

UvA-MT’s Participation in the WMT26 General Translation Shared Task

Di Wu* Sergey Troshin* Wafaa Mohammed Seth Aycock
Evgeniia Tokarchuk Vlad Niculae Christof Monz
Language Technology Lab, University of Amsterdam
{d.wu, s.troshin, w.m.a.mohammed, s.aycock}@uva.nl
{e.tokarchuk, v.niculae, c.monz}@uva.nl

Abstract

This paper presents UvA-MT’s submission to the WMT 2026 General Machine Translation shared task, competing in the unconstrained track across all 23 translation directions. This year, we fully leverage the test-time methods of Large Language Models (LLMs) for machine translation. Specifically, (1) we use sequential sampling with a refinement prompt to generate a pool of translation candidates for each source sentence, a strategy shown to outperform traditional parallel sampling (Wu et al., 2026); and (2) we employ LLMs as pairwise judges to select the best candidates via a round-robin voting mechanism, with offline evaluation on the WMT25 benchmark showing a **very clear** improvement over point-wise best-of- N selection. Lastly, we simply apply these two strategies for the submission of WMT26.

1 Introduction

In this paper, we describe our submission to the WMT 2026 General Machine Translation shared task (unconstrained track), which covers 23 translation directions. Motivated by recent advances in Large Language Models (LLMs), particularly test-time scaling techniques (Tan et al., 2025; Wu et al., 2025a, 2026), we investigate to what extent translation quality can be improved by relying solely on general-purpose LLMs at inference time.

Figure 1 illustrates an overview of our approach. Our submission uses two test-time technologies:

- **Sequential Sampling:** Given a source input to an LLM, a pool of translations is sequentially sampled, with each subsequent translation instructed to improve upon its predecessor. This recursive sampling strategy exploits the LLM’s self-improvement capability and has been shown by Wu et al. (2026) to be more effective than traditional parallel sampling, especially under limited inference budgets.

- **Contrastive Round-Robin Voting:** Given a pool of translation candidates, every pair is compared head-to-head by an LLM judge that explicitly contrasts their differences (rather than scoring each independently). Wins are tallied across all pairwise match-ups, and the candidate with the most victories is selected as the final output. We argue that this strategy is more reliable than traditional best-of- N sampling because pairwise judgments require only relative preferences rather than absolute quality scores, reducing the need for calibration and enabling finer-grained discrimination among candidates (Ouyang et al., 2022).

Section 2 describes our main methodology, which consists of two components: 1) sequential sampling, i.e., how to generate a pool of translation candidates for each source sentence; and 2) contrastive round-robin voting with a pairwise LLM judge, i.e., how to select the best candidate from the resulting pool.

In Section 3.1, we conduct offline evaluations on the WMT24 benchmark (Kocmi et al., 2024a; Deutsch et al., 2025), following Wu et al. (2026), showing that **sequential sampling** yields substantial translation improvements over parallel sampling, especially under limited sampling budgets.

In Section 3.2, we show that **contrastive round-robin voting** mechanism, powered by an LLM-based pairwise (contrastive) judge, consistently outperforms the traditional best-of- N sampling paradigm with a pointwise judge (metric model, e.g., CometKiwi). This is reflected in both improved Best@ N performance and stronger rank correlation, as measured by Kendall’s τ and Spearman’s ρ , on the WMT25 metrics benchmark (Lavie et al., 2025).

Lastly, we apply these two strategies for the submission of WMT26, see details in Section 4.

*Equal first-author contribution.

