

Yandex at WMT26: Translation Adaptation at 235B and Hard-Label Distillation to 8B

Dmitry Popov Elizaveta Kozlova Ekaterina Taracheva
Elizaveta Makhotina Artem Mekhraliev Miron Baratelia
Ekaterina Enikeeva Nikolay Karpachev
Yandex

Abstract

We describe YANDEX and YANDEX-8B, our unconstrained and constrained submissions to the WMT26 General Machine Translation task. Our previous WMT25 submission, a 7B translation specialist, shows strong Fluency but comparatively weaker Accuracy. To combine its translation behavior with the capacity of a much larger model without replaying the full 7B training pipeline at 235B scale, we add two forms of supervision to the instruction-tuning data of an in-house 235B pretrained model: direct translation targets generated by the 7B specialist and in-house translation-related tasks. The resulting YANDEX model improves both Accuracy and Fluency and reduces the LLM-based error score relative to the same 235B model tuned only on general instructions. Relative to the 7B specialist, it improves Accuracy and reduces the error score, while their Fluency point estimates remain similar. YANDEX then provides hard targets for YANDEX-8B, which is trained for English-to-Russian, English-to-Belarusian, English-to-Kazakh, and English-to-Armenian using direct and instruction-conditioned translation data. Across four directions, YANDEX-8B obtains the highest paragraph-level ChrF++ and the highest Accuracy and Fluency under our LLM-based evaluation among five public baselines.

1 Introduction

We present the Yandex submissions to the WMT26 General MT shared task (Kocmi et al., 2026). General-purpose large language models (LLMs) are capable translation systems, but competitive machine translation (MT) systems still benefit from task-specific post-training. Existing recipes (Xu et al., 2024a; Karpachev et al., 2025; Wang et al., 2025; Kocmi et al., 2025a) combine monolingual adaptation, supervised fine-tuning (SFT) on parallel data, preference optimization, teacher-generated data, and quality-aware decoding. Our WMT24 (Elshin et al., 2024) and

WMT25 (Karpachev et al., 2025) submissions combine different subsets of supervised fine-tuning, preference optimization, structure-preserving supervision, MT post-pretraining, and local-context decoding. Other WMT25 systems (Zheng et al., 2025; Wang et al., 2025; Kocmi et al., 2025a) additionally use multi-teacher distillation, reinforcement learning, post-editing, and Minimum Bayes Risk decoding. Repeating a complete multi-stage recipe becomes expensive for a model with hundreds of billions of parameters. We ask whether an existing translation specialist can instead transfer useful translation behavior through the instruction-tuning pipeline of a much larger model.

Our starting point is the team’s 7B English–Russian WMT25 submission, published as Yandex+WMT and referred to here as YANDEX-7B-WMT25 (Karpachev et al., 2025). In our current ORBIT-SC evaluation (Section 3.3), it shows high Fluency but substantially weaker Accuracy than the evaluated frontier LLMs. The system also used local decoding blocks because translation quality declined for inputs of several hundred tokens. We use it as a source of fluent translation targets and ask whether a larger model can improve accuracy and context-sensitive translation choices.

For the unconstrained submission, we use an in-house 235B pretrained model. It was initialized from Qwen3-235B weights (Yang et al., 2025) and then trained using an in-house pretraining pipeline. We tune two models from this checkpoint. GENERAL-235B uses general Russian- and English-language instruction data. YANDEX uses the same general data augmented with direct translation targets from YANDEX-7B-WMT25 and in-house translation-related tasks. This keeps YANDEX in the instruction setup anticipated for WMT26.

YANDEX then provides hard targets for our constrained 8B model. YANDEX-8B is trained for English-to-Russian, English-to-Belarusian,

English-to-Kazakh, and English-to-Armenian. Its data combines direct translation examples, matched instruction-conditioned examples, and dedicated JSON and HTML subsets.

Our contributions are:

- We adapt an in-house 235B model using targets from our 7B WMT25 specialist and translation-related instruction data. The resulting YANDEX model improves Accuracy and Fluency over the same 235B model tuned only on general instructions. Relative to the 7B model, it improves Accuracy while their Fluency point estimates remain similar.
- We use YANDEX as the teacher to train an 8B model for the constrained track on direct and instruction-conditioned examples in four translation directions. YANDEX-8B obtains the highest paragraph ChrF++ and the highest LLM-based Accuracy and Fluency among five evaluated public baselines in every direction.

