Google Translate's Research Submission to WMT2025

Mara Finkelstein, Geza Kovacs, Isaac Caswell, Tobias Domhan,
Jan-Thorsten Peter, Juraj Juraska, Markus Freitag, David Vilar
Google Translate

Abstract

Large Language Models have shown impressive multilingual capabilities, where translation is one among many tasks. Google Translate's submission to the 2025 WMT evaluation tries to research how these models behave when pushing their translation performance to the limit. Starting with the strong Gemma 3 model, we carry out supervised fine tuning on high quality, synthetically generated parallel data. Afterwards we perform an additional Reinforcement Learning step, with reward models based on translation metrics to push the translation capabilities even further. Controlling the combination of reward models, including reference-based and quality estimation metrics, we found that the behaviour of the model could be tailored towards a more literal or more creative translation style. Our two submissions correspond to those two models. We chose the more creative system as our primary submission, targetting a human preference for better sounding, more naturally flowing text, although at the risk of losing on the accuracy of the translation. It is an open question to find the sweet spot between these two dimensions, which certainly will depend on the specific domain to handle and user preferences.

1 Introduction

In this paper we present Google Translate's research submission to the General MT track for the WMT 2025 shared task. Starting with Gemma 3 (Gemma Team, 2025), a strong multilingual LLM, we focus on improving its translation capabilities through supervised fine-tuning (Section 2) and reinforcement learning (RL) (Section 3). We use a mix of human- and synthetically-generated parallel data for boosting translation performance, as well as general domain post-training data in order to mostly retain the general capabilities of the original model. Through combinations of different reward models in the reinforcement learning step, we were able to generate two candidates: one more targeted towards fluent translations, the other towards more literal but sometimes slightly unnatural translations. In the end we chose to submit the more fluent system as our primary submission.

2 Supervised Fine-Tuning

For supervised fine-tuning (SFT), we begin with the released Gemma 3 27B model. We use parallel data including both human-generated texts as well as synthetic data generated by Gemini (Gemini Team, 2025). In addition we include human-generated Multidimensional Quality Metrics (MQM) translation error annotation data as made available from the WMT evaluation campaigns,[1] as well as generic instruction-following data. We use the public Gemma Kauldron SFT tooling[2] to fine-tune the Gemma 3 27B pretrained checkpoint. For fine-tuning we use the AdaFactor (Shazeer and Stern, 2018) optimizer with a learning rate of 0.0001 and a batch size of 64, running for 20k steps.

2.1 SFT Data

We used 4 different types of data for the Supervised Fine Tuning step.

Synthetic Gemini-Generated Translation Data Our synthetic data is generated using MADLAD-400 as the monolingual source (Kudugunta et al., 2023). The MADLAD-400 sources are first bucketed by length, and then sampled in each bucket to obtain 1 million source segments for each language pair we wish to generate synthetic data for. We then run a preliminary filtering step across these source segments where we take 2 samples from Gemini 2.5 Flash (1 greedy decoding, 1 sampled at temperature=1.0) and compare their scores according to MetricX 24-QE (Juraska et al., 2024). We

[1] https://github.com/google/wmt-mqm-human-evaluation

[2] https://kauldron.readthedocs.io/en/latest/

Proceedings of the Tenth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 723–731
November 8–9, 2025.

select the 60k sources where the sample achieves the largest improvement over the greedy decoding. The intuition behind this source filtering approach is that we wish to select sources that will benefit the most from 128-sample QE decoding, so we use 2 samples as a low-cost approximation. We generate at two distinct lengths this way: individual sentences and text blobs of up to 512 tokens. This way we aim to support both translations of individual segments as well as longer texts.

After this selection process, for each of the 60k sources for each language pair we generate 128 samples from Gemini 2.5 Flash and then apply a MetricX 24-QE filter to select the best-performing examples. In order to avoid formatting issues or erroneous translations, we apply an additional filtering step, based again on Gemini 2.5 Flash. This methodology was applied to the language pairs listed in Table 1. First SFT experiments were carried out on a subset of this data, marked in bold in the table.

