UvA-MT's Participation in the WMT25 General Translation Shared Task

Di Wu **Yan Meng** **Maya Nachesa** **Seth Aycock** **Christof Monz**
Language Technology Lab
University of Amsterdam
{d.wu, y.meng, m.k.nachesa, s.aycock, c.monz}@uva.nl

Abstract

This paper presents UvA-MT's submission to the WMT 2025 shared task on general machine translation, competing in the unconstrained track across all 16 translation directions. Unusually, this year we use only WMT25's blind **test set** (source sentences only) to generate synthetic data for LLM training, and translations are produced using pure beam search for submission. Overall, our approach can be seen as a special variant of data distillation, motivated by two key considerations: (1) perfect domain alignment, where the training and test domains are distributionally identical; and (2) the strong teacher model, GPT-4o-mini, offers high-quality outputs as both a reliable reference and a fallback in case of mere memorization.

Interestingly, the outputs of the resulting model, trained on Gemma3-12B using Best-of-N (BoN) outputs from GPT-4o-mini, outperform both original BoN outputs from GPT-4o-mini and Gemma3-12B in some high-resource languages across various metrics. We attribute this to a successful model ensemble, where the student model (Gemma3-12B) retains the strengths of the teacher (GPT-4o-mini) while implicitly avoiding its flaws.

1 Introduction

In this paper, we describe the details of our submission to the WMT 2025 shared task on the general machine translation (unconstrained track), which includes 16 translation directions. With recent advances in Large Language Models (LLMs), particularly the emergence of stronger multilingual models, our focus in this paper is on effectively and efficiently adapting a general-purpose LLM for translation-specific tasks with limited training.

Unusually, this year we use only the **test set** to build synthetic data for model training and generate translations again based on the test set using pure beam search for submission, as shown in Figure 1. This is based on several considerations:

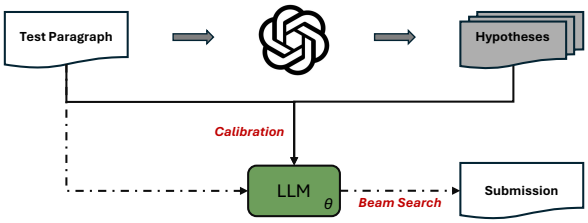


Figure 1: We use GPT-4o-mini to generate 16 hypotheses per sample on the WMT25 test set (at the paragraph level) using nucleus sampling. The resulting hypotheses are then used to train Gemma3-12B with the calibration method of Wu et al. (2025b). Finally, the calibrated Gemma model is used to translate the WMT25 test set for submission using pure beam search.

- **Very small sample sets** with a strong base model can effectively boost translation performance (Wu et al., 2024; Xu et al., 2024a).
- **Perfect domain alignment** since the training and test domains are inherently identical.
- **A strong teacher model** such as GPT-4o-mini[1] offers high-quality outputs as both a reliable reference and a fallback in case of mere memorization.

In the following sections, we test whether the student model (Gemma-3-12B) retains the strengths of GPT-4o-mini's outputs while implicitly avoiding certain flaws—effectively acting as a model ensemble.

More specifically, our strategy consists of two main steps:

Synthetic data building. We feed the WMT25 test set into GPT-4o-mini (Hurst et al., 2024), using the prompts provided with the offical test set[2], to generate 16 hypotheses per sample. The hypotheses are decoded using nucleus sampling with a top-p of 0.98[3] and a temperature of 1.0. Each sample

[1]As reported by Wu et al. (2025a), GPT-4o-mini can serve as a strong translation system.

[2]Official prompts are found here.

[3]We found that slightly lowering the top-p value effectively eliminates the off-targeting issue while preserving diversity.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 241–247
November 8–9, 2025.

is at the paragraph level, where "$\backslash n$" remains in the original data as a separator.

Post-training. We apply the calibration method (Wu et al., 2025b) only to post-train Gemma-3-12B, which has been shown to be more effective than supervised fine-tuning or recent preference optimization methods, like CPO (Xu et al., 2024b). The calibration method aims to improve the correlation between translation likelihood and quality scores as measured by a reference metric model, enhancing the effectiveness of beam search decoding. Following Wu et al. (2025b), we use CometKiwi-XXL to score each one-to-many translation pair in our synthetic dataset.

Finally, the resulting model, trained on synthetic data derived from WMT25 test set, is used to again translate the WMT25 test set. We observe that for some high-resource languages, the resulting model's outputs even surpass the best hypotheses in the synthetic data—demonstrating a successful form of model ensemble.

