IRB-MT at WMT25 Translation Task: A Simple Agentic System Using an Off-the-Shelf LLM

Ivan Grubišić*
Division of Electronics
Ruđer Bošković Institute
Zagreb, Croatia
name.surname@irb.hr

Damir Korenčić*
Division of Electronics
Ruđer Bošković Institute
Zagreb, Croatia
name.surname@irb.hr

Abstract

Large Language Models (LLMs) have been demonstrated to achieve state-of-the-art results on machine translation. LLM-based translation systems usually rely on model adaptation and fine-tuning, requiring datasets and compute. The goal of our team's participation in the "General Machine Translation" and "Multilingual" tasks of WMT25 was to evaluate the translation effectiveness of a resource-efficient solution consisting of a smaller off-the-shelf LLM coupled with a self-refine agentic workflow. Our approach requires a high-quality multilingual LLM capable of instruction following. We select `Gemma3-12B` among several candidates using the pretrained translation metric `MetricX-24-XL` and a small development dataset. WMT25 automatic evaluations place our solution in the mid tier of all WMT25 systems, and also demonstrate that it can perform competitively for approximately 16% of language pairs.

1 Introduction

Machine translation (MT) is an important yet unsolved NLP task with significant practical applications Kocmi et al. (2024a). Large language models (LLMs) have become a basis for state-of-the-art MT solutions, and the best approaches rely either on commercial LLMs or on open-weights models adapted using translation-specific data (Kocmi et al., 2024a). However, commercial cloud-based models may introduce cost constraints and dependency issues, while the adaptation of open-weight models commonly requires substantial computational resources and specialized training datasets.

Recent developments in LLMs yielded smaller yet capable models such as Gemma3 (Team et al., 2025) and Qwen3 (Yang et al., 2025), which are multilingual and support instruction following and reasoning. In parallel, research in multi-agent systems led to task-independent workflows, such as self-refine (Madaan et al., 2023), and task-oriented workflows where individual agents assume natural task-specific roles (Wu et al., 2024). Both approaches have demonstrated the capability of the agentic workflows to outperform individual LLMs (Madaan et al., 2023; Wu et al., 2024).

We hypothesize that the combination of capable smaller LLMs and agentic workflows has the potential to create a resource-effective translator with solid performance. Our participation in the WMT25 Translation Task (Kocmi et al., 2025a) is oriented toward testing this hypothesis in the controlled environment of the "constrained" track, which allows only openly available datasets and models below 20B parameters.

We evaluate our approach on the WMT25 General Machine Translation task, which assesses MT systems across four domains (news, social media, speech, and literary) with document-level context and multi-modal resources including video, image, and speech data (Kocmi et al., 2025a). The task comprises 16 language pairs covering major language groups including morphologically rich, low-resource, and diverse script languages. We also participate in the Multilingual subtask, which extends evaluation to 15 additional target languages. Overview of the dataset statistics can be found in Table 3.

As the first step in designing our system we tested several multilingual generative models and encoder-decoder models specialized for translation, evaluating them on a subset of pairs from the WMT24++ dataset (Deutsch et al., 2025). `Gemma3-12B` (Team et al., 2025) proved to be the best solution in terms of `MetricX-24-XL` metric (Juraska et al., 2024) so our final system is based on this model. We enhance the model with a version of the self-refine workflow (Madaan et al., 2023) based on a prompt adapted for machine translation. Our system uses as input only the text modality, and works with paragraph-sized text segments.

*Equal contribution.

Proceedings of the Tenth Conference on Machine Translation, pages 753–764
November 8-9, 2025

WMT25 evaluations using a number of automatic translation metrics (Kocmi et al., 2025b) show that our system achieves mid-level performance when compared with all the participating systems that include team-submitted solutions (both constrained and unconstrained), as well as benchmarks added by the organizers (individual LLMs and commercial solutions). The system achieves competitive performance for five language pairs (en→zh, en→de, en→id, en→sv, en→vi) and human ESA annotations (Kocmi et al., 2025a) show that it often generates good translations. We make the code of the system freely available.[1]

2 Related Work

Several multi-agent LLM systems for machine translation (MT) have been proposed recently, often inspired by human collaborative problem-solving and professional translation workflows (Wu et al., 2024; Peter et al., 2024; Briakou et al., 2024; Wang et al., 2025b; Anonymous, 2025). These systems aim to address the limitations of single-model MT systems, including in handling linguistic nuances, context, and idiomatic expressions. They consist of autonomous LLM-based agents assigned to specialized tasks, organized in a workflow and sometimes embedded in an iterative loop.

