Marco Large Translation Model at WMT2025: Transforming Translation Capability in LLMs via Quality-Aware Training and Decoding

Hao Wang, Linlong Xu, Heng Liu, Yangyang Liu, Xiaohu Zhao
Bo Zeng, Longyue Wang, Weihua Luo, Kaifu Zhang

Alibaba International Digital Commerce

https://github.com/AIDC-AI/Marco-MT
https://huggingface.co/AIDC-AI/Marco-MT-Algharb

Abstract

This paper presents the **Marco-MT-Algharb** system, our submission to the WMT2025 General Machine Translation Shared Task from **Alibaba International Digital Commerce** (AIDC). Built on a large language model (LLM) foundation, the system's strong performance stems from novel quality-aware training and decoding techniques: (1) a two-step supervised fine-tuning (SFT) process incorporating data distillation, (2) a two-step reinforcement learning (RL) framework for preference alignment, and (3) a hybrid decoding strategy that integrates word alignment with Minimum Bayes Risk (MBR) re-ranking to improve faithfulness. These approaches jointly ensure high accuracy and robustness across diverse languages and domains. **In the official human evaluation, our system secured six first-place finishes, four second, and two third-place results** in the constrained category across the 13 directions we participated in. Notably, for the **English-Chinese**, our results surpassed all open/closed-source systems.

Lang. Pair	Human Evaluation
en→zh	Rank 1
en→ja	Rank 1
en→uk	Rank 1
ja→zh	Rank 1
en→bho	Rank 1
en→et	Rank 1
en→cs	Rank 2
en→ko	Rank 2
en→ru	Rank 2
cs→de	Rank 2
en→arz	Rank 3
cs→uk	Rank 3

Table 1: Human evaluation rankings of Marco-MT-Algharb at WMT2025.

1 Introduction

The Conference on Machine Translation (WMT) continues to be the primary arena for benchmarking the advancements in machine translation technology (Kocmi et al., 2024; Freitag et al., 2023). For years, the field was dominated by the Transformer architecture (Vaswani et al., 2017), which set a high standard for translation quality through its powerful attention mechanism. However, the recent advent of Large Language Models (LLMs) has sparked a paradigm shift. These models, pre-trained on vast amounts of text data, have demonstrated remarkable capabilities in understanding context, generating fluent text, and leveraging world knowledge, making them exceptionally promising candidates for complex multilingual translation tasks (Achiam et al., 2023; Ouyang et al., 2022; Ming et al., 2024; Alves et al., 2024). However, adapting these powerful, general-purpose LLMs for high-fidelity, specialized translation presents a significant challenge (Jiao et al., 2023; Hendy et al., 2023). To bridge this gap, we propose quality-aware training and decoding techniques designed to systematically enhance both the fluency and faithfulness of LLM-based translation. To this end, we present the Marco-MT-Algharb system.

Marco-MT-Algharb is our submission to the WMT 2025 General Machine Translation Shared Task. Our participation covers 13 diverse language pairs.[1] Built upon the Qwen3-14B foundation

[1]The 13 language pairs are: English to Chinese (en→zh), English to Japanese (en→ja), English to Korean (en→ko), English to Egyptian Arabic (en→arz), English to Bhojpuri (en→bho), English to Czech (en→cs), English to Estonian (en→et), English to Russian (en→ru), English to Ukrainian (en→uk), English to Serbian (Latin script) (en→sr_Latn), Czech to German (cs→de), Czech to Ukrainian (cs→uk), and Japanese to Chinese (ja→zh).

Proceedings of the Tenth Workshop on Machine Translation (WMT), pages 587–593
November 8–9, 2025.

