

# HY-MT1.5 Technical Report

Tencent Hunyuan Team

## Abstract

In this report, we introduce our latest translation models, **HY-MT1.5-1.8B** and **HY-MT1.5-7B**, a new family of machine translation models developed through a holistic training framework tailored for high-performance translation. Our methodology orchestrates a multi-stage pipeline that integrates general and MT-oriented pre-training, supervised fine-tuning, on-policy distillation, and reinforcement learning. **HY-MT1.5-1.8B**, the 1.8B-parameter model demonstrates remarkable parameter efficiency, comprehensively outperforming significantly larger open-source baselines (e.g., Tower-Plus-72B, Qwen3-32B) and mainstream commercial APIs (e.g., Microsoft Translator, Doubao Translator) in standard Chinese-foreign and English-foreign tasks. It achieves approximately 90% of the performance of ultra-large proprietary models such as Gemini-3.0-Pro, while marginally trailing Gemini-3.0-Pro on WMT25 and Mandarin-minority language benchmarks, it maintains a substantial lead over other competing models. Furthermore, **HY-MT1.5-7B** establishes a new state-of-the-art for its size class, achieving 95% of Gemini-3.0-Pro's performance on Flores-200 and surpassing it on the challenging WMT25 and Mandarin-minority language test sets. Beyond standard translation, the HY-MT1.5 series supports advanced constraints, including terminology intervention, context-aware translation, and format preservation. Extensive empirical evaluations confirm that both models offer highly competitive, robust solutions for general and specialized translation tasks within their respective parameter scales.

**HY-MT1.5-1.8B:** <https://huggingface.co/tencent/HY-MT1.5-1.8B>

**HY-MT1.5-7B:** <https://huggingface.co/tencent/HY-MT1.5-7B>

**Code Repository:** <https://github.com/Tencent-Hunyuan/HY-MT>

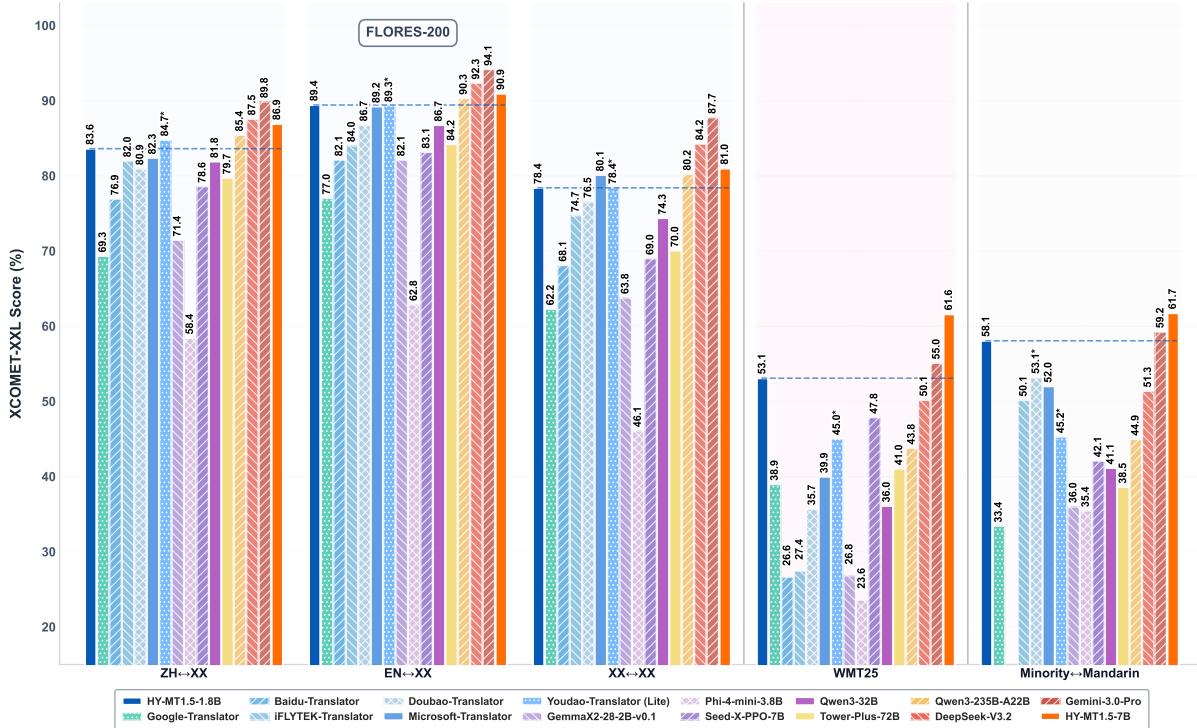
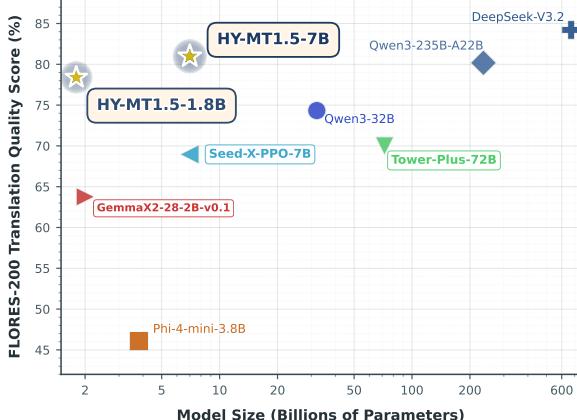
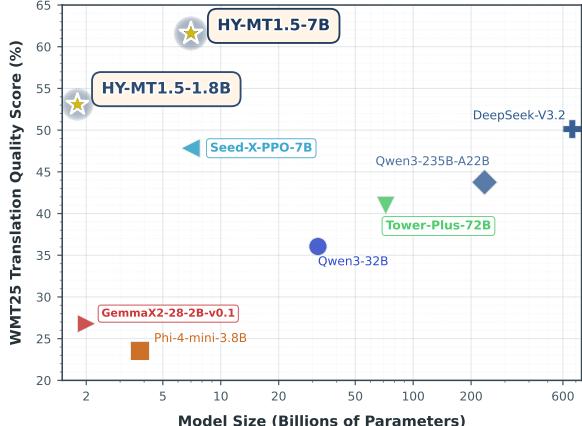


Figure 1: Benchmark performance of HY-MT1.5 models and state-of-the-art baselines.