For translations into Serbian we created a synthetic data variant in both Cyrillic and Latin script with some post-processing filters based on unicode ranges to make sure the translations are in the correct script. The goal of the synthetic data is to cover all languages relevant for the shared task. Except for Bhojpuri, Bengali and Maasai, the data covers all languages of the primary translation task as well as the multilingual subtask: synthetic data generation for Bhojpuri and Bengali did not finish in time for the shared task submission, and we decided to exclude Maasai due to quality concerns given the extremely low-resource nature of the language. In addition, Maasai was not covered by the multilingual pre-training of the MetricX base model, so that QE scores are likely not reliable.

Human-Generated Translation Data To increase the diversity and script coverage of the data we also include data for additional lower-resource languages. For these languages, due to uncertainty about the quality of Gemini-generated synthetic data, we opt to use human-generated parallel data instead. This data comes from the SMOL (Caswell et al., 2025) and GATITOS (Jones et al., 2023) datasets. SMOL covers 221 languages and GATITOS covers 170. This data was only used for the SFT stage, not RL.

Human-Generated MQM Data We include MQM data from WMT 2020 - 2023 (Lommel et al., 2014; Freitag et al., 2021) in the general training data mix. The intention is to increase the diversity of the training data and add information on translation error scoring. The model response is formatted as JSON as seen in Figure 1. A model fine-tuned only on the MQM portion is used as an AutoMQM model (Fernandes et al., 2023) for RL (see Section 3).

Generic Instruction-Following Data Our SFT mixture also includes 40% generic instruction-following data from the original Gemma 3 mixture. The purpose of including this data is to prevent the model from overfitting to the translation task and to maintain generic instruction-following capabilities.

2.2 Translation Performance

In order to measure the performance during the development cycle, we used a subset of the WMT24++ (Deutsch et al., 2025) corpus. We selected those language pairs (starting from English) that are also included in the WMT25 evaluation campaign, i.e. English to Arabic (Egypt), Chinese (Simplified), Czech, Estonian, Icelandic, Italian, Japanese, Korean, Russian, Serbian and Ukrainian.

In Table 2 results for the SFT approach are shown. We first experimented with running SFT on a set of 17 languages, some of which were included in the WMT25 set of languages. On this setup we saw improvements both on MetricX-24-XXL (Juraska et al., 2024) and COMET22 (Rei et al., 2022) (although only slight), but we saw a significant degradation on CHRF (Popović, 2015). Examining the produced translations, we saw a typical case of overfitting: the languages covered by our dataset saw improvements in translation quality, while those not included suffered from important degradations, especially those with alphabets not included in the data (which explained the big drop in CHRF).

With this observation, we designed a setup that tried to balance the improvements while keeping the overall performance. We lowered the learning rate, froze the embeddings and added generic SFT data derived from the Gemma 3 post-training setup, as well as the MQM data. With this setup, we were able to improve the quality as measured with MetricX and COMET22, while also recovering the original CHRF score. We were able to not only avoid drops but even see gains for languages that were not covered by the translation data mix. For example Bengali dropped from 41.83 CHRF

English (en)	↔	**Arabic (ar)**, **Chinese (zh)**, Czech (cs), **Dutch (nl)**, Estonian (et)*, Farsi (fa)*, **French (fr)**, **German (de)**, Greek (el)*, **Hindi (hi)**, **Indonesian (id)**, Indonesian (id)*, Icelandic (is)*, **Italian (it)**, **Japanese (ja)**, Kannada (kn)*, **Korean (ko)**, Lithuanian (lt)*, Marathi (mr)*, **Polish (pl)**, **Portuguese (pt)**, Romanian (ro)*, **Russian (ru)**, Serbian (sr)*, **Spanish (es)**, Swedish (sv)*, **Thai (th)**, **Turkish (tr)**, Ukrainian (uk), **Vietnamese (vi)**
Japanese (ja)	↔	Chinese (zh)

Table 1: List of language pairs for which we generated synthetic data. For language pairs marked with * formatting filtering was not applied. Languages in bold were included in the first set of experiments