2 Related Work

2.1 Multi-stage MT adaptation

Decoder-only LLMs can be adapted to MT through task-specific training. ALMA (Xu et al., 2024a) combines monolingual language adaptation with fine-tuning on a small, high-quality parallel corpus. Our previous English–Russian system (Karpachev et al., 2025) uses MT post-pretraining, parallel-data SFT, and two alignment stages. Synthetic supervision includes decoded teacher sequences, established in sequence-level knowledge distillation (Kim and Rush, 2016), and LLM outputs used as SFT targets by recent WMT systems.

Learning from human preferences is a common post-training paradigm for LLMs. DPO (Rafailov et al., 2023) directly optimizes a model on preferred and dispreferred responses without fitting a separate reward model. For MT, CPO (Xu et al., 2024b) uses a contrastive objective over preferred and dispreferred translations, although improvements under an alignment metric need not remain stable across neural and lexical metrics (Gisserot-Boukhlef et al., 2024). Command-A-Translate (Kocmi et al., 2025a) combines preference optimization with post-editing and limited Minimum Bayes Risk decoding. Shy-hunyuan-MT (Zheng et al., 2025) distills candidates from multiple LLMs and subsequently applies Group

Relative Policy Optimization with a composite reward. Marco-MT (Wang et al., 2025) similarly assembles a multi-stage pipeline comprising two-stage SFT with data distillation, two-stage reinforcement learning, and hybrid alignment-aware and Minimum Bayes Risk decoding. Together, these systems combine multiple data, training, alignment, and decoding stages whose repetition at hundreds of billions of parameters is computationally costly.

2.2 Instruction-conditioned MT

Instruction-conditioned MT extends language-direction prompts with requirements concerning style, terminology, domain, or output form. ParroT (Jiao et al., 2023) represents such requirements through translation hints in an instruction-tuned LLM. The Tower model family (Alves et al., 2024) builds TOWERBASE through multilingual continued pretraining and TOWERINSTRUCT through subsequent instruction tuning on tasks from the translation workflow. Raunak et al. (2024) distill instruction-following behavior into a substantially smaller NMT model. Their setup combines formality control, domain adaptation, and multimodal translation, and supports multiple instructions at once. Together, these methods encode translation requirements beyond the source and target languages.

2.3 MT evaluation protocols and metrics

Human MT evaluation ranges from fine-grained expert error analysis under MQM (Freitag et al., 2021) to ESA (Kocmi et al., 2024), which combines error-span marking with a continuous quality score. Automatic evaluation includes learned scalar metrics such as MetricX-24 (Juraska et al., 2024) and error-aware metrics such as xCOMET (Guerreiro et al., 2024), which also localizes and categorizes translation errors. LLM-based evaluators can instead elicit MQM-style annotations directly, as in GEMBA-MQM (Kocmi and Federmann, 2023). Scalar metrics support aggregate comparison, while span-based outputs localize predicted errors.

Translation quality is multidimensional. RATE (Popov et al., 2025) separates Accuracy and Fluency and uses a finer error-severity scale than conventional MT evaluation protocols. Its expert English–Russian evaluation found that strong MT outputs could exceed human translations in Accuracy while remaining below them in Fluency.

System	$z_{\text{Acc}} \uparrow$	System	$z_{\text{FL}} \uparrow$
Gemini-3.1-Pro	1.04	Gemini-2.5-Pro	0.61
GPT-5.5	0.83	Gemini-3.1-Pro	0.45
Claude-4.6-Opus	0.72	Yandex-7B-WMT25	0.37
Gemini-2.5-Pro	0.69	Yandex	0.34
Yandex	0.22	Claude-4.6-Opus	0.17
GPT-4.1	-0.37	GPT-4.1	0.09
General-235B	-0.69	General-235B	-0.12
Claude-4	-1.38	GPT-5.5	-0.32
Yandex-7B-WMT25	-2.05	Claude-4	-2.83

System	$z_{\text{A/F}} \uparrow$	System	$z_{\text{MQM}} \downarrow$
Gemini-3.1-Pro	0.83	Gemini-3.1-Pro	-1.06
Gemini-2.5-Pro	0.74	Gemini-2.5-Pro	-0.78
Claude-4.6-Opus	0.49	Claude-4.6-Opus	-0.66
Yandex	0.32	GPT-5.5	-0.49
GPT-5.5	0.23	Yandex	-0.19
GPT-4.1	-0.13	GPT-4.1	0.22
General-235B	-0.44	General-235B	0.49
Yandex-7B-WMT25	-0.82	Yandex-7B-WMT25	1.66
Claude-4	-2.52	Claude-4	1.85

Table 1: Judge-normalized paragraph-level English-to-Russian results on WMT25. Systems are sorted separately in each panel. Green, gray, and red indicate significantly better, not distinguished, and significantly worse results than YANDEX under the paired bootstrap.