(a) Model size versus Flores-200 ( $XX \leftrightarrow XX$ ) translation quality for different-scale open-source models.



(b) Model size versus WMT25 translation quality for different-scale open-source models.

Figure 2: Comparison of model size versus translation quality across Flores-200 and WMT25 datasets for open-source models.

## 1 Introduction

Machine translation (MT) has long been a high-demand practical goal and a prominent research challenge pursued by the computational linguistics community over the past few decades (Brown et al., 1990; 1993; Papineni et al., 2002; Sutskever et al., 2014; Bahdanau et al., 2015; Wu et al., 2016; Vaswani et al., 2017). The rapid advancement of large language models (LLMs) has revolutionized the learning paradigm of machine translation, shifting from traditional rule-based and statistical approaches to large-scale neural learning methodologies, and continuously pushing the boundaries of translation quality to unprecedented levels (Zhu et al., 2024; Kocmi et al., 2024; Pang et al., 2025). State-of-the-art closed-source models, such as Gemini-3.0-Pro (DeepMind, 2025), have demonstrated capabilities that approach or surpass those of expert human translators in specific language pairs.

Nevertheless, significant challenges persist in machine translation (Kocmi et al., 2025). First and foremost, the balance between translation quality and efficiency remains a critical unaddressed issue. State-of-the-art large-scale closed-source models often deliver high translation quality but incur prohibitive deployment costs and low inference efficiency due to their enormous parameter sizes, making them inaccessible for widespread practical applications (e.g., edge device deployment and high-throughput translation scenarios). Meanwhile, existing lightweight open-source models typically sacrifice translation quality to achieve efficiency, failing to match the performance of large closed-source models, thereby exacerbating the disparity between practical availability and the demands for high-quality translation. Second, current translation systems are predominantly limited to basic text translation tasks and lack support for customized translation requirements through flexible interaction with prompts. For instance, key capabilities such as contextual translation (maintaining coherence across multi-turn or long-document translation) and formatted translation (preserving the original document structure, such as tables, lists, and formulas) are insufficiently addressed. These limitations hinder the adaptation of translation systems to diverse real-world scenarios, in which customized requirements are increasingly prevalent.

These two core challenges—quality-efficiency imbalance and inadequate customized translation support—severely restrict the further advancement and widespread adoption of machine translation technology, highlighting the urgent need for innovative solutions that can simultaneously address efficiency, quality, and customization demands.

To directly tackle the aforementioned two core challenges, we present the HY-MT1.5 models and corresponding technical solutions, with three key contributions that closely align with the pain points:

1. **High-performance and efficient HY-MT1.5 models:** Targeting the core challenge of balancing translation quality and efficiency, we propose HY-MT1.5-1.8B and HY-MT1.5-7B that achieve a superior performance-efficiency balance. As shown in Figure 4 and 2, the 1.8B-parameter HY-MT1.5-1.8B comprehensively outperforms mainstream medium-sized open-source models (e.g., Tower-Plus-72B Rei et al., 2025, Qwen3-32B (Team, 2025)) and commercial translation APIs, reaching the 90th percentile of ultra-large closed-source models like Gemini-3.0-Pro (DeepMind, 2025). The 7B-parameter HY-MT1.5-7B further reaches the 95th percentile of Gemini-3.0-Pro (DeepMind, 2025) on the Flores-200 dataset (Team et al., 2022) and even surpasses it on WMT25 (Kocmi et al., 2025) and Mandarin-minority

language test sets, while maintaining efficient deployment capabilities that facilitate widespread practical application.

2. **Holistic and effective training scheme:** We develop a tailored training framework for machine translation, integrating general pre-training, MT-oriented pre-training, supervised finetuning, on-policy distillation, and reinforcement learning. This framework enables the models to excel in both general and low-resource translation scenarios, laying a solid foundation for the superior performance of the HY-MT1.5 models.
3. **Practical distinctive features:** Addressing the limitation of insufficient customized translation support in existing systems, the HY-MT1.5 models are equipped with practical features including terminology intervention, contextual translation, and formatted translation. These capabilities enable flexible response to customized translation demands through prompt interactions, significantly enhancing adaptability to diverse real-world application scenarios and bridging the gap between academic performance and industrial needs.

The remainder of this report is organized as follows: we first elaborate on the holistic training framework and development details of the HY-MT1.5 models. Subsequently, we present extensive experimental evaluations to validate the models' performance across various representative translation benchmarks, focusing on their performance-efficiency balance and customized translation capabilities. Finally, we discuss key findings and outline future research directions.

## 2 Methodology

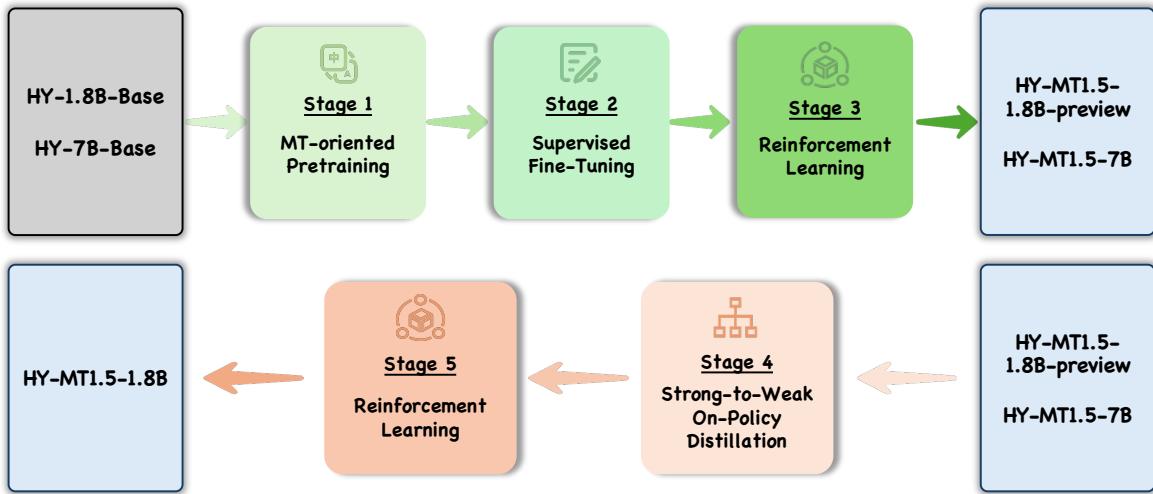


Figure 3: Training pipeline of HY-MT1.5-1.8B and HY-MT1.5-7B.