Such multi-agent systems can demonstrate superior performance compared to non-agentic baselines (Briakou et al., 2024; Wang et al., 2025a; Anonymous, 2025), . For example (Briakou et al., 2024) reported large improvements over conventional zero-shot prompting and even outperformed top-performing WMT 2024 systems in some cases. Furthermore, human evaluations frequently show a preference for translations produced by these multi-agent systems (Wu et al., 2024; Anonymous, 2025).

While effective, the proposed systems rely on powerful LLMs (such as GPT-4o and Gemini 1.5 Pro) and often involve complex and computation-intensive workflows, which entail latency and computational overhead. In contrast, our system relies on a smaller open-weights model and a simple one-step self-refine workflow (Madaan et al., 2023). This makes it resource efficient with translation time comparable to zero-shot inference with a single LLM.

[1] https://github.com/igrubi/irb-mt-wmt2025

3 Dataset

The WMT25 translation evaluation dataset encompasses 31 language pairs with diverse characteristics, enabling assessment across different language families, resource levels, and translation scenarios (Kocmi et al., 2025a). The list of language pairs and the statistics of associated sub-datasets is displayed in Table 3.

The General MT task comprises 16 language pairs covering both large and small languages. The task includes both English-centric and non-English language pairs, with English-to-target directions covering Arabic (Egyptian), Bhojpuri, Chinese (Simplified), Czech, Estonian, Icelandic, Italian, Japanese, Korean, Maasai (Kenya), Russian, Serbian (Latin), and Ukrainian. Additionally, the task features non-English source languages with Czech-to-German, Czech-to-Ukrainian, and Japanese-to-Chinese pairs.

The dataset exhibits significant variation in size and text complexity. Russian dominates with 7,804 texts, followed by Hindi with 5,087 texts, while smaller language pairs like Italian contain only 87 texts. The dataset contains texts from four domains: news articles, transcripts of video speech associated with audio data, social media posts associated with printscreen images, and literary texts. A significant number of texts from the General MT subtask does not belong to any domain.

The Multilingual (sub)task extends evaluation to 15 additional target languages: Bengali, German, Greek, Persian, Hindi, Indonesian, Italian, Kannada, Lithuanian, Marathi, Romanian, Serbian (Cyrillic), Swedish, Thai, Turkish, and Vietnamese. English is the only source language, and all the texts belong to one of the four domains described above.

As the statistics in Table 3 show, the General MT texts tend to be short, predominantly with 100–200 tokens, and mostly consist of a single paragraph. The Multilingual subtask texts are longer (with the exception of Hindi), having over 400 tokens on average, and tend to consist of at least several paragraphs that are longer than the General MT paragraphs. Pair with the largest document collection (English-Hindi) is an exception since it contains mostly short texts.

4 System

The goal of our submission was to examine how a lightweight, simple, and computationally effi-

Qwen3		NLLB		Gemma3		EuroLLM		Aya	
Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
6.98	3.40	6.54	3.70	4.59	2.59	5.03	2.83	5.01	2.88

Table 1: Average performance of LLMs benchmarked on the development set, measured by `MetricX-24-XL` (lower is better).

cient agentic workflow for MT compares against a range of other approaches . To enable a fair comparison with systems utilizing a similar level of resources, we submitted our solution to the "constrained" track (Kocmi et al., 2025a) which allows only open-weights models with a combined size below 20B parameters.

We always apply our translation methods on the level of a paragraph, i.e., no document-wide context is used. Additionally, only text is used as input, i.e., image and speech data provided for parts of the dataset is not used.

4.1 Model Selection

A self-refine (Madaan et al., 2023) agentic workflow for MT requires a translation model and a model able to implement the refinement of a translation. For both steps strong multilingual capabilities are desirable, and the refinement model should have instruction following capabilities in order to properly implement the refinement instructions.

In order to select the appropriate models we tested several LLMs on a subset of language pairs from the WMT24++ dataset (Deutsch et al., 2025), which we used to create the development dataset. To this end, we used pairs from WMT24++ that matched pairs from the WMT25 General MT track (see Table 3 for a list of WMT25 pairs). Since WMT24++ contains only en→X pairs, we created the non-English development pairs (cs→uk, cs→de, ja→zh) by matching via English texts. WMT24++ does not contain English-Bhojpuri (en→bho) and English-Massai (en→mas) data, so we did not test on these pairs. The final development set was constructed by subsampling 200 texts for each language pair.