Because this trade-off is central to our analysis, we report Accuracy and Fluency separately.

3 Yandex: Unconstrained System

3.1 Starting Points and Motivation

We use three names for the large-system experiments. YANDEX-7B-WMT25 is our previously published 7B translation specialist. GENERAL-235B and YANDEX are tuned from the same in-house 235B pretrained model, which was initialized from Qwen3-235B weights (Yang et al., 2025) and then trained using an in-house pretraining pipeline. GENERAL-235B is tuned on an in-house general instruction dataset composed primarily of Russian- and English-language instructions. YANDEX uses the same general data augmented with the translation and translation-related data described below. It is our WMT26 unconstrained submission.

The human RATE evaluation reported in our WMT25 system paper (Karpachev et al., 2025) shows a fluency-oriented quality profile for YANDEX-7B-WMT25: it receives higher Fluency and Style but lower Accuracy than the evaluated GPT-4.1 and Claude 4 outputs. Replaying its complete post-training pipeline at 235B scale would be costly. Instead, we use its translations as targets within the instruction tuning of the larger model and test whether this setup can combine the 7B model’s Fluency with the higher Accuracy and

better context-sensitive translation choices of the general-purpose 235B model.

3.2 Translation-Augmented Instruction Tuning

Most of the added supervision consists of direct English-to-Russian translation. We pair diverse English source texts with prompts that request only the translated text. YANDEX-7B-WMT25 generates the targets.

The mixture also includes in-house translation-related tasks, for example, explaining a word in context, changing the style of a translation, applying a glossary, editing a translation, or assessing translation quality. We retain general instruction data so that YANDEX remains an instruction-following model. This design reflects our expectation that WMT26 inputs would express diverse translation requirements through instructions.

3.3 Paragraph-Level Translation Quality

We evaluate English-to-Russian translation on 332 paragraphs from the WMT25 General MT test set (Kocmi et al., 2025b) spanning 87 documents and four domains. We use two ORBIT-SC (Popov et al., 2026) runs backed by GPT-5.4 and Claude-4.6-Opus. ORBIT-SC is our single-call LLM implementation of the RATE protocol (Popov et al., 2025). It assigns separate Accuracy and Fluency

scores and predicts error spans. For each translation, we compute the corresponding MQM score as $5N_{\text{major}} + N_{\text{minor}}$.

For each judge, we first compute A/F as the arithmetic mean of Accuracy and Fluency. We then macro-average the four domains and standardize Accuracy, Fluency, A/F, and MQM separately over the systems shared by both evaluation runs. Finally, we average the corresponding judge-wise z -scores. Unlike the four-direction 8B evaluation in Section 4, which uses GPT-5.4 alone, this comparison gives equal weight to two judges with different score scales.

Table 1 shows that the translation-augmented mixture performs significantly better than GENERAL-235B on all four measurements under the paired bootstrap.

Compared with YANDEX-7B-WMT25, YANDEX has higher z_{Acc} and $z_{\text{A/F}}$ and lower z_{MQM} . Their z_{FL} point estimates are close. This profile shows improved Accuracy with a similar Fluency point estimate to the 7B specialist. Since all translation-related components are added together, the result cannot be attributed to the 7B targets alone.

The external systems illustrate the same accuracy–fluency trade-off. GPT-5.5 has higher z_{Acc} than YANDEX, whereas YANDEX has higher z_{FL} . Their $z_{\text{A/F}}$ and z_{MQM} point estimates are close. YANDEX outperforms GPT-4.1 and Claude 4 on $z_{\text{A/F}}$ and z_{MQM} . Claude-4.6-Opus has higher z_{Acc} and lower z_{MQM} than YANDEX, while their $z_{\text{A/F}}$ and z_{FL} point estimates are close. The two Gemini systems obtain the strongest aggregate scores under this evaluation.