Our training framework for HY-MT1.5-1.8B follows a multi-stage pipeline designed to maximize the performance of smaller-parameter models through knowledge transfer and rigorous alignment. The overall pipeline consists of four main stages: MT-oriented Pre-training, Supervised Fine-Tuning (SFT), On-Policy Distillation, and Reinforcement Learning (RL).

### 2.1 MT-oriented Pretraining and Supervised Fine-Tuning

The initial phases of our training strategy align with the methodology described in our previous Hunyuan-MT Technical Report (Zheng et al., 2025). We use the HY-1.8B-Base<sup>1</sup> and HY-7B-Base<sup>2</sup> models as our base model to obtain HY-MT1.5-1.8B-preview and HY-MT1.5-7B.

- **Data Strategy.** We curate a massive dataset comprising high-quality multilingual monolingual corpora and parallel texts.
- **Process.** The base model undergoes Continuous Pretraining (CPT) followed by Supervised Fine-Tuning (SFT).

<sup>1</sup><https://huggingface.co/tencent/Hunyuan-1.8B-Pretrain>

<sup>2</sup><https://huggingface.co/tencent/Hunyuan-7B-Pretrain>

- **Objectives.** These stages are designed to enhance the model’s multilingual domain knowledge, translation capabilities, and adherence to translation instructions. For further details on the data curation and base training protocols, please refer to our previous work ([Zheng et al., 2025](#)).

## 2.2 Reinforcement Learning

To further align the model with human preferences and refine translation quality, we employ Reinforcement Learning. We adopt the GRPO (Group Relative Policy Optimization) ([Shao et al., 2024](#)) algorithm, which is also used in Hunyuan-MT-7B. GRPO updates the policy based on relative comparisons within groups of outputs, reducing training complexity while maintaining optimization stability.

We improve the reward modeling in the RL training of HY-MT1.5. Instead of relying on a single holistic score, we introduce a Rubrics-based Evaluation System. This multi-dimensional evaluation guides the LLM to evaluate translations with greater granularity.

We construct a structured scoring criterion set where an LLM-based evaluator scores translations across the following key dimensions:

- Accuracy: Evaluates whether the translation remains faithful to the original semantics, ensuring there are no omissions, mistranslations, or hallucinations.
- Fluency: Assesses whether the linguistic expression is natural and conforms to the grammar and idiomatic usage of the target language.
- Consistency: Checks for the consistent use of terminology, style, and context throughout the text.
- Cultural Appropriateness: Examines whether the translation adapts appropriately to the cultural background and expression habits of the target language.
- Readability: Evaluates how easy the text is to understand, ensuring clear sentence structures and distinct hierarchy.

Each dimension is assigned specific scoring standards and weights. The scores from these dimensions are aggregated to form the final reward signal. This fine-grained feedback mechanism provides the RL process with richer, more precise signals, enabling the model to improve simultaneously across multiple facets—resulting in translations that are not only correct but also natural, coherent, and culturally context-aware.

## 2.3 Strong-to-Weak On-Policy Distillation

While CPT and SFT significantly improve the 1.8B model’s performance, a performance gap relative to our larger HY-MT-7B model remains due to the inherent limitations imposed by its parameter size. To bridge this gap, we employ on-policy distillation.

Recent research ([Agarwal et al., 2024](#); [Gu et al., 2025](#); [Lu & Lab, 2025](#)) suggests that on-policy distillation is more effective than off-policy methods for improving student models. Consequently, we adopt this approach after SFT.

- **Teacher Model.** We utilize the fully trained HY-MT1.5-7B as the teacher model.
- **Data.** We collect approximately 1 million monolingual samples, covering all 33 supported languages, including specific ethnic minority languages and dialects.
- **Loss Function.** We employ per-token reverse KL divergence to align the student’s output distribution with the teacher’s. The loss function is defined as:

$$KL(\pi_\theta \parallel \pi_{\text{teacher}}) = \mathbb{E}_{x \sim \pi_\theta} [\log \pi_\theta(x_{t+1} \mid x_{1..t}) - \log \pi_{\text{teacher}}(x_{t+1} \mid x_{1..t})]$$

This process enables the 1.8B model to inherit the 7B model’s superior translation performance. Upon completion of this phase, we employ the same reinforcement learning approach utilized in the third stage to optimize the model, yielding the final model.

## 2.4 Quantization

Recent advancements in large language models (LLMs) have demonstrated remarkable success across a wide range of applications, from conversational chatbots to creative writing. However, growing concerns over data privacy, the need for offline functionality, and the high costs of large-scale cloud deployment necessitate the direct deployment of these models on edge devices, which are typically resource-constrained. Quantization has emerged as a promising technique to achieve this goal by

---

reducing model size and computational requirements through the use of lower-precision representations of model weights.

For HY-MT1.5-1.8B, the adoption of the W8A8C8-FP8 strategy effectively meets accuracy requirements, as FP8 provides strong support for LLM precision. For even lower bit-widths, the Weight-Int4 quantization strategy can further compress the 1.8B model to occupy less memory, catering to more demanding edge device constraints—though this comes at the cost of significant accuracy degradation. After comparing various quantization algorithms, we selected GPTQ (Frantar et al., 2023b) as the post-training quantization (PTQ) calibration strategy to minimize quantization error. This algorithm processes model weights layer by layer, leveraging a small amount of calibration data to minimize the reconstruction error of quantized weights, adjusting weights via an optimization process that approximates the inverse Hessian matrix. The workflow does not require model retraining; only a small calibration dataset is needed to quantize the weights, thereby improving inference efficiency and lowering the barrier to deployment.