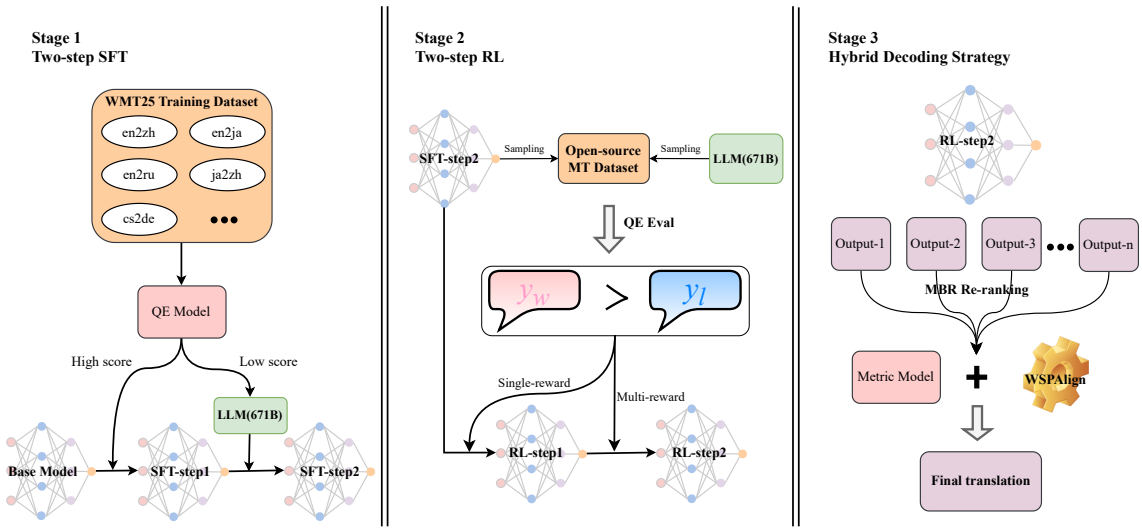

Figure 1: Overall pipeline of the Marco-MT-Algharb system. **SFT Stage:** Fine-tuning on QE-filtered and teacher-distilled data. **RL Stage:** Preference alignment via CPO and dynamic multi-reward optimization. **Decoding Stage:** Hybrid MBR re-ranking with a WSPAlign faithfulness score.

model (Yang et al., 2025), we propose quality-aware training and decoding techniques to enhance translation quality through three key phases: (1) a two-step Supervised Fine-Tuning (SFT) process with data distillation to expand data coverage; (2) a two-step Reinforcement Learning (RL) framework to align the model with quality estimation metrics; and (3) a hybrid decoding strategy that combines quality scores with word alignment to improve faithfulness.

Our work makes several key contributions to the development of state-of-the-art, LLM-based translation systems:

- **A progressive, two-step supervised fine-tuning (SFT) strategy.** We begin by training on high-quality, rigorously cleaned parallel data. Subsequently, we employ data distillation, using a powerful teacher model to regenerate translations for the data filtered out during the cleaning process. This allows our model to learn from a broader data distribution without being compromised by noise.
- **A two-step reinforcement learning (RL) framework for preference alignment.** We first utilize Contrastive Preference Optimization (CPO) (Xu et al., 2024) for initial alignment. We then introduce a novel dynamic multi-reward preference optimization method, which leverages a combination of quality metrics to achieve a more holistic improvement in translation adequacy and fluency.
- **A novel hybrid decoding strategy to mitigate omission errors.** We observed that models optimized heavily on neural quality estimation (QE) metrics like COMET (Rei et al., 2020) can sometimes produce fluent but incomplete translations. To address this, we developed a hybrid decoding algorithm that integrates a word-alignment-based penalty into the Minimum Bayes Risk (MBR) re-ranking framework, ensuring both semantic fidelity and lexical completeness.

In the official human evaluation (Kocmi et al., 2025), our system achieved top-3 rankings in 10 out of 13 participated directions within the constrained category, including five first-place finishes (see Table 1). Notably, our English-Chinese system surpassed all other submissions, including proprietary systems.

The remainder of this paper is structured as follows: Section 2 provides a detailed overview of our system's architecture and training methodology. Section 3 presents our experimental setup and results. Finally, Section 4 concludes the paper.

2 System Overview

Our translation system, Marco-MT-Algharb, is an end-to-end pipeline built upon a powerful foundation model. The overall workflow consists of four key stages: (1) selection of a base model architecture; (2) a two-step SFT process to impart translation capabilities; (3) a two-step RL process to align outputs with automated translation quality metrics; and (4) a hybrid decoding strategy for

robust and accurate inference. A schematic of our system is shown in Figure 1.