In our next version, we provide detailed experimental settings and results, including: (1) offline experiments demonstrating the effectiveness of the calibration method; (2) offline experiments evaluating this test-time model ensemble strategy; and (3) our evaluation results for the final submission.

2 Calibration Method

We now briefly describe our post-training method, namely calibration (Wu et al., 2025b). This method addresses the miscalibration problem in machine translation, where translation quality deteriorates as search approximations improve and higher-probability hypotheses are potentially worse translations.

Prior studies have tried to mitigate this miscalibration issue by introducing an additional optimization step during inference time, known as quality-aware decoding (QAD) (Fernandes et al., 2022). These approaches typically involve generating multiple candidate translations through sampling, followed by reranking or voting using reference-free and/or reference-based machine translation metrics, such as Best-of-N (BoN) sampling (Rei et al., 2024; Faria et al., 2024) and Minimum Bayes Risk (MBR) decoding (Kumar and Byrne, 2004; Freitag et al., 2022).

The calibration approach mitigates this issue by optimizing the Pearson correlation between translation likelihood and quality during **training time**. Extensive experiments from Wu et al. (2025b) show several key advantages of this method, including:

1. Substantial translation performance gains with limited training.

2. Clear enhancements for maximum *a posterior* decoding, like beam search.

3. A unified framework for both translation quality optimization and estimation. Notably, we also apply this method to participate in the Quality Estimation task at WMT25[4].

In this shared task, we employ calibration as our only post-training method. For further technical details, please refer to (Wu et al., 2025b).

3 Online Evaluation

We thoroughly evaluate our system's outputs and compare them with those of several high-performance open-source and closed-source LLM-based translation systems, including GPT-4.1, Claude-4, Command-R+, DeepSeek-V3, TowerPlus-9B, TowerPlus-72B, Qwen2.5-7B, Qwen3-235B, and AyaExpanse-32B. We access these systems' results from the WMT25 MT evaluation test set[5], which was released a few weeks before the submission deadline of this paper.

We report results using three metrics: COMETKiwi$_{23}^{\text{DA}}$-XL, COMETKiwi$_{23}^{\text{DA}}$-XXL (Rei et al., 2023), and COMET$_{22}^{\text{DA}}$ (Rei et al., 2022). In addition, we conduct a light human evaluation for the English–Chinese track, comparing our system (UvA-MT) with our base model, Gemma-3-12B.

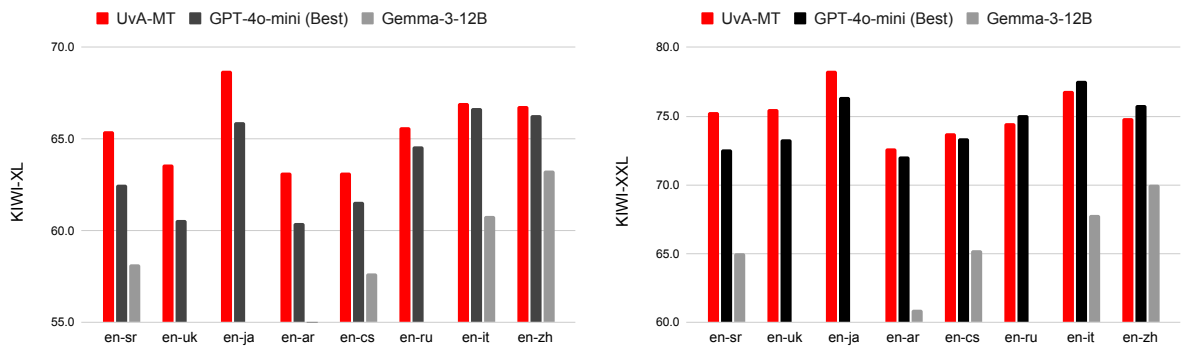
3.1 The Effectiveness of Ensembling

Figure 2 (a) and (b) show our system's results compared to those of 1) our base model, i.e., Gemma-3-12B, and 2) our teacher model, i.e., GPT-4o-mini, measured by CometKiwi-XL and CometKiwi-XXL, respectively.

Note that the reported results for GPT-4o-mini (best) are obtained using Best-of-N sampling. As described in the synthetic data building process in Section 1, we generate 16 hypotheses for each input and select the best one based on CometKiwi-XXL scores for evaluation.

[4]We cannot cite our QE system report at the time of submitting this paper; please refer to this year's findings paper for potential reference.

[5]The translation outputs from various systems are provided to human experts for annotation, which also serves as the evaluation task's test data.