We tested the following models: `Gemma3-12B` , `EuroLLM-9B-Instr` , `Qwen3-8B` , `Aya-101` , and `NLLB-200-3.3B` (Team et al., 2025; Martins et al., 2025; Yang et al., 2025; Üstün et al., 2024; NLLB Team, 2024). `Gemma3-12B` , `Qwen3-8B` and `EuroLLM-9B-Instr` are modern implementations of the GPT architecture, supporting multilinguality, instruction following, and, in the case of `Gemma3-12B` and `Qwen3-8B` , reasoning. These models can serve as basis for both the translator and the translation refinement agent. `Aya-101` is a massively multilingual instruction following model, but in this context we view it purely as a translation model since its effective input context length of 1024 limits the applicability to expectedly longer refinement prompts. `NLLB-200-3.3B` is a state-of-the-art encoder-decoder transformer trained specifically for multilingual machine translation. All models except `NLLB-200-3.3B` (trained to translate the entire input text) were equipped with simple translation prompts detailed in Appendix C.1.

`MetricX-24-XL` (Juraska et al., 2024) (the "metricx-24-hybrid-xl-v2p6" variant) was used to estimate the models' performance. `MetricX-24-XL` MetricX is a metric learned from parallel text with source, hypothesis, and reference segments that are annotated with human Direct Assessment (DA) and MQM scores. It is based on the mT5 transformer model (Xue et al., 2021) with a regression head, and it can estimate translation quality both with and without a reference translation.

Table 1 shows that Gemma3 has the best average translation performance (averaged over all language pairs), and that it has lowest average standard deviation among all tested models. This indicates that is should have the best and most stable translation performance. Per-pair results in Table 4 in Appendix B show that Gemma3 has superior or competitive performance for almost all language pairs. Additionally, Gemma3 has both instruction-following and reasoning capabilities (Team et al., 2025).

For these reasons, we decided that it is an optimal model for both translation and refinement. Additional benefit of this choice is the use of a single model for the entire workflow, which reduces the memory footprint.

4.2 The Agentic Workflow

The final workflow implements a two-stage translation process based on the self-refine workflow

Table 2: AutoRank translation scores formed by aggregating multiple automatic translation metrics (Kocmi et al., 2025b) and data on the relative position of IRB-MT, for both GeneralMT (top row group) and Multilingual (bottom row group) tracks. For each pair, AutoRank scores for IRB-MT and `Gemma3-12B` are given. Additionally, for each system subcategory, the total number of systems (#sys) and the number of systems ranked above IRB-MT (#above) is given. Constrained systems use openly available data and models below 20B parameters. Team systems are submitted by participating teams, while Benchmark systems are included by the organizers (Kocmi et al., 2025b).

Lang.	**IRB-MT**	**Gemma3**	**Constrained**				**Unconstrained**			
Pair			**Team**		**Benchmark**		**Team**		**Benchmark**	
			# sys	**# above**	**# sys**	**# above**	**# sys**	**# above**	**# sys**	**# above**
cs→de	12.1	**11.2**	9	5	9	2	7	3	15	13
cs→uk	**8.9**	9.7	10	6	9	1	8	3	15	11
ja→zh	**12.1**	17.1	9	6	9	1	8	6	15	9
en→ar	**10.8**	11.7	7	5	9	1	6	3	15	12
en→bho	**11.4**	12.3	7	4	9	1	6	3	13	10
en→zh	**9.3**	10.6	9	5	9	0	5	3	15	9
en→cs	**12.6**	13.4	12	8	9	1	6	3	15	13
en→et	**11.1**	12.1	8	7	9	0	6	4	15	9
en→is	**11.9**	13.8	4	3	9	1	6	5	14	9
en→it	**10.2**	15.5	4	2	9	0	5	3	15	11
en→ja	**10.3**	13.6	11	8	9	0	8	5	15	12
en→ko	**8.4**	9.0	7	4	9	0	5	3	15	8
en→mas	9.7	**8.8**	2	1	9	7	5	3	11	9
en→ru	**9.9**	14.7	9	6	9	0	8	4	14	6
en→sr	**6.3**	7.6	7	5	9	0	6	2	13	5
en→uk	**8.0**	14.4	9	5	9	0	9	3	15	6
en→bn	**5.1**	7.6	2	1	9	0	5	2	14	5
en→de	**9.8**	12.2	3	1	9	1	5	4	15	13
en→el	**5.9**	9.9	3	2	9	0	5	3	15	8
en→fa	**5.1**	5.7	2	1	9	0	5	3	14	8
en→hi	**5.3**	7.1	2	1	9	0	5	3	14	5
en→id	**5.5**	6.6	2	1	9	0	5	3	15	6
en→kn	**11.0**	13.4	2	1	9	0	5	4	14	9
en→lt	**8.9**	10.2	3	2	9	0	5	4	15	9
en→mr	**7.3**	12.4	2	1	9	0	5	3	14	6
en→ro	**6.4**	7.9	3	1	9	1	5	3	15	8
en→sr_Cy	**9.9**	12.1	4	2	9	0	6	5	13	5
en→sv	**5.8**	11.4	3	1	9	0	5	3	15	6
en→th	**4.8**	9.1	2	1	9	0	5	2	14	7
en→tr	**7.2**	8.7	2	1	9	0	5	3	15	7
en→vi	**5.1**	8.1	2	1	9	0	5	3	14	6