System	MetricX	COMET22	CHRF
Baseline (Gemma 3)	3.08	82.7	41.3
SFT on 17 langs	2.94	82.8	37.7
+ general setup	2.81	83.8	41.1
+ WMT25 langs	2.86	84.4	44.2

Table 2: Supervised fine-tuning results on WMT priority languages. "17 langs" refers to the 17 language pairs marked in bold in Table 1.

to 13.87 in the initial SFT setup, while improving to 45.40 in the general setup (and 46.10 when increasing the language coverage). That is despite not being covered by SFT data. This shows the importance of carefully selecting the fine-tuning setup and monitoring languages/scripts outside of the training data.

Lastly we expanded our parallel dataset to all languages included in WMT25 (except Bhojpuri, Maasai and Bengali), which provided an additional boost in COMET22 and CHRF, with a negligible drop in MetricX.

3 Reinforcement Learning

We performed reinforcement learning on top of the SFT checkpoint, using an ensemble of metrics as reward models, to further boost translation quality.

3.1 Reward Models

We used the following metrics as reward models during RL:

- MetricX-24-XXL-QE (Juraska et al., 2024), a learned, regression-based translation metric producing a floating point score between 0 (best) and 25 (worst), matching the standard MQM score range (Freitag et al., 2021). MetricX scores were linearly rescaled, using $5.0 - \texttt{score}$, when computing rewards, so that higher scores indicate better quality. Although MetricX can take source, reference, and hypothesis as input, we passed in an empty reference to use it as a QE score only.

- Gemma-AutoMQM-QE, a finetuned AutoMQM model (Fernandes et al., 2023). This model was initialized from the Gemma3-27B-IT checkpoint (Gemma Team, 2025), and was trained on MQM ratings data from WMT 2020 - WMT 2023 (Lommel et al., 2014; Freitag et al., 2021). Default MQM weights (Freitag et al., 2021) were used in computing (token-level) rewards from AutoMQM outputs. As with MetricX, it ignores the reference translation.

- Generalist reward model covering many tasks, including reasoning, instruction following, and multilingual abilities, adapted from the general Gemma 3 post-training setup (Gemma Team, 2025).

We used RL algorithms extended to support token-level advantages, which were added to the advantages computed from sequence-level rewards. This allowed us to use fine-grained, span-level reward signals from AutoMQM directly, for improved credit assignment and training efficiency in the spirit of Ramos et al. (2024). See Figure 1 for an illustration of how MetricX and AutoMQM rewards were (additively) combined during advantage computation. The combined advantages were then batch-normalized.

3.2 Language Distribution

For RL we used the same translation data as for SFT[3], but ignored the (synthetic) references, since

[3]Except for GATITOS and SMOL, which were used in SFT only.