(a) Performance measured by CometKiwi-XL

(b) Performance measured by CometKiwi-XXL

Figure 2: Performance comparison among the student model (Gemma-3-12B), the best-of-N outputs from the teacher model (GPT-4o-mini), and our ensemble model (UvA-MT). It is clear that UvA-MT surpasses all models in these 8 languages when measured with CometKiwi-XL, and outperforms some of them when measured with CometKiwi-XXL. Note that in some cases, such as **en-uk** and **en-ja**, the performance of Gemma-3-12B is either below the x-axis or not supported by the base model, so we did not show them in the figures.

For the other systems—namely Gemma-3-12B and our own (UvA-MT)—only a single hypothesis is generated per input using beam search with a beam size of 5.

As our system leverages both the base and teacher models, a successful ensemble would be expected to outperform each individually, at least in a portion of language directions.

Notably, we can observe in Figure 2 that:

- When evaluated with CometKiwi-XL (Figure 2-a), our system (UvA-MT) outperforms both the base model (Gemma-3-12B) and the teacher model (GPT-4o-mini) with Best-of-N sampling across all 8 language directions.
- When evaluating in CometKiwi-XXL (Figure 2-b), which is the very metric for Best-of-N selection, we can expect GPT-4o-mini's Best-of-N outputs to benefit from this evaluation due to greater potential for metric hacking. However, even under these conditions, our system still outperforms the teacher's results in a few language directions, such as `en-sr` (+2.7), `en-uk` (+2.0), and `en-ja` (+1.9), among others.

We acknowledge that some degree of metric gaming may also exist in our system's results above. However, we contend that a direct comparison between UvA-MT and GPT-4o-mini (Best) remains valid, because UvA-MT's training data is exactly the same as GPT-4o-mini's sampling data, we can therefore assume that UvA-MT could benefit from metric gaming *no more* than GPT-4o-mini's Best-of-N results; thus the relative gain over GPT-4o-mini (Best) should be considered realistic.

3.2 Overall Results

We now present a broader evaluation with detailed results measured by Comet22, CometKiwi-XL, and CometKiwi-XXL for 12 selected systems. For the complete set of results, please refer to the WMT25 findings paper. Note that the scores may differ slightly from those reported in the WMT25 findings paper, as variations in evaluation environments can introduce minor discrepancies (Zouhar et al., 2024).

Table 1 shows the results in CometKiwi-XL, one of this year's offical metrics. We show that UvA-MT achieves the best results in most of the language directions. When evaluated with CometKiwi-XXL (Table 2), the metric used for Best-of-N sampling, GPT-4o-mini (Best) obtains the highest scores in most cases. This is as expected with the previous discussion about metric gaming.

System ID	en-ru	en-it	en-zh	en-ar	en-cs	en-uk	en-ko
GPT-4.1	62.7	65.4	64.8	53.1	62.5	62.2	68.0
Claude-4	61.5	63.9	64.6	54.9	60.0	60.0	68.8
CommandA	–	64.4	64.2	53.1	60.3	60.6	68.6
DeepSeek-V3	62.5	65.0	61.3	57.0	62.1	61.3	67.5
UvA-MT	**65.6**	**67.0**	**66.8**	**63.2**	**63.2**	**63.6**	**70.0**
GPT-4o-mini (Best)	64.6	66.7	66.3	60.4	61.6	60.6	69.3
Gemma-3-12B	–	60.8	63.3	52.5	57.7	–	65.9
TowerPlus-9B	61.2	64.1	63.2	–	59.7	59.9	67.1
TowerPlus-72B	61.9	64.5	–	–	–	–	67.8
Qwen2.5-7B	–	58.4	62.1	–	–	–	–
Qwen3-235B	62.1	64.8	65.6	–	–	–	67.2
AyaExpanse-32B	–	63.9	–	58.2	59.9	–	66.7

System ID	cs-uk	en-ja	cs-de	en-et	ja-zh	en-is	en-sr
GPT-4.1	57.3	67.1	56.4	69.2	54.4	**64.7**	65.3
Claude-4	57.2	67.4	56.1	67.0	54.0	62.4	62.6
CommandA	56.4	67.1	56.5	–	53.4	–	–
DeepSeek-V3	55.8	66.2	56.3	–	52.6	54.7	60.1
UvA-MT	**57.7**	**68.7**	56.8	**69.3**	55.7	62.3	**65.4**
GPT-4o-mini (Best)	56.9	65.9	**58.4**	68.7	**56.2**	62.8	62.5
Gemma-3-12B	54.2	–	54.5	59.6	–	51.6	58.2
TowerPlus-9B	55.2	66.2	55.0	–	53.2	63.5	36.9
TowerPlus-72B	–	–	55.7	–	53.3	61.7	–
Qwen2.5-7B	–	–	–	–	52.2	–	–
Qwen3-235B	–	66.3	55.0	–	54.1	–	59.0
AyaExpanse-32B	55.4	–	55.0	–	–	–	–

Table 1: KIWI-XL scores across languages and systems. We highlight UvA-MT and GPT-4o-mini (Best), where the former uses the latter's output as training data. Bold indicates the highest score per column. We discard the results in two extremely low-resource directions, i.e., English to Bhojpuri and Maasai, as they are not supported by the base model and therefore lack meaningful comparability.