(Madaan et al., 2023). The initial translation is generated using the provided WMT25 prompts (slightly modified by dropping the instruction to respect the paragraphs structure). Details of the prompts used for the translation workflow can be found in Appendix C.2.

In the next step a refinement prompt, tailored for machine translation, is executed. The refinement prompt consists of the original translation prompt, the input text, the initial translation, and task-specific instructions. The instructions elicit the model to reason about the improvement and to produce the solution enclosed within "`<solution>` `</solution>`" tags. For efficiency, the model is instructed to keep the reasoning at "close to 300 words". The inference temperature was set to 0 (no sampling), and the maximum number of new tokens was set to 20K.

We evaluated the self-refine workflow by comparing it with the basic `Gemma3-12B` translator on both General MT and Multilingual language pairs from the WMT25 dataset. `MetricX-24-XL` scores showed that the self-refine approach has similar or slightly better scores across the majority of language pairs. We took this as evidence that the proposed agentic system does not perform worse then the baseline. Since the original self-refine experiments show improvements for a number of models and tasks (Madaan et al., 2023), we were confident that the agentic system would best the base translator when evaluated with other translation metrics.

5 Results

IRB-MT submitted translations for all of the 31 language pairs of the GeneralMT and Multilingual subtasks (Kocmi et al., 2025a). Automatic non-human evaluations show that IRB-MT is a mid-tier constrained system that outperforms baseline `Gemma3-12B` in most cases, which demonstrates the benefit of the self-refine approach. Out of the 16 pairs for which human evaluation was performed, IRB-MT's performance was high enough for it to be selected for human evaluation in the case of 13 pairs (81.25% of pairs (Kocmi et al., 2025b)).

To further analyze the relative performance of IRB-MT we rely on the AutoRank metric that aggregates the rankings of multiple translation quality metrics, in order to mitigate biases of individual metrics (Kocmi et al., 2025b). Other evaluated translation systems consist of systems submitted by participating teams, and additional organizer-chosen systems included for comparison (Kocmi et al., 2025b). We label these two groups "Team" and "Benchmark" systems, respectively. Benchmark systems include open-weight and cloud-based LLMs, and commercial MT systems.

The AutoRank statistics in Table 2 show that IRB-MT outperforms Gemma3 for all but two language pairs. In the constrained track, IRB-MT compares favorably with the Benchmark systems, being above most of them. This is not surprising since these are smaller (below 20B parameters) LLMs applied as zero-shot translators. On the other hand, when compared to Team constrained systems IRB-MT is, for most pairs, located at or below the median rank. Presumably, these systems are mostly data-based, i.e., they rely on model adaptation and fine-tuning.

As for the unconstrained track IRB-MT compares relatively favorably with the Team systems, often placed close to the middle of the list. However, this probably has most to do with the fact that, surprisingly, submitted unconstrained systems generally perform worse than the submitted constrained systems (Kocmi et al., 2025b). Unconstrained Benchmark systems consist of mid-sized LLMs (above 20B parameters), commercial LLMs, and commercial translation systems. These systems mostly outperform IRB-MT for large languages and for most European languages, while IRB-MT compares more favorably and sometimes competitively on mid-and lower-resourced languages and non-European languages.