2.1 Model Architecture

We selected Qwen3-14B-base[2] as the foundation for our system. Qwen3 is a series of advanced, multilingual large language models known for their strong performance across a wide range of natural language understanding and generation tasks. The 14-billion parameter variant provides a powerful balance between model capacity and computational feasibility for fine-tuning. Its extensive pre-training on a diverse corpus of multilingual data makes it an excellent starting point for developing a high-quality, multilingual translation system, as it already possesses a rich cross-lingual representation space. For our tasks, we utilized the base model and tailored it specifically for translation through the subsequent training stages.

2.2 Two-step Supervised Fine-tuning

The goal of our SFT process is to effectively adapt the general capabilities of Qwen3-14B to the specific domain of machine translation. We designed a two-step approach to maximize data utilization and model performance.

Step 1: SFT on High-Quality Parallel Data. In the first step, we focused on building a robust translation baseline using high-quality data. We collected all parallel data provided by the WMT 2025 organizers for the 13 language directions we participated in. This raw data underwent an intensive cleaning pipeline, which included:

- Normalization: Standardizing punctuation, spacing, and casing.
- Filtering: Removing sentence pairs based on length ratio mismatches and identifying sentence pairs with a high proportion of non-alphabetic characters.
- Language Identification: Ensuring that the source and target sentences correctly match their designated language labels.
- Quality Estimation Filtering: Employing a pre-trained QE model (CometKiwi-XXL[3]) to score sentence pairs and discarding those with predicted low translation quality.

After cleaning, the resulting high-quality dataset was used to perform a single, comprehensive multilingual SFT run. For 12 of our high-resource language directions, we compiled a substantial dataset of 10 million parallel sentences each. Due to data scarcity for the English-to-Bhojpuri (en→bho) direction, its data volume was significantly smaller. In total, this initial SFT step utilized a massive training corpus of approximately 120 million sentence pairs, training the model on all 13 language pairs simultaneously. This large-scale multilingual training encourages the model to develop robust shared representations and effectively leverage cross-lingual transfer learning.

Step 2: SFT with Distilled Noisy Data. A significant amount of data is typically discarded during aggressive cleaning. While noisy, this data often contains valuable lexical and syntactic diversity. To leverage this, we designed a second SFT step based on data distillation. We took the parallel data that was filtered out in Step 1 and used a powerful teacher model, DeepSeek-V3[4] (Liu et al., 2024), to regenerate the target-side translations. For each of the 13 language directions, we distilled approximately 800,000 sentence pairs. This process effectively "cleans" the noisy target text while preserving the original source text's diversity. The resulting distilled dataset was then used for a second round of SFT. This allowed our model to learn from a much broader set of source inputs, guided by the high-quality outputs of the teacher model, thereby enhancing its robustness and domain coverage.

2.3 Preference Alignment via Two-step Reinforcement Learning

Following SFT, we employed a two-step RL to further refine the model's output. The goal of this stage is to directly align the model's generations with scores from automated translation quality estimation metrics, which serve as a proxy for human judgment. This approach moves beyond the token-level supervision of SFT to optimize the holistic quality of the entire translated sentence based on established evaluation standards.

Step 1: Contrastive Preference Optimization with Diverse Candidate Translations. We began with CPO to align our model with high-quality translation preferences. The foundation for our RL training is a curated dataset of source sentences, created by randomly sampling 15,000 entries for each source language from high-quality, open-source

[2]https://huggingface.co/Qwen/Qwen3-14B-Base

[3]https://huggingface.co/Unbabel/wmt23-cometkiwi-da-xxl

[4]https://huggingface.co/deepseek-ai/DeepSeek-V3

corpora, including Flores-200 (nll, 2024) and historical WMT test sets (WMT08-23). We reuse this same dataset for both of our RL steps, not only for methodological consistency, but more importantly, due to the limited availability of high-quality data that closely mirrors the test domain.