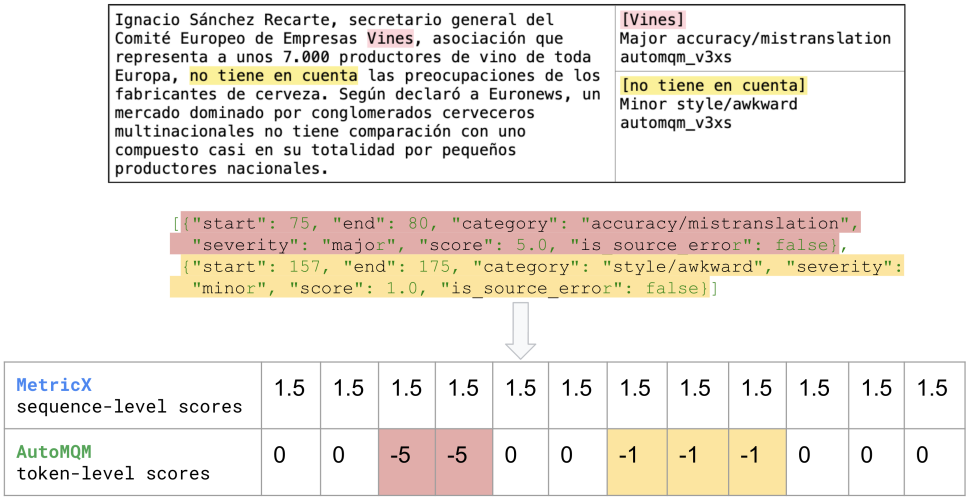
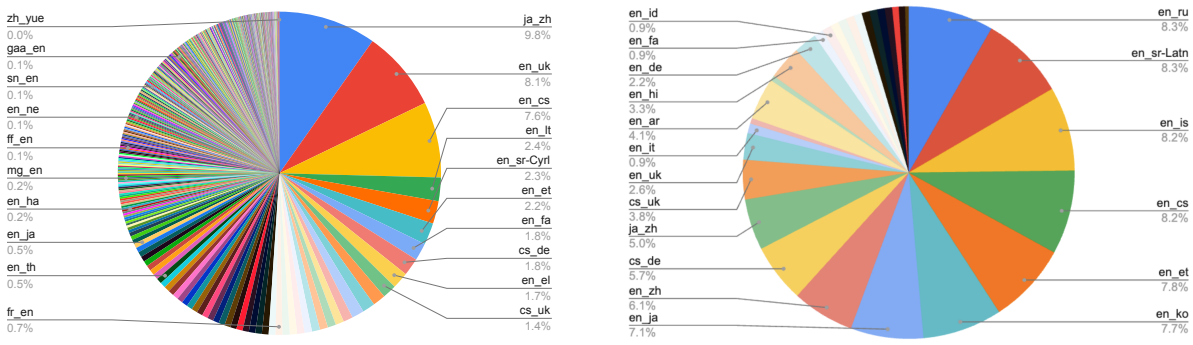


Figure 1: Illustration of how sequence-level and token-level rewards are additively combined during advantage computation in RL. Note that advantage is computed from sequence-level rewards as 'reward-to-go', meaning that rewards are broadcast uniformly to every token.



(a) SFT data mixture.

(b) RL data mixture.

Figure 2: Language distribution (token count) in the GemTrans data mixtures.

the rewards we used were reference-free. Furthermore, we included the WMT languages missing from the SFT step (*bn* and *bho*).

We re-balanced the RL data with UniMax sampling (Chung et al., 2023) to balance the distribution of language pairs in the final mixture, resulting in subsampling the high resource language pairs so that all have the same weight. The final proportion of languages for the SFT and RL phases can be found in Figure 2. While the RL prompt set was (approximately) a subset of the SFT data (note that the RL prompt set excluded the non-translation SFT split), we hypothesized that further gains in performance were still possible, given that the RL learning objective is very different from that of SFT, and improvements from RL (e.g., learning to not hallucinate) should generalize independently of the prompt set used.

System	MetricX	COMET22	CHRF
Baseline (Gemma 3)	3.08	82.7	41.3
+ full SFT	2.86	84.4	44.2
+ RL GEMTRANS1	2.40	83.7	37.4
+ RL GEMTRANS2	2.85	83.7	40.4

Table 3: Reinforcement learning results on WMT priority languages.

3.3 Translation Performance

We experimented with different combinations of reward functions and data conditions on small scale experiments, starting from the fine-tuned model described in Section 2. Our primary submission, "GEMTRANS1", used the ensemble of reward models described in Section 3.1. Our secondary submission, which we refer to as "GEMTRANS2", used

the reference-based version of MetricX-24-XXL (with the same synthetic references used for SFT, as described in Section 2.1), and used a prompted, rather than finetuned, AutoMQM reward model. This prompted AutoMQM model used the same prompt format as the finetuned model. Translation results are shown in Table 3. It can be seen that GEMTRANS1 managed to improve significantly on MetricX, although at the cost of drops in COMET22 and CHRF. GEMTRANS2 was not able to improve over the base SFT system in any metric. However, we still kept it as a candidate for submission, keeping in mind that GEMTRANS1 may potentially be overfitting to MetricX.