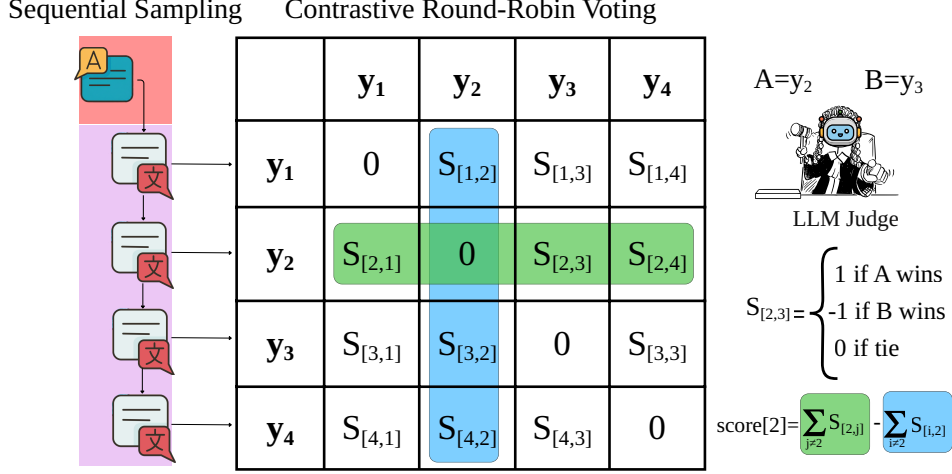


Figure 1: An overview of our pipeline: (left) sequential sampling to generate a list of candidates, see Section 2.1 for details; (right) contrastive round-robin voting with a LLM-as-a-Judge to select the best sample. Note that, to eliminate positional bias in LLM-based pairwise judging, each pair of samples is evaluated twice with their order swapped. The final score is computed by combining the scores that a hypothesis achieves in both directions, see Section 2.2 for details.

2 Methodology

2.1 Generation of N candidates

Our generation setting follows the previous and concurrent line of work, which we refer to as *sequential* sampling (Wu et al., 2025a, 2026). The key idea of this sampling strategy is to refine the previously generated hypotheses, instead of generating fully in parallel, which demonstrated higher sampling efficiency given smaller sampling budget. In particular, to obtain a set of candidates for reranking, we translate each document sequentially 8 times, including the initial translation and 7 refinement rounds. Formally, for a sampling turn t , *sequential* sampling generates a translation y as follows:

$$y^{(t)} \sim p_{\theta} \left(\cdot \mid I^{(r)}, y^{(t-1)}, I(x) \right), \quad (1)$$

where $I^{(r)}$ is a prompt requesting a model to revisit a translation to improve it (Wu et al., 2025a); and $I(x)$ includes a system prompt and the initial turn translation. To reduce the computational cost, here we use the history -1 formulation of the sequential sampling, shown to be most efficient and effective in Wu et al. (2026), feeding only the most recent previous translation to the model, thus capping the context size growth.

We use the following prompt template for the candidate generation step:

[system]
You are a professional translator. Follow the instruction and output only the translation, with no explanations.

Task-specific instruction:
{instruction}

[user turn initial]
Please translate the following text from {src_lang} to {tgt_lang}. Provide only one translation on the first line and do not output anything else after that.

{src_lang}: {source}
{tgt_lang}:

[assistant turn]
{assistant response}

[user turn revisiting]
Please again translate the following text from {src_lang} to {tgt_lang} to make it better. Provide only one translation on the first line and do not output anything else after that.

{src_lang}: {source}
{tgt_lang}:

[assistant turn]
{assistant response}

2.2 Contrastive Judge

We use the following prompt template for the pairwise judges:

```
[system]
You are an expert, impartial bilingual
translation evaluator. You compare two
candidate translations of the same source and
decide which is better. Judge strictly on
quality; ignore which candidate is labelled A
or B and ignore their order or length.
```

```
[user]
Compare two candidate translations and decide
which is better.

Source language: {src_lang}
Target language: {tgt_lang}

Judge against this rubric (earlier criteria
matter more):
Pick the translation that is more accurate and
more natural in the target language.

{instruction_block}=== SOURCE ===
{source}

=== TRANSLATION A ===
{trans_a}

=== TRANSLATION B ===
{trans_b}

Return only a single JSON object on the last
line and nothing after it:
{"winner": "A" | "B" | "tie"}

[assistant]
{assistant response}
```

The ‘{instruction_block}’ here is populated when the record carries a non-trivial instruction. Then it resolves to ‘The translations were produced under this task-specific instruction; also judge how faithfully each candidate follows it: \n{instruction}’. Trivial “Translate from X to Y.” instructions are dropped, as we incorporate the source and target language information directly into the user prompt. We use the constrained generation for json, implemented on the server side, whereas we provide the json schema.

We use temperature 0 for the judge, and we choose the lowest possible reasoning effort for each model.