3.4 Terminology and Named Entities

We additionally evaluate the three systems on an in-house 820-example terminology and named-entity benchmark. The set targets translation choices that require domain context or model knowledge. For example, *opening* has a specialized translation in Russian chess terminology that differs from its translation outside the chess context.

Errors are categorized as Critical, Major, or Minor. For N examples, we compute

$$\text{Penalty} = \frac{6N_{\text{critical}} + 3N_{\text{major}} + N_{\text{minor}}}{N}, \quad (1)$$

where lower is better.

GENERAL-235B reduces the penalty from 0.765 for the 7B specialist to 0.378 on this context- and

System	Penalty ↓	Critical	Major	Minor
Yandex	0.340	2	27	186
General-235B	0.378	1	35	199
Yandex-7B-WMT25	0.765	18	87	258

Table 2: Results on the in-house terminology and named-entity benchmark ($N = 820$ for every system).

knowledge-sensitive evaluation. The translation-augmented YANDEX model further reduces it to 0.340, with fewer Major and Minor errors than GENERAL-235B. Together with Table 1, these results show the intended complementary profile: the 235B models obtain lower terminology penalties than the 7B specialist, while YANDEX retains a similar z_{FL} point estimate.

4 Yandex-8B: Constrained System

4.1 Task Setting and Starting Model

Our constrained submission, YANDEX-8B, is trained for English-to-Russian, English-to-Belarusian, English-to-Kazakh, and English-to-Armenian translation. We initialize it from the approximately 8B-parameter [YandexGPT-5-Lite-8B](#) pretrained checkpoint (Yandex, 2025). We release the submitted weights on Hugging Face.¹

During development, the WMT26 General MT task description emphasized instruction-conditioned translation and ESA-based evaluation (WMT26 General MT Organizers, 2026). We expected prompts to specify domain, style, terminology, or structure while requiring only the translation as the response. The final test input contained four prompt templates, but we trained on a broader range of translation requirements.

4.2 Instruction and Translation Data

We collect English source texts from news, social, spoken-dialogue, and literary domains, together with JSON and HTML inputs. Translation instructions, usually written in Russian, are selected from an in-house instruction collection. We retain requests whose response should contain only a translation. For each source text, the 235B model selects ten candidate instructions. A second model rates their compatibility with the text on a 1–10 scale. We retain pairs scoring above 8. The instructions are independent of the target language. We specify the translation direction separately, allowing the

¹[foksly/wmt26-constrained-submission](#)

System	en-ru		en-be		en-kk		en-hy	
	ChrF++ ↑	MetricX ↓	ChrF++ ↑	MetricX ↓	ChrF++ ↑	MetricX ↓	ChrF++ ↑	MetricX ↓
Yandex-8B	62.3	1.58	58.0	3.87	55.3	3.44	53.5	5.41
Yandex-8B-CPO	63.0	1.54	58.1	3.86	55.2	3.56	54.8	5.09
TranslateGemma-12B	62.2	1.52	54.9	3.83	47.9	5.60	50.1	6.30
Qwen-3.5-9B	60.1	2.22	49.1	5.73	48.3	5.52	49.6	7.32
MADLAD-400-10B	59.4	4.79	54.2	5.56	46.8	5.86	50.3	7.13
NLLB-200-3.3B	57.4	6.01	52.1	7.63	47.8	7.10	52.8	6.35
GPT-OSS-20B	51.9	2.72	29.3	6.16	49.0	5.82	12.9	9.18

Table 3: Paragraph-level ChrF++ (↑) and MetricX (↓) on BOUQuET. Best scores are in bold.

System	en-ru			en-be			en-kk			en-hy		
	Accuracy ↑	Fluency ↑	MQM ↓	Accuracy ↑	Fluency ↑	MQM ↓	Accuracy ↑	Fluency ↑	MQM ↓	Accuracy ↑	Fluency ↑	MQM ↓
Yandex-8B	81.6	85.3	11.3	69.1	69.5	24.2	69.9	72.0	22.3	52.2	54.0	37.6
Yandex-8B-CPO	81.6	84.1	11.2	69.2	68.7	23.5	70.4	71.8	22.0	51.5	53.8	35.9
TranslateGemma-12B	73.9	81.2	17.9	59.1	64.4	30.8	47.2	52.5	40.9	41.8	48.1	43.8
Qwen-3.5-9B	69.6	72.7	21.4	51.1	51.8	38.5	46.2	51.4	40.8	39.1	44.8	46.4
MADLAD-400-10B	32.5	32.5	34.5	40.5	44.8	38.6	40.8	48.4	40.1	28.3	37.5	31.7
NLLB-200-3.3B	39.7	40.3	39.5	33.8	33.8	46.0	32.2	35.7	45.5	35.0	36.4	46.2
GPT-OSS-20B	62.4	63.5	27.2	40.9	39.3	46.8	40.8	42.3	44.5	30.2	29.4	52.4