As the table Table 2 contains only information on relative performance, we provide statistics on the GEMBA-ESA translation scores (Kocmi and Federmann, 2023) computed using GPT4.1 (Kocmi et al., 2025b). The scores, contained in Table 5 in Appendix D, lay out the scores of IRB-MT and top-performing systems from all categories. IRB-MT scores range between approx. 50 and approx. 75 for most systems (100 being the perfect performance). However, for 5 language pairs (en→zh, en→de, en→id, en→sv, en→vi) IRB-MT both achieves a score close to 80 and is approximately 10 points below the top-performing system, which shows the potential of the approach.

Human evaluation (Kocmi et al., 2025a) of selected translation systems was performed by applying the ESA annotation method (Kocmi et al., 2024b) and, for two language pairs, by applying the MQM method (Freitag et al., 2021). The evaluated systems were clustered based on the statistical significance of performance differences (Kocmi et al., 2025a).

When compared to other human-evaluated systems, IRB-MT is located at or below the median for the majority of language pairs. Out of 11 pairs for which the systems were ESA-annotated, for 6 pairs IRB-MT either has a score above 66%, or it is not significantly different from systems scoring above 66% (Kocmi et al., 2025a). According to ESA annotation guidelines the score of 66% is the threshold for translations with "Most meaning preserved and few grammar mistakes" (Kocmi et al., 2024b). We take this as an argument for our system's ability to generate good translations for a non-trivial percentage of language pairs. For the challenging English-Arabic pair, IRB-MT outperforms all of the constrained systems and compares favorably to 50% of the unconstrained systems (Kocmi et al., 2024b).

6 Conclusion and Future Work

We proposed a simple lightweight "self-refine" workflow for machine translation, based on the multilingual `Gemma3-12B` LLM with instruction-following and reasoning capabilities. Our approach was included in the WMT25 (Kocmi et al., 2025a) evaluation with automatic metrics, performed on a large set of translation systems. The results place our system, on average, in the mid-tier, but there is a significant performance variations across language pairs, with the tendency of better per-

formance on mid- and low-resource languages. While the IRB-MT fails to come close to the top-performing systems, human ESA annotations show that it often produces good translations (Kocmi et al., 2025a).

Future research on the improvement of our approach should tackle the issue of performance variability. Human or LLM-assisted examination of language pairs with the lowest scores would reveal the type of errors and suggest improvements. Although the agentic workflow by-and-large outperforms the base Gemma3 model, there is a significant variation in the gap between the two. Analyzing the language pairs and texts for which IRB-MT fails to improve upon the base system could lead to the refinement of the agentic system.

Other directions for further improvement of IRB-MT include: a larger thinking budget, an iterative improvement loop, and a more granular agentic system with specialized roles. Refinement of the agentic structure could combine elements of the existing MT workflows (Wu et al., 2024; Peter et al., 2024; Briakou et al., 2024; Wang et al., 2025b; Anonymous, 2025). The key challenge is to boost performance without relying on overly complex and long workflows, or on too large LLMs. Examining different combinations of the translator LLM and the refiner LLM could also lead to a performance boost, and to insights into the models' behavior.

In general, it would be interesting to examine how close can purely agentic approaches come to the data-driven approaches based on LLM adaptation and fine-tuning, and how does the compute-vs-performance tradeoff look like.

Acknowledgements

This paper was supported by the European Union's NextGenerationEU program. We would like to thank Tomislav Šmuc, Ph.D., and Prof. Sonja Grgić, Ph.D., for support and valuable discussions. We acknowledge EuroHPC Joint Undertaking for awarding us access to MareNostrum5 hosted by BSC, Spain, under the project ID EHPC-DEV-2025D05-087.

References

Anonymous. 2025. Agentdiscotrans: Agentic LLMs for discouse-level machine translation. In *Submitted to ACL Rolling Review - February 2025*. Under review.

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. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1301–1317, Miami, Florida, USA. Association for Computational Linguistics.

Daniel Deutsch, Eleftheria Briakou, Isaac Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. 2025. Wmt24++: Expanding the language coverage of wmt24 to 55 languages & dialects.

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.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. Metricx-24: The google submission to the wmt 2024 metrics shared task.

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. 2025a. 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.