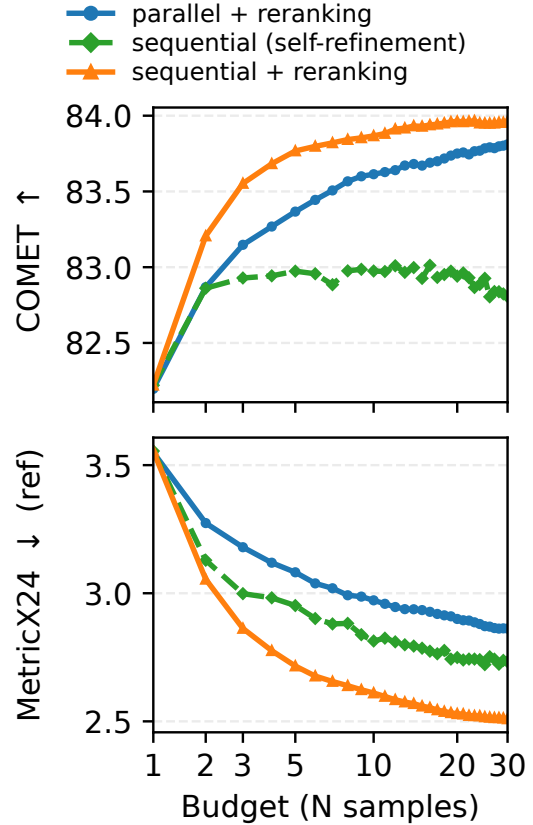


Figure 2: Sequential + reranking demonstrates superior quality@budget trade-off, compared to the parallel sampling + reranking, and sequential sampling (self-refinement), following Wu et al. (2026). Here we report the quality of the methods on WMT24++ with metrics aggregated over the 6 languages.

2.3 Contrastive Round-Robin Voting

The idea of a contrastive round-robin voting manifests itself as early as in the works of Llull (c. 1274-83),¹ where they propose to evaluate the candidates during election via full round-robin contrastive voting, such that a candidate who is victorious in most of the pairwise comparisons is named the winner (also known as Copeland’s method).

In our formulation, the contrastive round-robin voting with LLMs-as-a-judge consists of the pairwise judge scoring and the score aggregation. By solely relying on contrastive LLMs-as-a-judge prompt templates to obtain pairwise votes, we use the following counting aggregation strategy for our robust evaluation. Formally, given a pool of N candidates, we define a $N \times N$ score matrix S , where $S[i, j] \in \{1, -1, 0\}$ denotes a voting result for a candidate i against a candidate j , where the verdict

¹assembled and translated by Hägele and Pukelsheim (2001)

1 stands for a win, -1 for a lose, and 0 for a tie. Then a ranking score of each candidate is a sum of its points: $\text{score}[i] = \sum_j S[i, j]$. LLM judges are known to have strong position biases (Wang et al., 2024), i.e., they can be inconsistent when changing a candidate order. To mitigate the position bias, we evaluate both i vs j and j vs i directions (the order of these is swapped in the judge prompt template). The de-biased score we use is as following:

$$\text{score}[i] = \sum_{j=1}^N S[i, j] - \sum_{i=1}^N S[j, i]. \quad (2)$$

3 Offline Evaluation

3.1 On the Effectiveness of Sequential Sampling

We choose sequential sampling as the main generation strategy to form a pool of samples for subsequent reranking. Here, we briefly report an abridged version of the results from Wu et al. (2026) in Figure 2, demonstrating the comparison of parallel and sequential strategy for the GPT-4o-mini² model.

Here, we compare **parallel + reranking**, **sequential (self-refinement)** and **sequential + reranking**. These results report the average machine translation quality on WMT24++, aggregated over 6 target language directions: Dutch, German, Mandarin, Standard Arabic, Russian, Romanian; measured using a reference-based COMET (Rei et al., 2022), and reference-based MetricX-24-Hybrid-XXL (ref) (Juraska et al., 2024). Here, CometKiwi-XL (Rei et al., 2023) is used as a metric for reranking, following Wu et al. (2026). The results (Figure 2) show that the combination of **sequential sampling with reranking** outperforms both the parallel sampling with reranking and simple self-refinement, motivating our choice for the sampling method.