Table 4: ORBIT-SC results on WMT25 General MT source paragraphs. Best scores at the displayed precision are in bold, except for the marked MADLAD value discussed in the text.

same matched instruction–source pair to be reused across all four directions.

The resulting mixture contains matched instruction-conditioned examples and direct translation examples. It also contains dedicated JSON and HTML examples with format-preservation instructions. All targets are hard labels generated by the 235B YANDEX teacher. The combined dataset contains approximately one million examples across the four directions. We tune the 8B starting checkpoint on this mixture. YANDEX-8B denotes this supervised checkpoint and is the model submitted to WMT26.

We also tested a two-stage route to adapt the model more explicitly to Belarusian, Kazakh, and Armenian. The first stage used only URL-mined parallel data, with at least 50 million examples per direction, and the second used the instruction mixture above. We identify multilingual versions of web pages through language-specific URL variants and align the corresponding pages at the paragraph or sentence level. This route did not improve our development results. The hard-label mixture may already provide sufficient coverage, while URL-aligned web pages can contain machine-translated or otherwise low-quality parallel text and may require stronger filtering.

4.3 Contrastive Preference Optimization

Starting from the supervised checkpoint, we test an additional Contrastive Preference Optimization (CPO) (Xu et al., 2024b) stage. We generate candidate translations with Gemma-4-31B-IT (Gemma Team, 2026), DeepSeek-V4-Pro (DeepSeek-AI, 2026), GLM-5.2 (Z.ai, 2026), Qwen-3.7 (Qwen Team, 2026b), and YANDEX. DeepSeek-V4-Pro compares sampled candidate pairs side by side to produce preferred and dispreferred translations. We report one representative development checkpoint as YANDEX-8B-CPO. It was not submitted.

4.4 Evaluation

For reference-based evaluation, we use paragraph-level BOUQuET (Andrews et al., 2025), which provides references for all four directions. Table 3 reports ChrF++ (Popović, 2017) with its default configuration and MetricX-24-XXL (Juraska et al., 2024). To reflect the constrained track’s 20B-parameter limit, all five public baselines have at most 20B parameters: TranslateGemma-12B (Finkelstein et al., 2026), Qwen-3.5-9B (Qwen Team, 2026a), MADLAD-400-10B (Kudugunta et al., 2023), NLLB-200-3.3B (NLLB Team, 2024), and GPT-OSS-20B (OpenAI, 2025).

For reference-free evaluation, we evaluate translations of official English source paragraphs from WMT25 General MT (Kocmi et al., 2025b) with ORBIT-SC, using GPT-5.4 as a single fixed judge.

We report Accuracy, Fluency, and the MQM score defined in Section 3.3, and macro-average each score within domain and then equally across domains.

4.5 Results and Analysis

As shown in Table 3, the submitted YANDEX-8B obtains the highest paragraph ChrF++ among the five public baselines in all four directions. Its MetricX score is best among those baselines for English-to-Kazakh and English-to-Armenian, while TranslateGemma-12B is slightly better on English-to-Russian and English-to-Belarusian. Table 4 shows that YANDEX-8B also obtains the highest ORBIT-SC Accuracy and Fluency scores among the public baselines in every direction.

The CPO checkpoint produces mixed changes. ChrF++ improves on three directions and decreases slightly on English-to-Kazakh. MetricX improves on three directions but also worsens on English-to-Kazakh. ORBIT-SC Accuracy moves in both directions, and Fluency is lower on all four directions. MQM improves on all four directions, but Accuracy and Fluency do not show a consistent benefit. We therefore selected the supervised checkpoint for the submission.