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, Steinthór Steingrímsson, and Vilém Zouhar. 2024a. 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, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Natalia Fedorova, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Kenton Murray, Masaaki Nagata, Stefano Perrella, Lorenzo

Proietti, Martin Popel, Maja Popović, Parker Riley, Mariya Shmatova, Steinþór Steingrímsson, Lisa Yankovskaya, and Vilém Zouhar. 2025b. Preliminary ranking of wmt25 general machine translation systems.

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.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024b. Error span annotation: A balanced approach for human evaluation of machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1440–1453, Miami, Florida, USA. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594. Curran Associates, Inc.

Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno M. Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M. Alves, José Pombal, Nicolas Boizard, Manuel Faysse, Pierre Colombo, François Yvon, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2025. Eurollm-9b: Technical report.

NLLB Team. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–850.

Anishka Peter, Mai Dang, Michael Liu, Joaquin Dominguez, and Nibhrat Lohia. 2024. Multi-agent translation team (matt): Enhancing low-resource language translation through multi-agent workflow. SMU Data Science Review, Vol. 8, No. 3, Article 3. Available at SMU Scholar.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 technical report.

George Wang, Jiaqian Hu, and Safinah Ali. 2025a. Maats: A multi-agent automated translation system based on mqm evaluation.

Xi Wang, Jiaqian Hu, and Safinah Ali. 2025b. Maats: A multi-agent automated translation system based on mqm evaluation.

Minghao Wu, Jiahao Xu, and Longyue Wang. 2024.

TransAgents: Build your translation company with language agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 131–141, Miami, Florida, USA. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 15894–15939.

A Dataset Statistics

Table 3: WMT25 Dataset Statistics by Language Pair, for the General MT task (top half), and the "multilingual" subtask (bottom half). For each pair the statistics pertain to the texts in the source language (predominantly English). The statistics include average number of tokens in the text, statistics on the number of paragraphs per text, and on the number of tokens per paragraph. Tokenization is done by splitting on whitespaces, while the paragraphs are separated by a double newline character.

Language Pair	# Texts	Avg Text-Tokens	# Paragraphs			# Para-Tokens		
			Q1	Avg	Q3	Q1	Avg	Q3
cs→uk	230	157.1	1.0	1.8	2.0	64.0	89.2	113.0
cs→de	256	171.1	1.0	1.8	2.0	66.0	95.8	117.0
ja→zh	106	413.8	1.0	3.2	4.0	51.0	131.3	188.0
en→ar	1,251	96.3	1.0	1.2	1.0	37.0	80.5	110.0
en→bho	1,251	96.3	1.0	1.2	1.0	37.0	80.5	110.0
en→zh	1,251	96.3	1.0	1.2	1.0	37.0	80.5	110.0
en→cs	1,277	101.6	1.0	1.2	1.0	39.0	83.8	113.0
en→et	1,251	96.3	1.0	1.2	1.0	37.0	80.5	110.0
en→is	1,607	84.0	1.0	1.2	1.0	22.0	72.9	101.0
en→it	87	422.6	1.0	3.8	4.0	82.0	110.8	137.0
en→ja	1,251	96.3	1.0	1.2	1.0	37.0	80.5	110.0
en→ko	1,251	96.3	1.0	1.2	1.0	37.0	80.5	110.0
en→mas	1,251	96.3	1.0	1.2	1.0	37.0	80.5	110.0
en→ru	7,804	52.6	1.0	1.0	1.0	13.0	51.0	46.0
en→sr	2,251	137.9	1.0	1.1	1.0	61.0	124.3	184.0
en→uk	1,251	96.3	1.0	1.2	1.0	37.0	80.5	110.0
en→bn	87	422.6	1.0	3.8	4.0	82.0	110.8	137.0
en→de	113	407.7	1.0	3.4	2.0	84.0	120.0	146.0
en→el	87	422.6	1.0	3.8	4.0	82.0	110.8	137.0
en→fa	87	422.6	1.0	3.8	4.0	82.0	110.8	137.0
en→hi	5,087	36.2	1.0	1.0	1.0	18.0	34.6	44.0
en→id	87	422.6	1.0	3.8	4.0	82.0	110.8	137.0
en→kn	87	422.6	1.0	3.8	4.0	82.0	110.8	137.0
en→lt	87	422.6	1.0	3.8	4.0	82.0	110.8	137.0
en→mr	87	422.6	1.0	3.8	4.0	82.0	110.8	137.0
en→ro	87	422.6	1.0	3.8	4.0	82.0	110.8	137.0
en→sr_Cyrl	87	422.6	1.0	3.8	4.0	82.0	110.8	137.0
en→sv	87	422.6	1.0	3.8	4.0	82.0	110.8	137.0
en→th	87	422.6	1.0	3.8	4.0	82.0	110.8	137.0
en→tr	87	422.6	1.0	3.8	4.0	82.0	110.8	137.0
en→vi	87	422.6	1.0	3.8	4.0	82.0	110.8	137.0