In the next section (§3.2), we revisit the reranking step, the main focus of this submission.

3.2 On the Effectiveness of Contrastive Round-Robin Voting

Given a pool of hypotheses for each source text, we use Round-Robin Voting with commercial LLMs to select the winning hypothesis for the final submission. To show the effectiveness, we compare our strategy with traditional Best-of- N sampling,

known as Quality-aware decoding in MT (Fernandes et al., 2022), which uses machine translation quality estimation models for point-wise ranking, e.g., MetricX-QE or Comet-Kiwi.

We use the WMT25 General MT Human Evaluation dataset³ as the benchmark. We evaluate two translation directions, EN→ZH and EN→RU, which contain 198 and 200 paragraph-level source texts, respectively. Each source text is accompanied by translations produced by at most 18 MT systems, including last year’s leading systems i.e. GPT-4.1, Claude-4, UvA-MT, and Qwen3-235B (Achiam et al., 2023; Wu et al., 2025b; Yang et al., 2025; Kocmi et al., 2025).

For each source text in the dataset, we randomly sample 8 translations. We retain the corresponding human evaluation Error Span Annotation Kocmi et al. (2024b, ESA) scores for each translation (an ordinal score from 0 to 100). For translations annotated by multiple experts, we average the individual scores to obtain the final human evaluation score.

We evaluate pairwise judgments using contrastive Round-Robin Voting across different LLMs and compare them with traditional machine translation (MT) evaluation metrics by measuring their Kendall rank correlation (τ_a) with human judgments. The resulting rankings are presented in Table 1 and Table 2. Note that Kendall rank correlation is comparable across all of these systems. Additionally, we report Spearman’s rank correlation coefficient (ρ), Best@1, Best@2, rate of a tie, and disagreement rate.

We find that:

- Traditional MT metrics **completely fail** to rank last year’s translation hypotheses. Even the best-performing metric, COMET-Kiwi-23-XXL, achieves Kendall’s τ_a scores of only 5.2 and 8.6 in EN-ZH and EN-RU. Its Best@1 and Best@2 accuracies are approximately 12.5% and 25%, respectively, suggesting performance that is close to random guessing on this ranking task.
- Our pairwise LLM-as-a-Judge framework with Round-Robin Voting **substantially outperforms** all point-wise translation metrics, achieving consistent and very clear improvements in Kendall’s τ_a and Spearman’s ρ , as well as Best@1 and Best@2 accuracy.

²gpt-4o-mini-2024-07-18

³<https://github.com/wmt-conference/wmt25-general-mt>

Rank	Type	System	τ_a	ρ	Best@1	Best@2	Tie	Disagree
1	judge	Gemini-3.5-Flash (reasoning)	23.1	30.9	17.7	32.8	12.8	12.6
2	judge	GPT-5.4 (reasoning)	22.6	30.6	22.2	36.9	17.2	16.9
3	judge	Gemini-3.5-Flash (direct)	22.4	31.1	20.2	33.8	13.9	13.5
4	judge	GPT-5.4 (direct)	20.8	28.7	19.7	32.3	18.4	18.0
5	judge	GPT-4o-mini (direct)	16.7	25.9	21.7	33.8	31.6	31.3
6	judge	GPT-4o-mini (reasoning)	12.4	23.7	13.6	28.8	53.7	53.5
7	judge	Gemini-2.5-Flash (reasoning)	10.3	26.5	21.2	36.9	62.1	47.7
8	judge	Gemini-2.5-Flash (direct)	10.0	23.8	20.7	33.8	59.3	51.4
9	metric	COMET-Kiwi-23-XXL	5.2	6.9	14.1	25.8	0.3	0.0
10	metric	MetricX-24-Hyb-XL	4.9	6.8	17.2	29.3	0.3	0.0
11	metric	MetricX-24-Hyb-XXL	2.3	3.3	18.2	29.8	0.3	0.0
12	metric	COMET-Kiwi-22	1.6	2.6	15.2	25.8	0.3	0.0
13	metric	COMET-Kiwi-23-XL	1.6	2.6	15.7	25.3	0.3	0.0
14	metric	MetricX-23-QE-L	-2.8	-4.4	15.7	27.8	0.3	0.0

Table 1: EN-ZH results ($n = 198$ documents), reported as percentages. τ_a is the per-pair-verdict Kendall, which is the coefficient comparable across the judge/metric boundary and is used for system ranking. The best value among judge models is shown in bold. For Tie and Disagree, lower is better.