Extremely low-bit quantization (e.g., 2-bit, 1.58-bit) has recently attracted considerable interest from researchers and shows great potential. While ultra-low-bit quantization can compress models to an extreme degree, it also leads to substantial performance degradation. To mitigate this accuracy loss, we employ quantization-aware training (QAT) to reduce precision-related degradation through training. Unlike traditional QAT approaches, and considering the distribution characteristics of smaller models, we introduce an offset to better align the distribution of quantized weights with the original pre-quantization distribution. For 2-bit quantization, we apply symmetric quantization with a bias to achieve better results, while adopting per-channel granularity to ensure both inference performance and quantization accuracy. The weights for the model HY-MT1.5-1.8B-2BIT will be released in the near future.

### 3 Experiments

#### 3.1 Automatic Metrics

To comprehensively evaluate the multilingual translation capabilities, we conducted extensive experiments using the following test sets:

- **Flores-200<sup>3</sup>** (Team et al., 2022). We select 1,056 language pairs across 33 different languages from the Flores-200 dataset. These pairs are systematically categorized into three groups: Chinese  $\leftrightarrow$  XX, English  $\leftrightarrow$  XX, and XX  $\leftrightarrow$  XX translations.
- **WMT25<sup>4</sup>** (Kocmi et al., 2025). We incorporate human evaluation sets from WMT25 with 13 language pairs (Czech to German, Ukrainian and English to Bhojpuri, Czech, Egyptian Arabic, Estonian, Icelandic, Japanese, Maasai (Kenya), Russian, Serbian (Cyrillic script), Simplified Chinese, Ukrainian).
- **Mandarin  $\leftrightarrow$  Minority Testset**. This test set comprises translations between Chinese and minority languages, namely Tibetan, Mongolian, Uyghur, and Kazakh.

For automatic evaluation, we use the neural metrics XCOMET-XXL (Guerreiro et al., 2023) and CometKiwi (Rei et al., 2022), which generally correlate with human judgments.

As presented in Table 1, our experimental results indicate that the proposed HY-MT1.5-1.8B and HY-MT1.5-7B models achieve competitive performance across the XCOMET-XXL and CometKiwi evaluation metrics. On FLORES-200, HY-MT1.5-7B performs well across translation directions: it scores 0.8690 in ZH  $\leftrightarrow$  XX, outperforming translation-specialized models like iFLYTEK-Translator (0.8196) and Doubao-Translator (0.8091), and matching medium-sized general models such as Qwen3-235B-A22B (0.8539). Its 0.9093 score in EN  $\leftrightarrow$  XX surpasses most translation-specialized models and matches Qwen3-235B-A22B (0.9029), while its 0.8098 in XX  $\leftrightarrow$  XX outperforms all evaluated translation-specialized models. On the WMT25 benchmark, HY-MT1.5-7B achieves an XCOMET-XXL score of 0.6159, significantly outperforming all compared models across all three baseline categories. This result is 0.0654 higher than that of the top-performing ultra-large general model, Gemini 3.0 Pro (0.5505), and far exceeds that of translation-specialized models such as Seed-X-PPO-7B (0.4783) and Tower-Plus-72B (0.4100). Even the smaller HY-MT1.5-1.8B model achieves an XCOMET-XXL score of 0.5308 on WMT25, outperforming many medium to small-sized general models (e.g., Qwen3-32B: 0.3605) and translation-specialized models.

A particularly noteworthy advantage is the exceptional performance of the HY-MT1.5 models on Mand.  $\leftrightarrow$  Min. translation pairs, a critical task for Chinese-centric multilingual translation. HY-MT1.5-7B

<sup>3</sup><https://huggingface.co/datasets/Muennighoff/flores200>

<sup>4</sup><https://github.com/wmt-conference/wmt25-general-mt/blob/main/data/wmt25-genmt-humeval.jsonl>

<sup>5</sup><https://github.com/Tencent-Hunyuan/Hunyuan-MT>

Table 1: Performances of state-of-the-art models on Flores-200 (Team et al., 2022), WMT25 (Kocmi et al., 2025), and Mandarin $\leftrightarrow$ Minority translation. Specifically, we report the Chinese-centric ( $ZH \leftrightarrow XX$ ), English-centric ( $EN \leftrightarrow XX$ ),  $XX \leftrightarrow XX$ , and  $Mand. \leftrightarrow Min.$  performances of HY-MT1.5-1.8B, HY-MT1.5-7B, and prominent existing systems. Here,  $Mand. \leftrightarrow Min.$  denotes Mandarin $\leftrightarrow$ Minority translation. Values denoted with \* indicate that the metric scores for the corresponding model are computed only for the supported language pairs; approximately half of the total languages are unsupported. Values replaced by – indicate that the model does not support the language pairs of the corresponding test set. Models with open-source weights are marked with <sup>†</sup>. HY-MT1.0-7B refers to our previous model, Hunyuan-MT-7B<sup>5</sup>. Baselines are categorized into three groups: (1) ultra-large general models, (2) medium to small-sized general models, and (3) translation-specialized models.