The marked MADLAD-400-10B English-to-Armenian score illustrates a limitation of span-count aggregation. It has a low MQM score despite low Accuracy and Fluency. A diagnostic proxy flags 99 of 332 outputs as likely non-translations, including 91 with Accuracy zero. These outputs receive a lower mean MQM score than the remaining outputs (22.97 versus 35.64) because their critical spans are fewer but much longer on average (129 versus 37 characters). We therefore interpret MQM together with Accuracy and Fluency. The proxy is not a human label and is used only to diagnose the aggregation behavior.

5 Conclusion

We described two WMT26 systems connected through hard-label training. Our 7B WMT25 translation specialist generates English-to-Russian targets for the instruction tuning of an in-house 235B model. The resulting YANDEX system improves both Accuracy and Fluency over the same 235B model tuned only on general instructions. YANDEX then becomes the teacher for YANDEX-8B, our constrained model for four translation directions.

On BOUQuET, YANDEX-8B obtains the highest paragraph ChrF++ among five evaluated public baselines in every direction. On WMT25 General MT source paragraphs, it obtains the highest ORBIT-SC Accuracy and Fluency among the same baselines in every direction. An additional CPO stage produces small and inconsistent changes, so the submitted system is the supervised checkpoint. Overall, the results show how a compact MT specialist, a large instruction model, and a constrained multilingual student can be connected without repeating the complete multi-stage MT pipeline at the largest scale.

References

- 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](#). In *First Conference on Language Modeling*.
- Pierre Andrews, Mikel Artetxe, Mariano Coria Meglioli, Marta R. Costa-jussà, Joe Chuang, David Dale, Mark Duppenthaler, Nathaniel Paul Ekberg, Cynthia Gao, Daniel Edward Licht, Jean Maillard, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Eduardo Sánchez, Ioannis Tsiamas, Arina Turkatenco, Albert Ventayol-Boada, and Shireen Yates. 2025. [BOUQuET: Dataset, benchmark and open initiative for universal quality evaluation in translation](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27515–27535, Suzhou, China. Association for Computational Linguistics.
- DeepSeek-AI. 2026. [DeepSeek-V4: Towards highly efficient million-token context intelligence](#). *arXiv preprint arXiv:2606.19348*.
- Denis Elshin, Nikolay Karpachev, Boris Gruzdev, Ilya Golovanov, Georgy Ivanov, Alexander Antonov, Nickolay Skachkov, Ekaterina Latypova, Vladimir Layner, Ekaterina Enikeeva, Dmitry Popov, Anton Chekashev, Vladislav Negodin, Vera Frantsuzova, Alexander Chernyshev, and Kirill Denisov. 2024. [From general LLM to translation: How we dramatically improve translation quality using human evaluation data for LLM finetuning](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 247–252, Miami, Florida, USA. Association for Computational Linguistics.
- Mara Finkelstein, Isaac Caswell, Tobias Domhan, Jan-Thorsten Peter, Juraj Juraska, Parker Riley, Daniel Deutsch, Geza Kovacs, Cole Dilanni, Colin Cherry, Eleftheria Briakou, Elizabeth Nielsen, Jiaming Luo, Kat Black, Ryan Mullins, Sweta Agrawal, Wenda Xu, Erin Kats, Stephane Jaskiewicz, and 2 others. 2026.