Rank	Type	System	τ_a	ρ	Best@1	Best@2	Tie	Disagree
1	judge	GPT-5.4 (direct)	30.8	42.5	36.0	51.5	17.2	16.8
2	judge	Gemini-3.5-Flash (reasoning)	30.8	41.7	37.0	52.0	11.9	11.6
3	judge	Gemini-3.5-Flash (direct)	30.6	41.6	36.0	52.0	10.6	10.3
4	judge	GPT-5.4 (reasoning)	29.7	40.1	35.5	54.5	16.8	16.5
5	judge	GPT-4o-mini (direct)	22.4	34.2	29.5	41.0	30.5	30.2
6	judge	GPT-4o-mini (reasoning)	19.8	33.4	28.0	43.0	43.7	43.4
7	judge	Gemini-2.5-Flash (reasoning)	16.0	38.3	32.0	47.5	53.2	45.0
8	judge	Gemini-2.5-Flash (direct)	15.7	35.9	32.5	48.0	54.3	48.3
9	metric	COMET-Kiwi-23-XXL	8.6	11.4	17.0	28.5	0.3	0.0
10	metric	MetricX-24-Hyb-XXL	6.4	8.1	17.5	28.5	0.7	0.0
11	metric	MetricX-24-Hyb-XL	6.3	7.5	13.0	22.5	0.3	0.0
12	metric	COMET-Kiwi-23-XL	4.9	6.1	16.0	28.5	0.3	0.0
13	metric	COMET-Kiwi-22	2.2	2.7	11.0	23.0	0.3	0.0
14	metric	MetricX-23-QE-L	0.2	-0.3	16.5	25.5	0.3	0.0

Table 2: EN-RU results ($n = 200$ documents), reported as percentages. τ_a is the per-pair-verdict Kendall, which is the coefficient comparable across the judge-metric boundary and is used for system ranking. The best value among judge models is shown in bold. For Tie and Disagree, lower is better.

- Strong commercial LLMs, such as GPT-5.4 and Gemini-3.5-Flash, perform substantially better, exhibiting significantly lower tie rates. In contrast, weaker models, such as Gemini-2.5-Flash, produce a much higher tie rate, suggesting a limited ability to discriminate among high-quality translation candidates.
- Pairwise LLM judges exhibit position bias (Disagree): reversing the order of a translation pair can cause the preferred candidate to switch. However, strong LLMs, e.g., Gemini-3.5-Flash, performs much better than weaker models here, e.g., in Chinese, Gemini-3.5-Flash has a score of 12.6 while Gemini-2.5-Flash has a score of 51.4.
- Reasoning sometimes helps judging, but the gains are marginal compared with direct judging. For example, Gemini-3.5-Flash with reasoning only slightly outperforms its

direct-judge counterpart (23.1 vs. 22.4 in τ_a), while even achieving lower Spearman’s ρ and Best@1 accuracy.

Considering its strong performance and low cost, we selected gemini-3.5-flash (direct) as the judge model for our final submission.

4 Final Ensemble

For the final submission, we run the generation using gemini-3.5-flash and GPT-5.4, selecting a winning candidate from each family using gemini-3.5-flash (direct).

Ensembling 2 model families. For two models (a, b), let us denote $S_{a,a}, S_{b,b}$ as score matrices for intra-model evaluation, and $S_{a,b}, S_{b,a}$ as inter-model evaluation. We do not fully compute $S_{a,b}, S_{b,a}$, but only compute the cross-model scores

only for the selected winners (w_a and w_b):

$$\begin{aligned} \text{score}[w_a] = & \sum_{j \neq w_a} S_{a,a}[w_a, j] - \sum_{j \neq w_a} S_{a,a}[j, w_a] \\ & + \sum_j S_{a,b}[w_a, j] - \sum_j S_{b,a}[j, w_a]. \end{aligned} \quad (3)$$

$$\begin{aligned} \text{score}[w_b] = & \sum_{j \neq w_b} S_{b,b}[w_b, j] - \sum_{j \neq w_b} S_{b,b}[j, w_b] \\ & + \sum_j S_{b,a}[w_b, j] - \sum_j S_{a,b}[j, w_b]. \end{aligned} \quad (4)$$

We run $2N(N-1)$ judge calls within the groups and $4N$ cross-group judge calls for the winners (assuming all groups are of size N), in total we run $2N(N-1) + 4N - 2 = 142$ calls. A full $N = 16$ round robin tournament would require $2N(2N-1) = 240$ judge calls. Although other efficient tournament variants are possible, we use this strategy as a simple and easy to scale hierarchical approximation to the full round-robin tournament.