4 Automatic Post-editing

In order to minimize formatting errors, we run an additional post-editing pass on the resulting translations. Each model post-edits its own translations, i.e. they serve both as translation and post-edit systems. We prompt the model to just fix the formatting, without altering the text of the translation, using the prompt shown in Appendix A. The examples were chosen from errors we spotted during the development process of the model. On automatic metrics we saw small but consistent improvements for both GEMTRANS1 and GEMTRANS2 systems.

5 Final Submission

As described in Section 3.3, we ended up with two main candidates for submission. MetricX showed a clear preference for GEMTRANS1, although there was a clear drop in CHRF.[4] This might be a signal of the model overfitting to MetricX, so in order to get a better picture, we prompted Gemini to compare a subset of the translations of the WMT25 test sets, assigning a score of +1 if GEMTRANS1 was preferred, and -1 otherwise. The results can be found in Table 4, and show a preference for GEMTRANS1.

Additionally we also prompted Gemini to perform an MQM evaluation with additional quality scoring (similar to this year's setup in the shared task evaluation). In this case, the system preferred GEMTRANS2, both in terms of the quality score as well as MQM. When looking into the decomposition into accuracy and fluency scores, the MQM analysis shows better accuracy for GEMTRANS2, but at the cost of fluency.

[4]We had already discarded the failure cases described in Section 2.2 for this setup.

Language Pair	Mean
cs→de_DE	+0.22
cs→uk_UA	−0.07
en→ar_EG	+0.24
en→bho_IN	+0.23
en→cs_CZ	−0.07
en→et_EE	+0.15
en→is_IS	+0.26
en→ja_JP	+0.26
en→ko_KR	+0.31
en→mas_KE	+0.31
en→ru_RU	+0.26
en→sr_Latn_RS	+0.20
en→uk_UA	+0.31
en→zh_CN	+0.11
ja→zh_CN	+0.12
Average	+0.17

Table 4: Gemini side-by-side scores for the different language pairs. A positive score represents preference for GEMTRANS1, a negative score preference for GEMTRANS2.

Lastly we manually spot checked the translations (focusing on German outputs). We can confirm the conclusions of the MQM analysis, as we found the output of GEMTRANS1 to be more fluent than that of GEMTRANS2. Table 5 shows some examples comparing the two systems. In the first example, GEMTRANS1 uses the informal form "du" instead of the formal "Sie", which is more appropriate for the context of YouTube videos, and in general the flow of the text is more natural than that of GEMTRANS2. The second example shows a more interesting effect. We saw that GEMTRANS1 in several cases adds "explanations" to its output, e.g. it expands the UCL and LSE abbreviations in the example shown in Table 5. In general we found this added information to be useful and appropriate, and we felt that it indeed increased the translation quality. However, in a more strict sense, these could be interpreted as "accuracy" errors, which would explain the error numbers shown in Table 6. GEMTRANS2 on the other hand seemed indeed to be a more accurate model, but it sometimes produced "unnatural" translations.

Of course all that glitters is not gold, and we also found examples where the added text from GEMTRANS1 was incorrect or even hallucinated. As for the question of overfitting to MetricX, we can-

Source	So that's it for today's ikea hacks video. I hope that you enjoy this one and give it a big thumbs up if you did and let me know in the comment section below which one of these projects was your favorite. `[...]`
GEMTRANS1	Das war's für heute mit unserem Video zu IKEA-Hacks. Ich hoffe, es hat **euch** gefallen, und wenn ja, **gebt** uns gerne ein "Daumen hoch"! **Schreibt** uns auch in den Kommentaren, welches dieser Projekte **euch** am besten gefallen hat. `[...]`
GEMTRANS2	Das war's für heute mit diesem Video zu IKEA-Hacks. Ich hoffe, es hat **Ihnen** gefallen, und **geben Sie** ihm einen Daumen hoch, wenn das der Fall war, und **lassen Sie** mich im Kommentarbereich unten wissen, welches dieser Projekte **Ihr** Favorit war. `[...]`