| Models   | Metrics    | FLORES-200              |                         |                         | WMT25   | Mand. $\leftrightarrow$ Min. |
|--|------------|-------------------------|-------------------------|-------------------------|---------|------------------------------|
|  |            | $ZH \leftrightarrow XX$ | $EN \leftrightarrow XX$ | $XX \leftrightarrow XX$ |         |                              |
| <b>Gemini 3.0 pro</b><br>(DeepMind, 2025)                      | XCOMET-XXL | 0.8982                  | 0.9413                  | 0.8773                  | 0.5505  | 0.5921                       |
|  | CometKiwi  | 0.7882                  | 0.8809                  | 0.7530                  | 0.6552  | 0.5274                       |
| <b>DeepSeek-V3.2<sup>†</sup></b><br>(DeepSeek-AI et al., 2025) | XCOMET-XXL | 0.8752                  | 0.9231                  | 0.8421                  | 0.5013  | 0.5133                       |
|  | CometKiwi  | 0.7798                  | 0.8736                  | 0.7521                  | 0.6353  | 0.5253                       |
| <b>Qwen3-235B-A22B<sup>†</sup></b><br>(Team, 2025)             | XCOMET-XXL | 0.8539                  | 0.9029                  | 0.8018                  | 0.4375  | 0.4493                       |
|  | CometKiwi  | 0.7651                  | 0.8586                  | 0.7313                  | 0.5820  | 0.4456                       |
| <b>Qwen3-32B<sup>†</sup></b><br>(Team, 2025)                   | XCOMET-XXL | 0.8185                  | 0.8670                  | 0.7433                  | 0.3605  | 0.4110                       |
|  | CometKiwi  | 0.7429                  | 0.8329                  | 0.6965                  | 0.5016  | 0.3841                       |
| <b>Phi-4-mini-3.8B<sup>†</sup></b><br>(Microsoft et al., 2025) | XCOMET-XXL | 0.5839                  | 0.6284                  | 0.4606                  | 0.2357  | 0.3542                       |
|  | CometKiwi  | 0.4327                  | 0.6182                  | 0.3482                  | 0.2819  | 0.2003                       |
| <b>Tower-Plus-72B<sup>†</sup></b><br>(Rei et al., 2025)        | XCOMET-XXL | 0.7969                  | 0.8416                  | 0.7002                  | 0.4100  | 0.3855                       |
|  | CometKiwi  | 0.7182                  | 0.8113                  | 0.6553                  | 0.5554  | 0.3540                       |
| <b>Seed-X-PPO-7B<sup>†</sup></b><br>(Cheng et al., 2025)       | XCOMET-XXL | 0.7856                  | 0.8312                  | 0.6896                  | 0.4783  | 0.4206                       |
|  | CometKiwi  | 0.7145                  | 0.8160                  | 0.6436                  | 0.6623  | 0.4861                       |
| <b>GemmaX2-28-2B-v0.1<sup>†</sup></b><br>(Cui et al., 2025)    | XCOMET-XXL | 0.7142                  | 0.8208                  | 0.6376                  | 0.2679  | 0.3596                       |
|  | CometKiwi  | 0.6746                  | 0.8095                  | 0.6310                  | 0.3750  | 0.3981                       |
| <b>Google-Translator</b>                                       | XCOMET-XXL | 0.6929                  | 0.7700                  | 0.6225                  | 0.3893  | 0.3338                       |
|  | CometKiwi  | 0.6169                  | 0.7552                  | 0.5947                  | 0.5938  | 0.3209                       |
| <b>Baidu-Translator</b>  | XCOMET-XXL | 0.7690                  | 0.8209                  | 0.6807                  | 0.2662  | –                            |
|  | CometKiwi  | 0.6789                  | 0.7770                  | 0.6369                  | 0.3284  | –                            |
| <b>iFLYTEK-Translator</b>                                      | XCOMET-XXL | 0.8196                  | 0.8397                  | 0.7467                  | 0.2742  | 0.5011                       |
|  | CometKiwi  | 0.7326                  | 0.8035                  | 0.6868                  | 0.4747  | 0.4871                       |
| <b>Doubao-Translator</b>                                       | XCOMET-XXL | 0.8091                  | 0.8673                  | 0.7653                  | 0.3567  | 0.5314*                      |
|  | CometKiwi  | 0.7156                  | 0.8349                  | 0.6993                  | 0.5869  | 0.4061*                      |
| <b>Microsoft-Translator</b>                                    | XCOMET-XXL | 0.8234                  | 0.8917                  | 0.8007                  | 0.3993  | 0.5196                       |
|  | CometKiwi  | 0.7297                  | 0.8546                  | 0.7253                  | 0.5994  | 0.3218                       |
| <b>Youdao-Translator (Lite)</b>                                | XCOMET-XXL | 0.8474*                 | 0.8930*                 | 0.7840*                 | 0.4499* | 0.4525*                      |
|  | CometKiwi  | 0.7720*                 | 0.8656*                 | 0.7599*                 | 0.6520* | 0.5050*                      |
| <b>HY-MT1.0-7B<sup>†</sup></b>                                 | XCOMET-XXL | 0.8643                  | 0.9065                  | 0.7829                  | 0.6023  | 0.6082                       |
|  | CometKiwi  | 0.7913                  | 0.8610                  | 0.7210                  | 0.6735  | 0.4162                       |
| <b>HY-MT1.5-1.8B<sup>†</sup></b>                               | XCOMET-XXL | 0.8361                  | 0.8942                  | 0.7840                  | 0.5308  | 0.5806                       |
|  | CometKiwi  | 0.7655                  | 0.8411                  | 0.7182                  | 0.6195  | 0.4084                       |
| <b>HY-MT1.5-7B<sup>†</sup></b>                                 | XCOMET-XXL | 0.8690                  | 0.9093                  | 0.8098                  | 0.6159  | 0.6174                       |
|  | CometKiwi  | 0.7924                  | 0.8650                  | 0.7336                  | 0.6885  | 0.4455                       |

achieves an XCOMET-XXL score of 0.6174 in this direction, which outperforms all evaluated baselines. It exceeds the top-performing ultra-large general model Gemini 3.0 Pro by 0.0253 (0.6174 vs. 0.5921). Even the 1.8B variant (0.5806) outperforms most baselines in this setting, including ultra-large models like DeepSeek-V3.2 (0.5133) and translation-specialized models such as iFLYTEK-Translator (0.5011).

The HY-MT1.5 models balance superior performance with high parameter efficiency. For example, HY-MT1.5-7B (7B parameters) outperforms the larger Tower-Plus-72B (72B) across FLORES-200 and Mand.  $\leftrightarrow$  Min. translation. Both models also outperform commercial translators (e.g., Google-Translator) and smaller general models (e.g., Qwen3-32B), validating our model design.

Moreover, HY-MT1.5-7B outperforms HY-MT1.5-1.8B across all tasks, with the largest gain (16.0%) on

WMT25, indicating that moderate model scaling boosts translation quality. As open-source models, they enable broader academic and industrial adoption than closed-source alternatives such as Gemini 3.0 Pro.

### 3.2 Human Evaluation

Table 2: Human evaluation of translation quality for the Chinese-to-English ( $ZH \Rightarrow EN$ ) and English-to-Chinese ( $EN \Rightarrow ZH$ ) directions. The highest scores are shown in bold.

| Model                | $ZH \Rightarrow EN$ | $EN \Rightarrow ZH$ | Avg.        |
|----------------------|---------------------|---------------------|-------------|
| Baidu-Translator     | 2.75                | 2.46                | 2.55        |
| iFLYTEK-Translator   | 2.88                | 2.54                | 2.65        |
| Doubaotranslator     | 2.97                | 2.48                | 2.64        |
| Microsoft-Translator | 2.94                | 2.57                | 2.69        |
| Google-Translator    | 2.84                | 2.10                | 2.34        |
| <b>HY-MT1.5-1.8B</b> | <b>3.01</b>         | <b>2.61</b>         | <b>2.74</b> |

To address the limitations of automatic evaluation (Lavie et al., 2025), we conduct human evaluation in which multilingual experts rate translations on a 0–4 scale, focusing on pre-annotated error-prone points and considering accuracy, fluency, and idiomaticity.