B Development set results

Table 4: `MetricX-24-XL` results of benchmarked LLMs on the development set for the language pairs from WMT24++ that occur in the WMT25 General MT track. The models are: `Qwen3-8B` , `NLLB-200-3.3B` , `Gemma3-12B` , `EuroLLM-9B-Instr` , and `Aya-101` .

Language Pair	Qwen3		NLLB		Gemma3		EuroLLM		Aya	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Average	6.98	3.40	6.54	3.70	**4.59**	**2.59**	5.03	2.83	5.01	2.88
cs→de	3.31	2.25	3.45	2.52	2.68	1.97	2.51	1.87	2.73	1.78
cs→uk	7.06	3.39	5.84	3.45	4.40	2.47	4.84	2.67	5.10	2.62
ja→zh	4.44	1.85	5.62	2.96	3.24	1.64	3.40	1.60	3.89	1.96
en→ar	5.46	2.89	5.77	3.49	5.40	2.71	4.82	2.64	5.22	2.81
en→zh	4.18	1.68	5.60	3.50	2.62	1.71	2.88	1.88	3.70	2.25
en→cs	7.87	4.39	7.08	4.36	5.47	3.22	4.85	2.86	6.15	3.51
en→et	15.04	6.10	7.84	4.48	8.08	4.18	6.31	3.62	7.28	4.19
en→is	17.76	6.52	8.79	4.82	10.12	4.75	16.08	7.03	7.42	3.46
en→it	3.43	2.48	3.43	2.60	2.83	2.08	2.80	1.98	4.15	3.13
en→ja	4.55	2.07	7.09	3.34	4.20	2.08	4.46	2.11	4.72	2.18
en→ko	4.82	2.44	14.06	4.11	3.96	1.98	4.32	2.12	4.95	2.79
en→ru	5.05	3.35	5.94	4.54	3.30	2.45	4.50	3.49	4.70	3.13
en→sr	7.60	4.28	4.59	3.42	4.07	2.68	3.85	2.80	4.69	3.15
en→uk	7.14	3.98	6.43	4.22	3.94	2.37	4.74	2.94	5.43	3.29

C Prompts

C.1 Simple Translation Prompts for Model Selection

These prompts were applied for benchmarking `Gemma3-12B` , `EuroLLM-9B-Instr` , `Qwen3-8B` , and `Aya-101` on the development set. In the case of `Aya-101` the system prompt was not used. Only in the case of `Gemma3-12B` was "Output ONLY the translated text!" added to the end of the system prompt, since the initial tests revealed that the model almost always produces "thinking" tokens before producing the translation.

C.1.1 System Prompt

```
You are a professional translator with expertise in multiple languages.
Provide accurate, natural translations that preserve meaning and context.
[Output ONLY the translated text!]
```

C.1.2 User Prompt

```
Please translate the following text from {source_language} to {target_language}:

{text}
```

C.2 Prompts for the Self-Refine Translation Workflow

These prompts were used in combination with `Gemma3-12B` to produce the final translations.

Prompt for `Gemma3-12B` Translator

For the translation agent prompts provided as part of the WMT25 datasets were used, as they are well-formed and convey additional domain-specific information. One such prompt was given for every text in the test dataset. To illustrate the structure of these prompts we display two prompts for Czech-German translation, for the "news" and "social" domains, respectively. The only modification of

the original prompts was the removal of the sentence “Retain the paragraph breaks (double new lines) from the input text.”, which was done because we always applied our translator on individual paragraphs.