Source	Game 17 of #UniversityChallenge is the first of eight knockout games in the second round. This was UCL and LSE, and UCL gave an outstanding performance: `[...]`
GEMTRANS1	Das 17. Spiel der #UniversityChallenge ist das erste von acht K.O.-Spielen in der zweiten Runde. Es wurde zwischen der UCL **(University College London)** und der LSE **(London School of Economics)** gespielt, und die UCL zeigte eine herausragende Leistung: `[...]`
GEMTRANS2	Spiel 17 von #UniversityChallenge ist das erste von acht K.O.-Spielen in der zweiten Runde. Es handelte sich um UCL und LSE, und UCL zeigte eine herausragende Leistung: `[...]`

Table 5: Translation examples for both systems under consideration. GEMTRANS1 generally produces more fluent translations, sometimes with added explanations that help the understanding of the text.

		MQM			
System	Q	Total	Acc	Fl	O
GEMTRANS1	68.0	15.5	13.6	1.1	0.8
GEMTRANS2	75.6	11.4	8.4	1.9	1.0

Table 6: MQM evaluation by prompting Gemini for MQM and quality (Q) scores. The MQM scores are additionally split into accuracy (Acc), fluency (F) and "other" (O) categories.

not completely discard this hypothesis, but we did not see any evidence of pathological outputs that might be gaming the metric. We decided to move forward with this system, as the general quality indeed seemed to be superior to that of GEMTRANS2, and this constituted our primary submission for the shared task.

6 Findings of the Human Evaluation

The organizers of WMT shared the results of the human evaluation. The performance of GEMTRANS1

Language Pair	Cluster	Rank
cs→de_DE	2	9-14
cs→uk_UA	2	4-8
en→ar_EG	9	11-14
en→bho_IN	n/a	n/a
en→cs_CZ	4	13-16
en→et_EE	7	10-12
en→is_IS	9	12-12
en→it_IT	1	1-4
en→ja_JP	3	12-16
en→ko_KR	2	5-10
en→mas_KE	n/a	n/a
en→ru_RU	3	13-16
en→sr_Cyr_RS	6	8-9
en→uk_UA	2	4-8
en→zh_CN	2	5-10
ja→zh_CN	5	14-15

Table 7: Human evaluation results for the GEMTRANS1 system.

has been summarized in Table 7. It can be seen that GemTrans performs generally in the top 3 clusters for 8 of the 14 language pairs where it was evaluated by humans, being in the first one for English to Italian. The first clusters are usually taken by large systems with a much bigger parameter count. The ranks show a wider variance, due to the highly competitive landscape of this year's shared task and the big number of participating systems.

For English to Arabic GEMTRANS1 obtained a very low human score, along with all other systems in its cluster. This was the result of the system failing to produce the correct Arabic dialect (Egyptian). In a related fashion, Bhojpuri showed low automatic scores (thus GEMTRANS1 was not included in the human evaluation) and we suspect that GEMTRANS1 failed to generate Bhojpuri, falling back to Hindi instead.

In the report about the shared task, in the additional analysis of the Serbian translations, the organizers explicitly highlight the GEMTRANS1 system for a "notable amount of idiomatic translations, even more than humans" (Kocmi et al., 2025), although they also point out a "relatively high number of errors". These findings agree with our own observations about the system.

7 Conclusions

We have presented the Google Translate Research submission to the WMT25 evaluation campaign. Starting from Gemma 3 we used supervised fine-tuning and RL to boost translation performance, in addition to a small automatic post-editing step to improve the formatting of the translations.

Out of the two final candidate systems, we found that one was more tailored towards fluency while the other one was more tailored towards accuracy (possibly illustrating the tradeoff discussed by e.g. Flamich et al. (2025); Schleiermacher (1816); Dryden (1685)). After evaluation with automatic metrics as well as manual inspection, we decided to move forward with the more fluent system, as it seemed to produce generally higher quality translations.

Whether that was the correct decision may largely depend on the criterion of the human evaluators. Quality of machine translation is indeed in the eye of the beholder, and the more "free-style" translations produced by the system may not be the preferred ones if more literal translation are desired, closer to the source sentence. This dichotomy again highlights the difficulties of machine translation evaluation, and may be indeed point towards new research directions (for both MT generation and evaluation), where the intent of the translation can play a more relevant role.

