ($p = 0.09$)
“Buggy”	89.9 ± 0.4 ($p = 0.05$)*	30.5 ± 1.3 ($p = 0.02$)*	37.4 ± 1.3 ($p = 0.04$)*	51.6 ± 1.3 ($p = 0.05$)

Table 7: LLM1–3 combined outputs for 5_en_short: English-Japanese.

Metric	Details
BLEU	nrefs:1 bs:1000 seed:12345 case:mixed eff:no tok:13a smooth:exp version:2.3.1 mateo:1.1.3
chrF2	nrefs:1 bs:1000 seed:12345 case:mixed eff:yes nc:6 nw:0 space:no version:2.3.1 mateo:1.1.3
TER	nrefs:1 bs:1000 seed:12345 case:lc tok:tercom norm:yes punct:yes asian:no version:2.3.1 mateo:1.1.3
COMET	nrefs:1 bs:1000 seed:12345 c:Unbabel/wmt22-comet-da version:2.0.1 mateo:1.1.3

Table 8: Metrics configurations for English-Russian.

Metric	Details
BLEU	nrefs:1 bs:1000 seed:12345 case:mixed eff:no tok:ja-mecab-0.996-IPA smooth:exp version:2.3.1 mateo:1.1.3
chrF2	nrefs:1 bs:1000 seed:12345 case:mixed eff:yes nc:6 nw:0 space:no version:2.3.1 mateo:1.1.3
TER	nrefs:1 bs:1000 seed:12345 case:lc tok:tercom norm:yes punct:yes asian:yes version:2.3.1 mateo:1.1.3
COMET	nrefs:1 bs:1000 seed:12345 c:Unbabel/wmt22-comet-da version:2.0.1 mateo:1.1.3

Table 9: Metrics configurations for English-Japanese.

		BLEU-chrF2	BLEU-TER	BLEU-COMET
EN-RU	<i>r</i>	0.998	-0.999	0.806
	<i>p</i>	<0.0001	<0.0001	0.0345
EN-JA	<i>r</i>	0.990	-0.969	0.388
	<i>p</i>	<0.0001	0.0002	0.4321

Table 10: Pearson correlations (r) for BLEU-chrF2, BLEU-TER, and BLEU-COMET for 1-8_en.

Metric	Label	Sgmts	MT1	MT2	MT3	LLM1	LLM2	LLM3
COMET	229_en	229	87.7	91.0	90.0	88.7	89.6	88.9
	1143_en	1143	89.4	91.4	90.8	89.7	90.1	89.1
	2183_en	2183	87.5	90.5	89.9	88.4	88.9	88.3
	1-8_en	3555	88.1	90.8	90.2	88.8	89.3	88.6
BLEU	229_en	229	37.2	57.7	43.6	34.6	35.3	32.3
	1143_en	1143	45.8	60.1	49.0	42.5	40.8	36.2
	2183_en	2183	39.0	55.5	43.6	36.5	35.3	31.8
	1-8_en	3555	41.1	57.2	45.4	38.4	37.1	33.2
chrF2	229_en	229	62.2	74.7	65.6	61.5	62.3	59.4
	1143_en	1143	68.3	76.4	70.4	67.0	66.2	62.9
	2183_en	2183	62.5	73.0	66.0	61.7	61.4	58.6
	1-8_en	3555	64.4	74.2	67.4	63.5	63.0	60.1
TER	229_en	229	46.0	30.9	42.3	48.7	47.4	51.6
	1143_en	1143	38.7	28.5	36.5	41.4	42.7	47.3
	2183_en	2183	45.2	32.6	41.6	47.2	47.9	51.4
	1-8_en	3555	43.1	31.1	40.0	45.4	46.2	50.1

Table 11: Evaluation scores for documents of different sizes: English-Russian.

Metric	Label	Sgmts	MT1	MT2	MT3	LLM1	LLM2	LLM3
COMET	229_en	229	88.0	90.8	90.3	89.6	90.1	89.8
	1143_en	1143	88.6	89.8	90.8	89.7	90.3	89.8
	2183_en	2183	87.8	89.6	90.5	89.4	90.0	89.4
	1-8_en	3555	88.1	89.7	90.6	89.5	90.1	89.5
BLEU	229_en	229	30.8	36.3	35.7	31.3	30.3	29.8
	1143_en	1143	31.0	35.1	36.7	32.0	29.5	28.9
	2183_en	2183	31.2	40.3	36.9	31.9	30.6	28.9
	1-8_en	3555	31.1	38.6	36.8	31.9	30.2	28.9
chrF2	229_en	229	38.7	43.5	42.7	37.7	37.6	36.7
	1143_en	1143	38.7	42.1	43.4	37.6	36.5	35.4
	2183_en	2183	40.0	47.8	44.6	39.3	38.2	36.8
	1-8_en	3555	39.5	46.0	44.1	38.6	37.6	36.3
TER	229_en	229	54.1	49.2	48.2	50.9	51.4	52.7
	1143_en	1143	53.4	48.8	48.3	51.9	52.6	53.5
	2183_en	2183	56.5	47.0	50.7	53.9	54.9	56.5
	1-8_en	3555	55.3	47.5	49.7	53.0	53.9	55.2

Table 12: Evaluation scores for documents of different sizes: English-Japanese.