As shown in Table 2, the evaluated models are divided into two tiers: the lightweight specialized model HY-MT1.5-1.8B and mainstream commercial translation systems. HY-MT1.5-1.8B achieves the highest average score (2.74), outperforming all commercial systems, which is consistent with automatic evaluation. Most systems perform better in  $ZH \Rightarrow EN$  than  $EN \Rightarrow ZH$ , mainly due to the complexity of Chinese syntax generation.

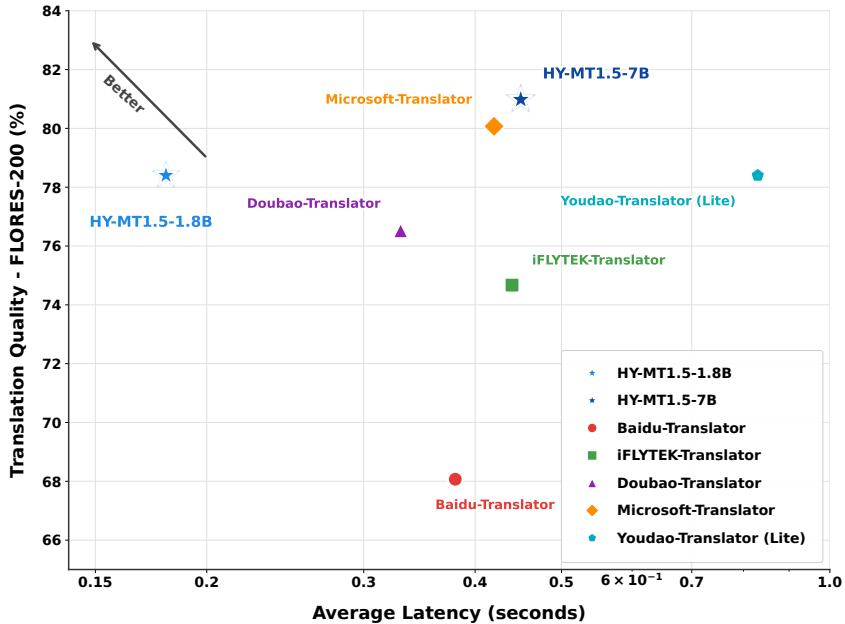


Figure 4: Average response time versus translation quality for different translation models.

### 3.3 Efficiency of HY-MT1.5 Models

To evaluate the translation efficiency of the HY-MT1.5 models, a standardized speed test was conducted: 100 Chinese texts (average length 50 tokens, covering daily and general business scenarios) were sequentially translated into English, with the average response time used as the primary metric.

As shown in Figure 4, the HY-MT1.5 models exhibit a superior balance between translation quality and response efficiency. Specifically, HY-MT1.5-1.8B achieves an FLORES-200 quality score of approximately 78% while maintaining an average response time of 0.18 seconds, indicating a clear speed advantage. The length of the test texts aligns with practical translation needs in daily communication, office work, and quick information retrieval, and the 0.18-second instant response time eliminates user waiting,

---

fully meeting real-time interaction requirements. Overall, HY-MT1.5-1.8B’s high efficiency, supported by optimized model design and inference logic, makes it well-suited for high-throughput, real-time translation scenarios such as instant messaging, intelligent customer service, and mobile translation applications.

For HY-MT1.5-7B, its quality score exceeds 80% (exceeding that of most models in the figure), and its average response time is 0.45 seconds. This response time is comparable to that of Microsoft-Translator, whereas its translation quality is distinctly higher than that of Microsoft-Translator. This result confirms that the HY-MT1.5 models combine high translation accuracy with fast inference speed, making them suitable for scenarios requiring both high-quality translation and real-time responsiveness.

### 3.4 Practical Distinctive Features

To address the core limitations of existing translation systems—overreliance on basic text translation and inadequate support for customized requirements via flexible prompts—we propose three distinctive features, validated through scenario-based case studies in Table 3. These features target terminology accuracy, context-aware disambiguation, and format preservation, which are critical for real-world scenario adaptation but underexplored in current frameworks.

First, terminology-guided translation resolves inaccurate rendering of cultural or domain-specific terms via a term-anchored prompt template. As shown in Scenario 1, the term “混元珠” is transliterated as the semantically ambiguous “Hunyuan” without terminology prompts; with the authorized mapping “Chaos Pearl” injected into the prompt, the model generates a precise, consistent translation.

Second, context-aware translation mitigates lexical ambiguity in contextualized tasks using a context-integrated prompt template. For the phrase “The Educational and Inspirational Poetics of Pilots” in Scenario 2, the model misinterprets “pilot” as “飞行员” in the absence of contextual cues; when TV series context is provided, it correctly identifies “pilot” as “试播集” and produces a contextually coherent result.

Third, format-preserving translation preserves the structural integrity of tagged text by using a format-constrained prompt template. In Scenario 3, the model outputs disorganized tags (e.g., <1> instead of <s1>) without format prompts. With explicit rules to preserve <sn>tags and to wrap outputs in <target>tags, it generates translations that maintain both semantic fidelity and structural consistency.

Collectively, these prompt-driven features enable the model to transcend basic translation tasks, delivering customized solutions for terminology-sensitive, context-dependent, and format-constrained scenarios.

### 3.5 Quantization Experiment

Large Language Models (LLMs) have achieved significant success across a range of applications, but high resource demands hinder their deployment on resource-constrained edge devices. Quantization, which reduces model size and computational cost by using low-precision weight representations, is a key solution (Kulkarni et al., 2022). For the HY-MT1.5-1.8B model, we tested two quantization strategies (Int4 and FP8) and adopted the GPTQ algorithm (Frantar et al., 2023a) for Post-Training Quantization (PTQ) to minimize errors. GPTQ processes weights layer-wise with small calibration data, avoiding retraining and improving deployment efficiency.