You are a professional Czech-to-German translator, tasked with providing translations suitable for use in Germany (de_DE). Your goal is to accurately convey the meaning and nuances of the original Czech text while adhering to German grammar, vocabulary, and cultural sensitivities. The original Czech text is a news article. Ensure the translation is formal, objective, and clear. Maintain a neutral and informative tone consistent with journalistic standards. Produce only the German translation, without any additional explanations or commentary. ~~Retain the paragraph breaks (double new lines) from the input text.~~ Please translate the following Czech text into German (de_DE):
{**text**}

You are a professional Czech-to-German translator, tasked with providing translations suitable for use in Germany (de_DE). Your goal is to accurately convey the meaning and nuances of the original Czech text while adhering to German grammar, vocabulary, and cultural sensitivities. The original Czech text is user-generated content from a social media platform. Ensure you do not reproduce spelling mistakes, abbreviations or marks of expressivity. Platform-specific elements such as hashtags or userids should be translated as-is. Produce only the German translation, without any additional explanations or commentary. ~~Retain the paragraph breaks (double new lines) from the input text.~~ Please translate the following Czech text into German (de_DE):
{**text**}

Prompt for `Gemma3-12B` Translation Refinement

The `reasoning_words` parameter was fixed to 300, and the text of the solution was extracted from the <solution> </solution> tags.

Your job is to review a translation, and correct it if necessary.
You will be given an original text, translation instructions,
and the translation created according to these instructions.
Respect the instructions!

These are the instructions according to which the translation was produced:
{**translation_prompt**}

Original text:
{**original_text**}

Translation:
{**translation**}

First, analyze the instructions, the original text, and the translation.
Then reason about the improved solution (if any), and produce the solution.
Try to keep the reasoning succinct, close to {**reasoning_words**} words!
End with your final solution, enclosed within the <solution> </solution> tags.

D Comparison with other Participating Systems

Table 5: GEMBA-ESA translation scores (Kocmi and Federmann, 2023) computed using GPT4.1 (Kocmi et al., 2025b), for both GeneralMT (top row group) and Multilingual (bottom row group) tracks. Pairs for which GEMBA-ESA were not computed are omitted. For each pair, scores for IRB-MT and `Gemma3-12B` are given, as well as scores for best-performing systems in each sub-category defined by the properties of the system. Constrained systems use openly available data and models below 20B parameters. Team systems are submitted by participating teams, while Benchmark systems are included by the organizers (Kocmi et al., 2025b).

Language	IRB-MT	Gemma3	Constrained		Unconstrained	
Pair			Team	Benchmark	Team	Benchmark
cs→de	75.4	77.5	88.3	77.5	87.5	91.0
cs→uk	74.8	75.9	85.3	76.7	84.3	89.5
ja→zh	70.4	64.1	85.5	69.8	81.8	84.8
en→ar	67.5	67.6	75.0	67.6	75.4	84.5
en→zh	77.5	76.6	88.3	76.6	81.5	88.7
en→cs	73.6	74.1	89.4	75.8	86.2	91.5
en→et	60.5	59.4	87.8	59.4	74.3	90.7
en→is	47.2	42.1	83.9	76.3	85.1	87.6
en→it	79.8	74.7	88.7	78.6	88.0	90.5
en→ja	77.9	73.8	89.6	76.3	86.3	91.2
en→ko	76.3	77.0	85.9	77.0	82.3	88.1
en→ru	76.5	73.2	85.9	73.2	80.6	87.8
en→sr	66.7	63.6	86.5	63.6	75.3	86.9
en→uk	76.9	65.8	86.0	75.2	82.4	89.8
en→bn	72.7	65.9	83.2	65.9	75.1	86.6
en→de	79.0	76.2	90.6	80.0	89.0	91.7
en→el	73.9	62.9	85.8	67.7	84.1	88.7
en→fa	73.1	72.5	84.1	72.5	80.4	88.4
en→hi	74.3	70.1	82.3	70.8	79.0	86.3
en→id	80.6	81.1	87.1	81.1	83.7	89.3
en→kn	57.6	49.4	78.8	54.2	67.3	81.6
en→lt	61.2	58.3	84.1	58.3	72.4	87.3
en→mr	68.1	51.8	81.6	55.6	72.4	84.7
en→ro	77.4	77.9	86.3	79.9	86.0	89.3
en→sr_Cyrl	64.2	61.8	83.3	61.8	74.5	87.2
en→sv	80.4	69.2	91.0	81.3	85.1	92.3
en→th	77.1	62.6	87.9	62.6	80.4	90.6
en→tr	71.8	69.4	85.2	69.4	80.2	87.9
en→vi	77.7	70.9	87.3	70.9	83.2	88.6