EN-RU	Correlation pairs		
	229 / 2183	1143 / 2183	229 / 1143
COMET			
<i>r</i>	0.978	0.942	0.891
<i>p</i> -value	0.0001	0.0014	0.0077
BLEU			
<i>r</i>	0.991	0.994	0.973
<i>p</i> -value	< 0.0001	< 0.0001	0.0001
chrF2			
<i>r</i>	0.990	0.990	0.966
<i>p</i> -value	< 0.0001	< 0.0001	0.0003
TER			
<i>r</i>	0.992	0.990	0.971
<i>p</i> -value	< 0.0001	< 0.0001	0.0002

Table 13: Pearson correlations (r) and p -values for EN-RU for three pairs of score value sets across six translation systems.

EN-JA	Correlation pairs		
	229 / 2183	1143 / 2183	229 / 1143
COMET			
<i>r</i>	0.869	0.989	0.797
<i>p</i> -value	0.0126	< 0.0001	0.0387
BLEU			
<i>r</i>	0.980	0.901	0.955
<i>p</i> -value	0.0001	0.0060	0.0007
chrF2			
<i>r</i>	0.984	0.929	0.967
<i>p</i> -value	< 0.0001	0.0024	0.0003
TER			
<i>r</i>	0.852	0.922	0.936
<i>p</i> -value	0.0174	0.0031	0.0018

Table 14: Pearson correlations (r) and p -values for EN-JA for three pairs of score value sets across six translation systems.

Letter grade	Numeric grade
A	4.00
A-	3.67
A-/B+	3.50
B+	3.33
B	3.00
B-	2.67
B-/C+	2.50
C+	2.33
C	2.00
C-	1.67
C-/D+	1.50
D+	1.33
D	1.00
F	0.00

Table 15: Manual grading scale.

5_en_short: English-Russian				
	Pearson Correlation		Spearman Correlation	
	<i>r</i>	<i>p</i>	ρ	<i>p</i>
MT1	0.689	< 0.0001	0.755	< 0.0001
MT2	0.549	0.0016	0.765	< 0.0001
MT3	0.660	0.0001	0.765	< 0.0001
LLM1	0.692	< 0.0001	0.806	< 0.0001
LLM2	0.795	< 0.0001	0.734	< 0.0001
LLM3	0.548	0.0016	0.567	0.0011

5_en_short: English-Japanese				
	Pearson Correlation		Spearman Correlation	
	<i>r</i>	<i>p</i>	ρ	<i>p</i>
MT1	0.609	0.0004	0.743	< 0.0001
MT3	0.626	0.0002	0.699	< 0.0001
MT4	0.738	< 0.0001	0.657	0.0001
LLM1	0.827	< 0.0001	0.811	< 0.0001
LLM2	0.767	< 0.0001	0.783	< 0.0001
LLM3	0.839	< 0.0001	0.910	< 0.0001

Table 16: Pearson (r) and Spearman (ρ) correlation coefficients between sentence-level COMET scores and numeric human grades for the select “10-10-10” translation outputs for 5_en_short.

3_en_short: English-Russian				
	Pearson Correlation		Spearman Correlation	
	<i>r</i>	<i>p</i>	ρ	<i>p</i>
MT1	0.927	< 0.0001	0.895	< 0.0001
MT2	0.845	< 0.0001	0.779	< 0.0001
MT3	0.878	< 0.0001	0.851	< 0.0001
LLM1	0.761	< 0.0001	0.830	< 0.0001
LLM2	0.697	< 0.0001	0.738	< 0.0001
LLM3	0.663	< 0.0001	0.774	< 0.0001

3_en_short: English-Japanese				
	Pearson Correlation		Spearman Correlation	
	<i>r</i>	<i>p</i>	ρ	<i>p</i>
MT1	0.558	< 0.0001	0.672	< 0.0001
MT3	0.831	< 0.0001	0.843	< 0.0001
MT4	0.630	< 0.0001	0.694	< 0.0001
LLM1	0.582	< 0.0001	0.589	< 0.0001
LLM2	0.462	0.0002	0.436	0.0004
LLM3	0.665	< 0.0001	0.646	< 0.0001

Table 17: Pearson (r) and Spearman (ρ) correlation coefficients between sentence-level COMET scores and numeric human grades for the select “20-20-20” translation outputs for 3_en_short.