We note that the primary focus of this paper is to explore whether an ensemble strategy can outperform the teacher's output—a trend that is clearly observed in most cases in Table 1 and in a few cases in Table 2.

A more convincing result is obtained with Comet22, the reference-based metric, where we additionally consider the translation references provided by WMT25, thus maximizing the metric difference between training and evaluation. In Table-3, we can see that UvA-MT achieves best results in **en-ru** and **en-it** among all systems.

4 Discussion and Conclusion

Beyond Metric Hacking. We acknowledge that some degree of metric gaming is present in the results above, although its extent is difficult to quantify. Our focus in this competition, however, is to demonstrate gains that go beyond mere metric hacking.

In the extreme case where UvA-MT simply memorized the best outputs from GPT-4o-mini (maximizing hacking), the latter's score would represent the upper bound of the former. Therefore, a direct comparison between UvA-MT and GPT-4o-mini (Best) remains realistic, and any gain over GPT-4o-mini (Best) would reflect genuine enhancements. Interestingly, we observe them in most of the language directions.

We attribute these gains to a form of successful model ensemble, in which the student LLMs integrate the strengths of the teacher model's outputs while discarding some of their shortcomings. Regarding the role of the post-training method applied here, including whether it is a critical component for this ensemble, we leave for future investigation.

Practical Significance. Our setting is not well-suited for real-time translation systems, as training a student model is required for each group of new inputs. Nevertheless, our findings point to a promising direction for ensembling the strengths of two models when the target domain is established in advance. This is particularly relevant in practical scenarios such as customized translation, where latency is secondary and effectiveness is the foremost priority.

System ID	en-ru	en-it	en-zh	en-ar	en-cs	en-uk	en-ko
GPT-4.1	70.4	74.8	72.3	62.8	72.9	74.5	78.7
Claude-4	69.7	72.8	72.3	63.6	69.0	71.0	78.9
CommandA	–	72.9	71.4	62.0	69.3	70.8	78.4
DeepSeek-V3	69.7	74.3	67.6	65.6	72.7	73.4	77.7
UvA-MT	74.5	76.9	74.8	**72.6**	**73.8**	**75.5**	79.9
GPT-4o-mini (Best)	**75.1**	**77.6**	**75.8**	72.1	73.4	73.3	**80.8**
Gemma-3-12B	–	67.8	70.0	60.9	65.2	–	75.3
TowerPlus-9B	68.8	71.9	69.9	–	68.2	70.3	76.3
TowerPlus-72B	70.3	73.3	–	–	–	–	77.4
Qwen2.5-7B	–	62.9	68.7	–	–	–	–
Qwen3-235B	70.0	73.9	73.7	–	–	–	77.4
AyaExpanse-32B	–	71.8	–	66.1	69.0	–	76.1

System ID	cs-uk	en-ja	cs-de	en-et	ja-zh	en-is	en-sr
GPT-4.1	62.6	76.1	66.4	81.0	65.1	**74.5**	75.8
Claude-4	63.5	77.2	66.5	77.9	65.2	70.3	72.5
CommandA	61.6	76.1	66.5	–	64.6	–	–
DeepSeek-V3	60.7	74.9	65.5	–	62.2	60.3	69.3
UvA-MT	63.6	**78.3**	67.2	80.2	67.1	68.5	**75.3**
GPT-4o-mini (Best)	**66.8**	76.4	**72.0**	**81.4**	**69.8**	72.4	72.6
Gemma-3-12B	59.7	–	63.3	68.0	–	51.9	65.1
TowerPlus-9B	60.5	74.8	64.5	–	63.8	72.1	36.8
TowerPlus-72B	–	–	65.1	–	64.2	69.3	–
Qwen2.5-7B	–	–	–	–	62.0	–	–
Qwen3-235B	–	75.5	64.5	–	64.5	–	66.9
AyaExpanse-32B	61.0	–	64.7	–	–	–	–

Table 2: KIWI-XXL scores across languages and systems. We highlight UvA-MT and GPT-4o-mini (Best), where the former uses the latter's output as training data. Bold indicates the highest score per column.