References

Isaac Caswell, Elizabeth Nielsen, Jiaming Luo, Colin Cherry, Geza Kovacs, Hadar Shemtov, Partha Talukdar, Dinesh Tewari, Baba Mamadi Diane, Koulako Moussa Doumbouya, et al. 2025. Smol: Professionally translated parallel data for 115 under-represented languages. *arXiv preprint arXiv:2502.12301*.

Hyung Won Chung, Noah Constant, Xavier Garcia, Adam Roberts, Yi Tay, Sharan Narang, and Orhan Firat. 2023. Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining.

Daniel Deutsch, Eleftheria Briakou, Isaac Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, et al. 2025. Wmt24++: Expanding the language coverage of wmt24 to 55 languages & dialects. *arXiv preprint arXiv:2502.12404*.

John Dryden. 1685. *Sylvae [Translator's preface]*. A Scolar Press facsimile. Scolar Press.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André FT Martins, Graham Neubig, Ankush Garg, Jonathan H Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. *arXiv preprint arXiv:2308.07286*.

Gergely Flamich, David Vilar, Jan-Thorsten Peter, and Markus Freitag. 2025. You cannot feed two birds with one score: the accuracy-naturalness tradeoff in translation.

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.

Gemini Team, Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, et al. 2025. Gemma 3 technical report.

Alexander Jones, Isaac Caswell, Orhan Firat, and Ishank Saxena. 2023. GATITOS: Using a new multilingual lexicon for low-resource machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 371–405, Singapore. Association for Computational Linguistics.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica M. Lundin, Christof Monz, Kenton Murray, Masaaki Nagata, Stefano Perrella, Lorenzo Proietti, Martin Popel, Maja Popović, Parker Riley, Mariya Shmatova, Steinþór Steingrímsson, Lisa Yankovskaya, and Vilém Zouhar. 2025. Findings of the wmt25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: A multilingual and document-level large audited dataset. *Advances in Neural Information Processing Systems*, 36:67284–67296.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica*, (12):0455–463.

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.

Miguel Moura Ramos, Tomás Almeida, Daniel Vareta, Filipe Azevedo, Sweta Agrawal, Patrick Fernandes, and André FT Martins. 2024. Fine-grained reward optimization for machine translation using error severity mappings. *arXiv preprint arXiv:2411.05986*.

Ricardo Rei, José GC De Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André FT Martins. 2022. Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585.

Friedrich Schleiermacher. 1816. *Über die verschiedenen Methoden des Übersetzens*. Abhandlungen der Königlichen Akademie der Wissenschaften in Berlin. Walter de Gruyter GmbH.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR.

Appendix

A APE Prompt

```
You are a copy-editor fixing formatting issues related to translation. You will see
a source in {src_lang} and its translation in a {tgt_lang}. If there are formatting
errors, fix them. Here are examples of things to fix:

{examples}

Do NOT fix the translations themselves. Only fix the sorts of minor errors described
above, IF they exist. MOST TRANSLATIONS WILL NOT NEED ANY CORRECTIONS!

Only output the fixed translation, with no additional formatting or chattiness. Here
is the example to fix:

source={src}
output={out}
corrected=
```

Examples

- Mismatching quotation marks:

  ```
  source=“Let her go!”
  output="¡Déjala ir!"
  corrected=“¡Déjala ir!”
  ```

- “user” omitted from beginning:

  ```
  source=@user38 heard this essay was good and it is
  output=Ich habe gehört, dieser Artikel ist gut, und das ist er auch.
  corrected=@user38: Ich habe gehört, dieser Artikel ist gut, und das ist er auch.
  ```

- HTML tag translated, instead of preserved as tag

  ```
  source=<contents for=sec2>section 2...</contents>
  output=<Inhalt für=sec2>Abschnitt 2...</Inhalt>
  corrected=<contents for=sec2>Abschnitt 2...</contents
  ```