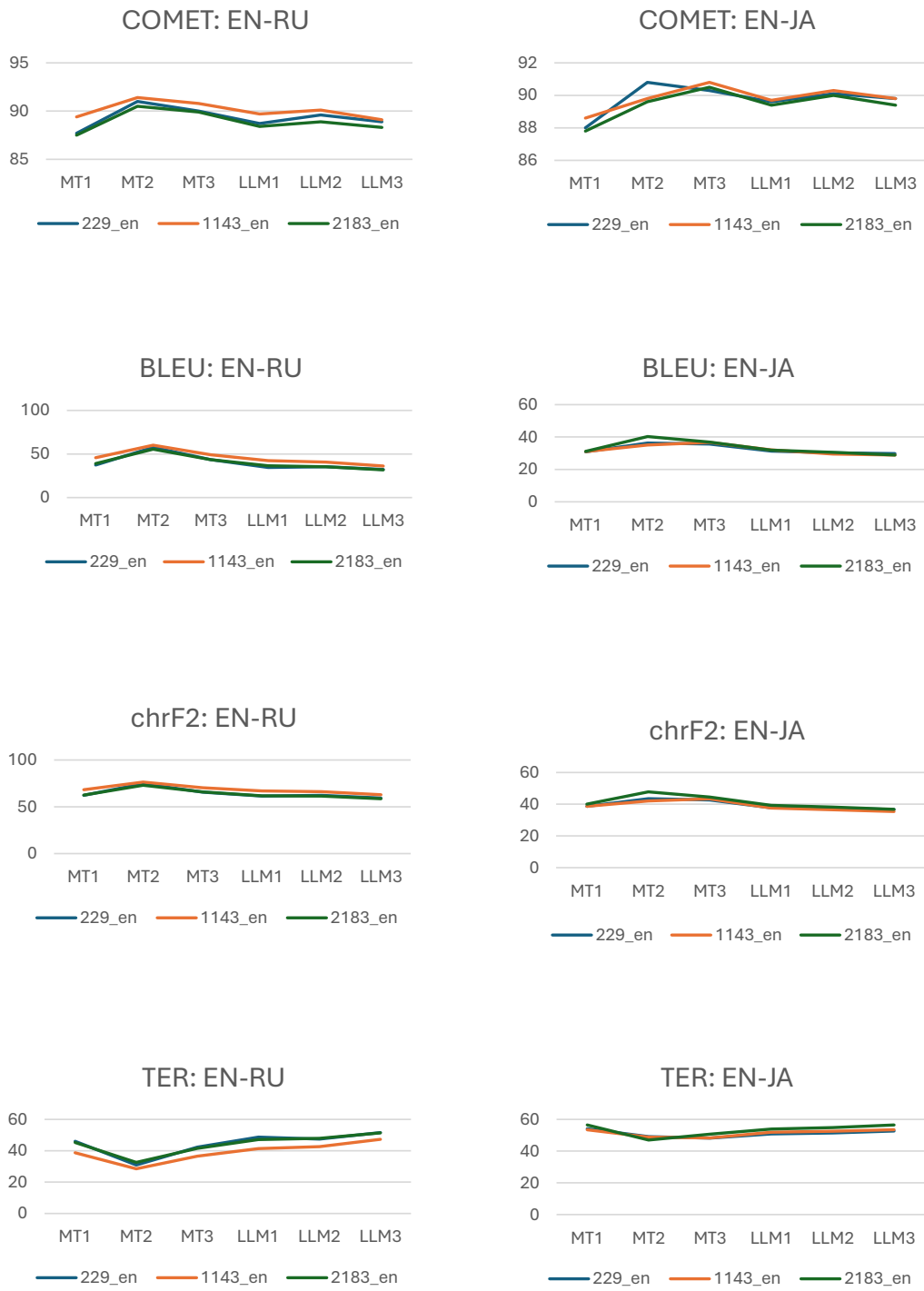


Figure 2: Automatic evaluation scores across the outputs of six translation systems for three non-overlapping parts of the RFTC corpus: 229, 1143, and 2183 segments.

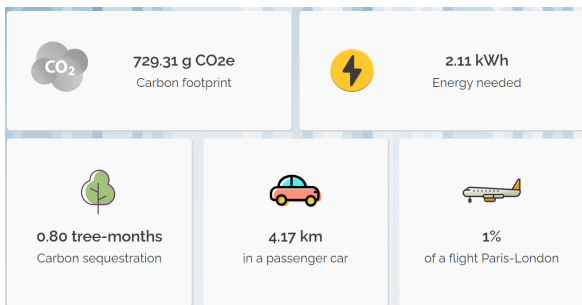


Figure 3: Carbon imprint for 13th Gen Intel(R) Core(TM) i9-13900KF 3.00 GHz 64.0 GB PC.

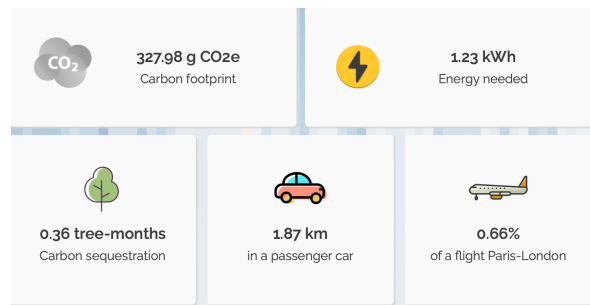


Figure 4: Carbon imprint for virtual server with Intel Xeon (Skylake) 6-core CPU, 16 GB of RAM.