To construct preference pairs for these source sentences, we adopted a teacher-augmented strategy. For each source sentence, we generated a pool of candidate translations populated from two distinct sources: (1) multiple sampled outputs from our own SFT-trained model, and (2) a high-quality translation from a powerful, external teacher model, DeepSeek-V3.

We then used a reference-free QE model CometKiwi-XXL to score every candidate in this combined pool. The preference pair was formed as follows:

- The **"chosen"** translation was the candidate with the highest evaluation score. In many cases, this was the output from the DeepSeek-V3 model, providing a strong, high-quality target.
- The **"rejected"** translation was a candidate from the same pool with a significantly lower evaluation score, often one of the less successful samples from our own model.

This teacher-augmented approach is highly effective as it provides a robust learning signal, allowing our model to learn from responses that are often superior to its own initial capabilities. This CPO step provided a stable initial alignment towards the quality standards set by a strong translation model.

Step 2: Dynamic Multi-Reward Optimization with Self-Distillation. To achieve more nuanced control and move beyond reliance on a single, static metric, we introduce a novel training framework in our second RL step. This framework combines a dynamic, hybrid reward signal with a knowledge distillation objective, encouraging the model to internalize the principles of translation quality.

First, our reward function is not static but a dynamic composite of two sources: an external QE metric (CometKiwi-XXL) and the model's own internal reward signal. The total reward R_{total} for a generated translation y from source x at training step t is defined as:

$$R_{\text{total}}(x, y, t) = (1 - w_{\text{self}}(t)) \cdot R_{\text{QE}}(x, y) + w_{\text{self}}(t) \cdot R_{\text{self}}(x, y) \quad (1)$$

where R_{QE} is the score from the QE model, and $R_{\text{self}}(x, y)$ is the model's own sequence log-probability ($\log P_\theta(y|x)$), serving as a measure of its confidence. The weight of the self-reward, $w_{\text{self}}(t)$, is annealed to increase gradually with the training step t. This curriculum strategy allows the model to initially anchor its learning on the reliable external metric and progressively trust its own refined judgment as it improves.

Second, to accelerate the refinement of the model's internal judgment, we introduce a Kullback-Leibler (KL) divergence loss term. This term explicitly distills the relational quality knowledge from the QE model into the model's probability space. For a pair of translations (y_w, y_l) where QE scores y_w higher than y_l, we define a target preference distribution based on their score difference. The KL loss then minimizes the divergence between the model's predicted preference probability and this QE-derived target distribution:

$$\mathcal{L}_{\text{KL}} = D_{\text{KL}}\left(\sigma(\tau \cdot \Delta\text{QE}) \parallel \sigma(\Delta \log P_\theta)\right) \quad (2)$$

where ΔQE is the difference in QE scores for the pair (y_w, y_l), $\Delta \log P_\theta$ is the difference in the model's log-probabilities for the same pair, σ is the sigmoid function, and τ is a temperature parameter controlling the sharpness of the target distribution.

The final objective combines the preference optimization loss with $\mathcal{L}_{\text{KL}}$. This synergistic approach allows the model to not only generate better translations based on the hybrid reward but also to simultaneously internalize the very principles of translation quality evaluation. This makes the self-reward signal more reliable over time and leads to significant, stable performance gains.

2.4 Hybrid Decoding Strategy

A notable pitfall of optimizing Large Language Models towards neural metrics is their tendency to produce translations that are highly fluent yet lexically or semantically incomplete (Freitag et al., 2022; Moghe et al., 2022). To combat this, we developed a hybrid decoding strategy that fuses the MBR re-ranking algorithm with a reward for lexical faithfulness.

Our approach is built upon Minimum Bayes Risk (MBR) decoding (Freitag et al., 2021). In standard MBR, we first generate a set of N candidate translations $\{y_1, y_2, \ldots, y_N\}$ for a given source text x. The optimal translation y^* is the one that has the highest expected utility score against all other can-