Table 4 presents the performance of different quantization schemes (original, Int4, FP8) on multiple translation tasks, evaluated by XCOMET-XXL and CometKiwi metrics. It shows that FP8 quantization preserves accuracy close to the original model (e.g., ZH ⇔ XX XCOMET-XXL score: 0.8379 for FP8 vs. 0.8361 for the original), whereas Int4 quantization reduces model size but causes noticeable accuracy degradation.

## 4 Conclusion

This report introduces the HY-MT1.5 models (1.8B and 7B) and a dedicated machine translation training framework that effectively addresses two core challenges: the quality-efficiency trade-off and inadequate support for customized translation needs. By integrating general pre-training, MT-oriented pre-training, supervised fine-tuning, on-policy distillation, and reinforcement learning with a rubrics-based evaluation system, the models achieve a superior balance of performance and efficiency, as well as strong adaptability to practical scenarios.

Experimental results on key benchmarks (Flores-200, WMT25, Mandarin-minority languages) confirm the competitiveness of HY-MT1.5 models. The 1.8B model outperforms mainstream medium-sized open-source models and commercial APIs, achieving performance comparable to that of ultra-large

Table 3: Case studies across different scenarios, including terminology, context and format translation.

| Scenario 1: Terminology Translation |   |
|-------------------------------------|---|
| <b>Example #1</b>                   | 孕育出一颗混元珠  |
| <b>Terminology</b>                  | "混元珠": Chaos Pearl  |
| <b>Prompt Template</b>              | 参考下面的翻译:<br>{terminology} 翻译成{terminology_target_language}  |
|                                     | 将以下文本翻译为{target_language}, 注意只需要输出翻译后的结果, 不要额外解释:<br>{source text}  |
| <b>Without Terminology</b>          | Give birth to a <b>Hunyuan</b> Pearl  |
| <b>With Terminology</b>             | Give birth to a <b>Chaos</b> Pearl  |
| Scenario 2: Context Translation     |   |
| <b>Example #2</b>                   | The Educational and Inspirational Poetics of Pilots   |
| <b>Prompt Template</b>              | {context}<br>参考上面的信息, 把下面的文本翻译成{target_language}, 注意不需要翻译上文, 也不要额外解释:<br>{source text}  |
| <b>Context</b>                      | However, given the rise of the boxed DVD model that enables viewers to consume a series chronologically, viewers are now more likely to start at the beginning of a series, as they recognize that many complex television programs are designed to be watched from the start. For commercial television programs, that start is the unusual entity known as the pilot. |
| <b>Without Context</b>              | 飞行员们身上的教育与激励性诗歌之美   |
| <b>With Context</b>                 | 《试播集: 教育与启示性的诗歌艺术》  |
| Scenario 3: Format Translation      |   |
| <b>Example #3</b>                   | <s1>The rain it raineth every day, </s1><s2>as Shakespeare noted, apparently even on Saturn. The cosmos, it seems, is no comfort at this moment.</s2><s3>Adam Gopnik on confronting uncertain times.</s3>   |
| <b>Prompt Template</b>              | 将以下<source></source>之间的文本翻译为中文, 注意只需要输出翻译后的结果, 不要额外解释, 原文中的<sn></sn>标签表示标签内文本包含格式信息, 需要在译文中相应的位置尽量保留该标签。输出格式为: <target>str</target>   |
|                                     | <source>{{{{source_text_with_format}}}}</source>  |
| <b>Without Format</b>               | <1> 雨天天都在下, </s1><s2>正如莎士比亚所言, 显然在土星上也是如此。此刻, 宇宙似乎并不能给人带来任何安慰。</s2><s3>亚当·戈普尼克谈如何面对充满不确定性的时代。</s3>  |
| <b>With Format</b>                  | <target><s1> 雨日日不停地下着, </s1><s2>正如莎士比亚所言, 看来在土星上也是如此。此刻的宇宙似乎并不能带来安慰。</s2><s3>亚当·戈普尼克谈面对不确定时代的挑战。</s3></target>  |

closed-source models such as Gemini-3.0-Pro, which rank in the 90th percentile. The 7B model further improves performance, surpassing Gemini-3.0-Pro on WMT25 and Mandarin-minority translation, and reaching its 95th percentile on Flores-200. With an average response time of 0.18 seconds, HY-MT1.5-1.8 B is well suited to high-throughput and real-time scenarios.

Beyond core translation performance, the models feature three practical capabilities: terminology intervention, contextual translation, and formatted translation. These prompt-enabled features flexibly meet customized demands, bridging the gap between academic research and industrial applications. Quantization experiments verify deployment potential—FP8 quantization maintains performance close to

Table 4: Translation performance metrics of different quantization schemes. The highest and second-best scores are shown in **bold** and underlined, respectively.

| Models             | Metrics    | FLORES-200    |               |               |
|--------------------|------------|---------------|---------------|---------------|
|                    |            | ZH ↔ XX       | EN ↔ XX       | XX ↔ XX       |
| HY-MT1.5-1.8B      | XCOMET-XXL | 0.8361        | <b>0.8942</b> | <b>0.7840</b> |
|                    | CometKiwi  | <u>0.7655</u> | <b>0.8411</b> | <b>0.7182</b> |
| HY-MT1.5-1.8B-FP8  | XCOMET-XXL | <b>0.8379</b> | 0.8905        | 0.7794        |
|                    | CometKiwi  | <u>0.7659</u> | <u>0.8396</u> | <u>0.7156</u> |
| HY-MT1.5-1.8B-Int4 | XCOMET-XXL | 0.8060        | 0.8665        | 0.7336        |
|                    | CometKiwi  | 0.7462        | 0.8234        | 0.6884        |

the original model while reducing resource consumption, facilitating application on resource-constrained devices.

Future work will focus on three directions: 1) Expanding language coverage to more low-resource and underrepresented languages, enhancing the inclusivity of MT technology; 2) Optimizing the training framework with more efficient knowledge distillation and reinforcement learning to improve the performance-efficiency ratio of small and medium-sized models; 3) Deepening customized translation research by integrating advanced prompt engineering and domain adaptation to better meet industry-specific needs. The HY-MT1.5 models are expected to provide high-quality, efficient, and flexible translation solutions for academic and industrial communities, advancing the development and popularization of MT technology.