System ID	en-ru	en-it	en-zh	en-ar	en-cs	en-uk	en-ko
GPT-4.1	82.4	45.7	82.9	**79.1**	**85.7**	**85.5**	87.4
Claude-4	80.6	44.2	82.1	76.4	82.3	82.3	85.9
CommandA	–	44.7	80.9	77.4	82.9	82.9	85.4
DeepSeek-V3	82.1	46.0	80.9	79.0	85.4	84.9	86.6
UvA-MT	**83.4**	**46.5**	82.7	78.9	85.2	84.5	86.3
GPT-4o-mini (Best)	–	–	–	–	–	–	–
Gemma-3-12B	–	44.6	80.7	77.0	80.4	–	84.7
TowerPlus-9B	80.8	44.7	80.6	–	82.5	83.1	83.8
TowerPlus-72B	80.9	44.6	–	–	–	–	84.3
Qwen2.5-7B	–	43.0	80.9	–	–	–	–
Qwen3-235B	82.1	46.1	**83.4**	–	–	–	**87.3**
AyaExpanse-32B	–	45.1	–	75.4	82.7	–	84.6

System ID	cs-uk	en-ja	cs-de	en-et	ja-zh	en-is	en-sr
GPT-4.1	**88.1**	**88.1**	**83.9**	**86.5**	**85.0**	**81.8**	**83.8**
Claude-4	87.0	86.4	82.3	83.2	83.7	78.3	79.6
CommandA	87.2	86.5	83.1	–	83.4	–	–
DeepSeek-V3	87.6	87.4	83.4	–	83.6	73.7	80.2
UvA-MT	86.9	86.9	82.5	85.0	82.8	78.5	67.7
GPT-4o-mini (Best)	–	–	–	–	–	–	–
Gemma-3-12B	85.1	–	80.6	79.3	–	70.4	74.1
TowerPlus-9B	86.7	85.7	81.2	–	82.6	81.0	49.0
TowerPlus-72B	–	–	81.3	–	82.5	79.5	–
Qwen2.5-7B	–	–	–	–	81.1	–	–
Qwen3-235B	–	87.9	82.3	–	83.8	–	78.3
AyaExpanse-32B	86.5	–	83.0	–	–	–	–

Table 3: Comet22 scores across languages and systems. We highlight UvA-MT and GPT-4o-mini (Best), where the former uses the latter's output as training data. Bold indicates the highest score per column.

References

Gonçalo Faria, Sweta Agrawal, António Farinhas, Ricardo Rei, José de Souza, and André Martins. 2024. Quest: Quality-aware metropolis-hastings sampling for machine translation. *Advances in Neural Information Processing Systems*, 37:89042–89068.

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.

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. 2022. High quality rather than high model probability: Minimum Bayes risk decoding with neural metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Shankar Kumar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 169–176, Boston, Massachusetts, USA. Association for Computational Linguistics.

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, JosÃ© Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André 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.

Ricardo Rei, Jose Pombal, Nuno M. Guerreiro, João Alves, Pedro Henrique Martins, Patrick Fernandes, Helena Wu, Tania Vaz, Duarte Alves, Amin Farajian, Sweta Agrawal, Antonio Farinhas, José G. C. De Souza, and André Martins. 2024. Tower v2: Unbabel-IST 2024 submission for the general MT shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 185–204, Miami, Florida, USA. Association for Computational Linguistics.

Di Wu, Seth Aycock, and Christof Monz. 2025a. Please translate again: Two simple experiments on whether human-like reasoning helps translation. *arXiv preprint arXiv:2506.04521*.

Di Wu, Yibin Lei, and Christof Monz. 2025b. Calibrating translation decoding with quality estimation on llms. *arXiv preprint arXiv:2504.19044*.

Di Wu, Shaomu Tan, Yan Meng, David Stap, and Christof Monz. 2024. How far can 100 samples go? unlocking zero-shot translation with tiny multi-parallel data. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15092–15108, Bangkok, Thailand. Association for Computational Linguistics.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024a. A paradigm shift in machine translation: Boosting translation performance of large language models. In *The Twelfth 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 *ICML*.

Vilém Zouhar, Pinzhen Chen, Tsz Kin Lam, Nikita Moghe, and Barry Haddow. 2024. Pitfalls and outlooks in using COMET. In *Proceedings of the Ninth*

Conference on Machine Translation, pages 1272–1288, Miami, Florida, USA. Association for Computational Linguistics.