- TranslateGemma technical report. *arXiv preprint arXiv:2601.09012*.
- 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.
- Gemma Team. 2026. [Gemma 4 technical report](#). *arXiv preprint arXiv:2607.02770*.
- Hippolyte Gisserot-Boukhlef, Ricardo Rei, Emmanuel Malherbe, Céline Hudelot, Pierre Colombo, and Nuno M. Guerreiro. 2024. [Is preference alignment always the best option to enhance LLM-based translation? an empirical analysis](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 1373–1392, Miami, Florida, USA. Association for Computational Linguistics.
- Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. [xCOMET: Transparent machine translation evaluation through fine-grained error detection](#). *Transactions of the Association for Computational Linguistics*, 12:979–995.
- Wenxiang Jiao, Jen-tse Huang, Wenxuan Wang, Zhiwei He, Tian Liang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. [ParroT: Translating during chat using large language models tuned with human translation and feedback](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15009–15020, 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.
- Nikolay Karpachev, Ekaterina Enikeeva, Dmitry Popov, Arsenii Bulgakov, Daniil Pantelev, Dmitrii Ulianov, Artem Kryukov, and Artem Mekhraliev. 2025. [Yandex submission to the WMT25 general machine translation task](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 740–752, Suzhou, China. Association for Computational Linguistics.
- Yoon Kim and Alexander M. Rush. 2016. [Sequence-level knowledge distillation](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.
- Tom Kocmi, Arkady Arkhangorodsky, Alexandre Berard, Phil Blunsom, Samuel Cahyawijaya, Théo Dehaze, Marzieh Fadaee, Nicholas Frosst, Matthias Galle, Aidan Gomez, Nithya Govindarajan, Wei-Yin Ko, Julia Kreutzer, Kelly Marchisio, Ahmet Üstün, Sebastian Vincent, and Ivan Zhang. 2025a. [Command-A-Translate: Raising the bar of machine translation with difficulty filtering](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 789–799, Suzhou, China. 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 Lakouga, Jessica Lundin, Christof Monz, Kenton Murray, and 10 others. 2025b. [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*, pages 355–413, Suzhou, China. Association for Computational Linguistics.
- Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Sergey Dukanov, Anton Dvorkovich, Mark Fishel, Markus Freitag, Samuel Frontull, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Stan Kharevich, Philipp Koehn, Zheng Li, Jean Maillard, Christof Monz, Alexander Murauski, and 12 others. 2026. [Findings of the WMT26 general machine translation shared task: Contrastive dynamic human evaluation at scale](#). In *Proceedings of the Eleventh Conference on Machine Translation*. Association for Computational Linguistics.
- Tom Kocmi and Christian Federmann. 2023. [GEMBA-MQM: Detecting translation quality error spans with GPT-4](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.
- Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024. [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.
- 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](#). In *Advances in Neural Information Processing Systems*, volume 36.
- NLLB Team. 2024. [Scaling neural machine translation to 200 languages](#). *Nature*, 630:841–846.
- OpenAI. 2025. [gpt-oss-120b and gpt-oss-20b model card](#).
- Dmitry Popov, Roman Bokhyan, Ekaterina Enikeeva, Artem Mekhraliev, Stepan Vysotsky, and Nikolay Karpachev. 2026. [Beyond prompts: A systematic study of LLM-based machine translation evaluators](#). In *Findings of the Association for Computational Linguistics: EMNLP 2026*.

- Dmitry Popov, Vladislav Negodin, Ekaterina Enikeeva, Iana Matrosova, Nikolay Karpachev, and Max Ryabinin. 2025. [Refined assessment for translation evaluation: Rethinking machine translation evaluation in the era of human-level systems](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22079–22095, Suzhou, China. Association for Computational Linguistics.
- Maja Popović. 2017. [chrF++: words helping character n-grams](#). In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.
- Qwen Team. 2026a. [Qwen3.5: Towards native multi-modal agents](#).
- Qwen Team. 2026b. [Qwen3.7 model release](#). Accessed 7 August 2026.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. [Direct preference optimization: Your language model is secretly a reward model](#). In *Advances in Neural Information Processing Systems*, volume 36.
- Vikas Raunak, Roman Grundkiewicz, and Marcin Junczys-Dowmunt. 2024. [On instruction-finetuning neural machine translation models](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 1155–1166, Miami, Florida, USA. Association for Computational Linguistics.
- Hao Wang, Linlong Xu, Heng Liu, Yangyang Liu, Xiaohu Zhao, Bo Zeng, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2025. [Marco large translation model at WMT2025: Transforming translation capability in LLMs via quality-aware training and decoding](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 587–593, Suzhou, China. Association for Computational Linguistics.
- WMT26 General MT Organizers. 2026. [WMT26 general machine translation shared task](#). Accessed 7 August 2026.
- Haoran Xu, Young Jin Kim, Amr Mohamed Nabil Aly Aly Sharaf, and Hany Awadalla. 2024a. [A paradigm shift in machine translation: Boosting translation performance of large language models](#). In *International Conference on Learning Representations*.
- Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024b. [Contrastive preference optimization: Pushing the boundaries of LLM performance in machine translation](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 55204–55224. PMLR.
- Yandex. 2025. [YandexGPT-5-Lite-8B pretrained model](#).
- An Yang et al. 2025. [Qwen3 technical report](#). *arXiv preprint arXiv:2505.09388*.
- Z.ai. 2026. [GLM-5.2 model card](#).
- Mao Zheng, Zheng Li, Yang Du, Bingxin Qu, and Mingyang Song. 2025. [Shy-hunyuan-MT at WMT25 general machine translation shared task](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 607–613, Suzhou, China. Association for Computational Linguistics.