Model	AVG	en→zh	en→arz	en→bho	en→cs	en→et	en→ja
Proprietary Models							
GPT-4o	74.20	75.70	74.85	30.19	85.44	76.63	78.35
Claude 3.7 Sonnet	74.24	76.79	76.08	29.12	85.10	76.71	79.48
Ablation Baselines (Marco-MT-Algharb)							
Two-step SFT	76.47	77.67	77.86	32.41	87.82	79.62	81.55
++Two-step RL	77.96	80.59	79.43	34.21	88.90	80.45	82.88
++Hybrid Decode	**79.33**	**82.39**	**80.13**	**38.61**	**90.50**	**83.39**	**83.24**

Model	en→ko	en→ru	en→uk	en→sr	cs→uk	cs→de	ja→zh
Proprietary Models							
GPT-4o	81.14	79.07	76.28	77.88	81.08	82.01	65.94
Claude 3.7 Sonnet	82.32	79.57	76.54	77.71	81.38	80.67	63.69
Ablation Baselines (Marco-MT-Algharb)							
Two-step SFT	84.52	81.35	78.49	80.58	82.47	82.81	66.99
++Two-step RL	84.50	82.87	80.02	82.02	83.37	83.54	68.27
++Hybrid Decode	**84.60**	**84.46**	**82.70**	**83.66**	**83.89**	**84.29**	**69.44**

Table 2: Main results on the WMT25 General test set, evaluated using XCOMET-XXL scores. We report the average (AVG) over all 13 language pairs. The best score in each column is in **bold**. For brevity, **en→sr** refers to the English-to-Serbian (Latin) direction.

didates:

$$y^* = \underset{y_i}{\text{argmax}} \sum_{j=1}^{N} U(y_i, y_j) \tag{3}$$

where the utility function $U(y_i, y_j)$ is realized using the COMET-22 metric model[5].

Our innovation is to incorporate an alignment-based score into this framework to explicitly reward source faithfulness. The final score for each candidate is a hybrid of its peer-based MBR utility and its source-based alignment score. We define the standard MBR score for a candidate y_i as:

$$S_{\text{MBR}}(y_i) = \sum_{j=1}^{N} U(y_i, y_j) \tag{4}$$

The alignment score, $S_{\text{align}}(x, y_i)$, is computed using our tool, `WSPAlign` (Wu et al., 2023), which measures the lexical faithfulness between the source x and the candidate y_i. A higher score indicates better faithfulness. The final hybrid score is then calculated as:

$$S_{\text{hybrid}}(y_i) = S_{\text{MBR}}(y_i) + \lambda \cdot S_{\text{align}}(x, y_i) \tag{5}$$

where λ is a hyperparameter that balances the MBR and alignment terms. During inference, we generate N candidates, compute S_{hybrid} for each, and select the one with the highest score as the final translation. This approach significantly reduces the frequency of omission errors in our final submissions.

3 Experiments

3.1 Experimental Setup

Implementation Details. For the Supervised Fine-Tuning (SFT) stage, we perform full-parameter fine-tuning. In contrast, for the Reinforcement Learning (RL) stage, we employ a parameter-efficient LoRA strategy (Hu et al., 2022), configuring the adapters with a rank of 64 and an alpha of 128. For the optimization process, we use the Adam optimizer (Kingma and Ba, 2014) with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. We adopt a carefully designed learning rate schedule that decreases with each stage of our pipeline: the learning rate was set to 2×10^{-5} for the first SFT step, 1×10^{-5} for the second SFT step, 2×10^{-6} for the first RL step (CPO), and 1×10^{-6} for the final RL step. A global batch size of 64 was maintained throughout all training phases. For our dynamic multi-reward

[5]We use the `Unbabel/wmt22-comet-da` implementation from Hugging Face.

optimization step, key hyperparameters include a KL distillation temperature τ of 0.3, and a self-reward weight w_{self} that is linearly annealed from an initial value of 0.05 to 0.8. In the subsequent Hybrid Decoding stage, the balancing weight λ is set to 0.5.

Evaluation Setup. We evaluate our final model on the official test sets for the WMT25 General Machine Translation Shared Task. Our participation covers 13 language pairs: en→{zh, arz, bho, cs, et, ja, ko, ru, uk, sr_Latn}, cs→{uk, de}, and ja→zh. Our decoding procedure implements the proposed hybrid re-ranking strategy. For each source sentence, we generate 100 candidate translations via stochastic sampling using the `vllm` library (Kwon et al., 2023) for efficient inference, which are then scored and re-ranked using our hybrid scoring function (Equation 5) to select the final output. The quality of this final translation is measured using the state-of-the-art reference-free metric XCOMET-XXL[6] (Guerreiro et al., 2024).