5 Conclusion

In sum, we implement a combination of sequential candidate sampling and contrastive round-robin voting for reranking. Our contrastive reranking strategy demonstrates much higher correlation with human judgements than more standard quality-aware decoding techniques.

6 Acknowledgments

This work is supported in part by project VI.Veni.212.228 of the research program ‘Veni’, which is financed by the Dutch Research Council (NWO); this research is also funded in part by the Netherlands Organization for Scientific Research (NWO) under project numbers VI.C.192.080 and 2023.017; and is part of ‘Hybrid Intelligence: augmenting human intellect’ (<https://hybrid-intelligence-centre.nl>) with project number 024.004.022 of the research program ‘Gravitation’ which is (partly) financed by the Dutch Research Council (NWO).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Daniel Deutsch, Eleftheria Briakou, Isaac Rayburn 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](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12257–12284, Vienna, Austria. Association for Computational Linguistics.

Patrick Fernandes, António Farinhas, Ricardo Rei, José G. C. de Souza, Perez Ogayo, Graham Neubig, and Andre Martins. 2022. [Quality-aware decoding for neural machine translation](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1396–1412, Seattle, United States. Association for Computational Linguistics.

Günter Hägele and Friedrich Pukelsheim. 2001. “*Lull’s writings on electoral systems*.”. *Studia Lulliana* 41: 3-38.

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 Lakouga, Jessica Lundin, Christof Monz, Kenton Murray, and 10 others. 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*, pages 355–413, Suzhou, 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, and 3 others. 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, 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.
- Alon Lavie, Greg Hanneman, Sweta Agrawal, Diptesh Kanojia, Chi-Kiu Lo, Vilém Zouhar, Frederic Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Deoghare, Archchana Sindhuja, Jiayi Wang, David Ifeoluwa Adelani, Brian Thompson, Tom Kocmi, Markus Freitag, and Daniel Deutsch. 2025. [Findings of the WMT25 shared task on automated translation evaluation systems: Linguistic diversity is challenging and references still help](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 436–483, Suzhou, China. Association for Computational Linguistics.
- Ramon Llull. c. 1274-83. *Artifitium electionis personarum (AEP)*, edited and translated in Günter Hägele and Friedrich Pukelsheim, “Llull’s Writings on Electoral Systems,” *Studia Lulliana* 41 (2001): 7–11; from *Biblioteca Apostolica Vaticana, Cod. Vat. lat. 9332, f. 11r-12v*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. [COMET-22: Unbabel-IST 2022 submission for the metrics shared task](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Ricardo Rei, Nuno M. Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André F. T. Martins. 2023. [Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore. Association for Computational Linguistics.
- Shaomu Tan, Ryosuke Mitani, Ritvik Choudhary, and Toshiyuki Sekiya. 2025. Investigating test-time scaling with reranking for machine translation. *arXiv preprint arXiv:2509.19020*.
- Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. [Large language models are not fair evaluators](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.
- Di Wu, Seth Aycok, and Christof Monz. 2025a. Please translate again: Two simple experiments on whether human-like reasoning helps translation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20435–20451.
- Di Wu, Yan Meng, Maya Konstantinovna Nachesa, Seth Aycok, and Christof Monz. 2025b. [UvA-MT’s participation in the WMT25 general translation shared task](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 688–694, Suzhou, China. Association for Computational Linguistics.
- Di Wu, Sergey Troshin, Christof Monz, Antske Fokkens, and Vlad Niculae. 2026. Ladders in chaos: When, how,(and perhaps why) does test-time scaling improve llm machine translation. *arXiv preprint arXiv:2608.28496*.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.