3.2 Main Results

Baselines. To demonstrate the effectiveness of our multi-stage pipeline, we compare our final system against several key baselines. We conduct an ablation study with two internal models: (1) Two-step SFT, our model after only the two SFT step, to measure the combined impact of our RL and hybrid decoding steps; and (2) Two-step RL, the model after full SFT and RL training, to isolate the performance gain from the hybrid decoding strategy. Furthermore, we benchmark our system against leading proprietary models, including GPT-4o and Claude 3.7 Sonnet.

Table 2 presents the main findings of our evaluation. The results clearly demonstrate the superiority of our final Marco-MT-Algharb system. On average, our final system achieves an XCOMET-XXL score of 79.33, significantly outperforming all other models in the comparison.

The effectiveness of our multi-stage pipeline is validated by the ablation study. Our full RL framework (Two-step RL) improves upon the SFT-only baseline by a substantial margin of 1.5 points on average. The final addition of our hybrid decoding strategy (Hybrid Decode) provides a further 1.3-point gain, highlighting the crucial and cumulative contribution of each component to the final performance.

[6]https://huggingface.co/Unbabel/XCOMET-XXL

Most notably, Marco-MT-Algharb not only surpasses the strong proprietary baselines of GPT-4o and Claude 3.7 Sonnet by a large margin (over 5.1 points on average), but it also achieves the highest score across every individual language pair. This consistently strong performance highlights the effectiveness of our specialized training and decoding approach. These results suggest that a carefully refined, open-source model can produce translations of exceptional quality, capable of outperforming even leading general-purpose proprietary systems on this benchmark.

4 Conclusion

This paper presented Marco-MT-Algharb, our system for the WMT25 General Machine Translation Shared Task. We introduced a novel quality-aware framework that enhances LLM-based translation through a synergistic combination of two-step supervised fine-tuning with data distillation, dynamic multi-reward reinforcement learning, and a hybrid alignment-aware decoding strategy. Our methodology was validated on the WMT25 test set, where Marco-MT-Algharb substantially outperformed strong baselines and leading proprietary models. These results were corroborated by the official human evaluation, which placed our system in the top three across 12 of our 13 language pairs, including six first-place victories. Notably, in the highly competitive English-Chinese direction, our system ranked first among all open and closed-source submissions.

References

2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Duarte M Alves, José Pombal, Nuno M Guerreiro, Pedro H Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, and 1 others. 2024. Tower: An open multilingual large language model for translation-related tasks. *arXiv preprint arXiv:2402.17733*.

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. 2021. Minimum bayes risk decoding with neural metrics of translation quality. *arXiv preprint arXiv:2111.09388*.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, and 1 others. 2023. Results of wmt23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André FT Martins. 2022. Results of wmt22 metrics shared task: Stop using bleu–neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68.

Nuno M Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André FT Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*, 1(10).

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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, 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*, 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, and 1 others. 2024. 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.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pages 611–626.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Lingfeng Ming, Bo Zeng, Chenyang Lyu, Tianqi Shi, Yu Zhao, Xue Yang, Yefeng Liu, Yiyu Wang, Linlong Xu, Yangyang Liu, and 1 others. 2024. Marco-llm: Bridging languages via massive multilingual training for cross-lingual enhancement. *arXiv preprint arXiv:2412.04003*.

Nikita Moghe, Tom Sherborne, Mark Steedman, and Alexandra Birch. 2022. Extrinsic evaluation of machine translation metrics. *arXiv preprint arXiv:2212.10297*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *arXiv preprint arXiv:2009.09025*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Qiyu Wu, Masaaki Nagata, and Yoshimasa Tsuruoka. 2023. Wspalign: Word alignment pre-training via large-scale weakly supervised span prediction. *arXiv preprint arXiv:2306.05644*.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *arXiv preprint arXiv:2401.08417*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.