## 5 Contributions

### 5.1 Core Contributors

Mao Zheng, Zheng Li, Tao Chen, Mingyang Song, Di Wang

### 5.2 Contributors

Feng Zhang, Tinghao Yu, Chengcheng Xu, Zhenyu Huang, Liya Zhan, Jun Xia, Jiaqi Zhu, Xingwu Sun, Yufei Wang, Can Xu, Liang Dong, Huxin Peng, Fei Cheng, Zheng Zhang, Lili He, Huashuo Li, Decheng Wu, Guanghua Yu, Kai Wang, Haozhao Kuang

## References

- Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes, 2024. URL <https://arxiv.org/abs/2306.13649>.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL <http://arxiv.org/abs/1409.0473>.
- Peter F. Brown, John Cocke, Stephen Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. A statistical approach to machine translation. *Comput. Linguistics*, 16(2):79–85, 1990.
- Peter F. Brown, Stephen Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Comput. Linguistics*, 19(2):263–311, 1993.
- Shanbo Cheng, Yu Bao, Qian Cao, Luyang Huang, Liyan Kang, Zhicheng Liu, Yu Lu, Wenhao Zhu, Jingwen Chen, Zhichao Huang, Tao Li, Yifu Li, Huiying Lin, Sitong Liu, Ningxin Peng, Shuaijie She, Lu Xu, Nuo Xu, Sen Yang, Runsheng Yu, Yiming Yu, Liehao Zou, Hang Li, Lu Lu, Yuxuan Wang, and Yonghui Wu. Seed-x: Building strong multilingual translation llm with 7b parameters, 2025. URL <https://arxiv.org/abs/2507.13618>.

---

Menglong Cui, Pengzhi Gao, Wei Liu, Jian Luan, and Bin Wang. Multilingual machine translation with open large language models at practical scale: An empirical study, 2025. URL <https://arxiv.org/abs/2502.02481>.

DeepMind. Introducing gemini 3. <https://blog.google/products/gemini/gemini-3-collection/>, 2025. Accessed: 2025-12-29.

DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Erhang Li, Fangqi Zhou, Fangyun Lin, Fucong Dai, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Hao Li, Haofen Liang, Haoran Wei, Haowei Zhang, Haowen Luo, Haozhe Ji, Honghui Ding, Hongxuan Tang, Huanqi Cao, Huazuo Gao, Hui Qu, Hui Zeng, Jialiang Huang, Jiashi Li, Jiaxin Xu, Jiewen Hu, Jingchang Chen, Jingting Xiang, Jingyang Yuan, Jingyuan Cheng, Jinhua Zhu, Jun Ran, Junguang Jiang, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kaige Gao, Kang Guan, Kexin Huang, Kexing Zhou, Kezhao Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Wang, Liang Zhao, Liangsheng Yin, Lihua Guo, Lingxiao Luo, Linwang Ma, Litong Wang, Liyue Zhang, M. S. Di, M. Y Xu, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Panpan Huang, Peixin Cong, Peiyi Wang, Qiancheng Wang, Qihao Zhu, Qingyang Li, Qinyu Chen, Qiushi Du, Ruiling Xu, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runqiu Yin, Runxin Xu, Ruomeng Shen, Ruoyu Zhang, S. H. Liu, Shanghao Lu, Shangyan Zhou, Shanhua Chen, Shaofei Cai, Shaoyuan Chen, Shengding Hu, Shengyu Liu, Shiqiang Hu, Shirong Ma, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, Songyang Zhou, Tao Ni, Tao Yun, Tian Pei, Tian Ye, Tianyuan Yue, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjie Pang, Wenjing Luo, Wenjun Gao, Wentao Zhang, Xi Gao, Xiangwen Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaokang Zhang, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xingyou Li, Xinyu Yang, Xinyuan Li, Xu Chen, Xuecheng Su, Xuehai Pan, Xuheng Lin, Xuwei Fu, Y. Q. Wang, Yang Zhang, Yanhong Xu, Yanru Ma, Yao Li, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Qian, Yi Yu, Yichao Zhang, Yifan Ding, Yifan Shi, Yiliang Xiong, Ying He, Ying Zhou, Yinmin Zhong, Yishi Piao, Yisong Wang, Yixiao Chen, Yixuan Tan, Yixuan Wei, Yiyang Ma, Yiyuan Liu, Yonglun Yang, Yongqiang Guo, Yongtong Wu, Yu Wu, Yuan Cheng, Yuan Ou, Yuanfan Xu, Yuduan Wang, Yue Gong, Yuhan Wu, Yuheng Zou, Yukun Li, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehua Zhao, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhixian Huang, Zhiyu Wu, Zhuoshu Li, Zhuping Zhang, Zian Xu, Zihao Wang, Zihui Gu, Zijia Zhu, Zilin Li, Zipeng Zhang, Ziwei Xie, Ziyi Gao, Zizheng Pan, Zongqing Yao, Bei Feng, Hui Li, J. L. Cai, Jiaqi Ni, Lei Xu, Meng Li, Ning Tian, R. J. Chen, R. L. Jin, S. S. Li, Shuang Zhou, Tianyu Sun, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xinnan Song, Xinyi Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, Dongjie Ji, Jian Liang, Jianzhong Guo, Jin Chen, Leyi Xia, Miaojun Wang, Mingming Li, Peng Zhang, Ruyi Chen, Shangmian Sun, Shaoqing Wu, Shengfeng Ye, T. Wang, W. L. Xiao, Wei An, Xianzu Wang, Xiaowen Sun, Xiaoxiang Wang, Ying Tang, Yukun Zha, Zekai Zhang, Zhe Ju, Zhen Zhang, and Zihua Qu. Deepseek-v3.2: Pushing the frontier of open large language models, 2025. URL <https://arxiv.org/abs/2512.02556>.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers, 2023a. URL <https://arxiv.org/abs/2210.17323>.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers, 2023b. URL <https://arxiv.org/abs/2210.17323>.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models, 2025. URL <https://arxiv.org/abs/2306.08543>.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. xcomet: Transparent machine translation evaluation through fine-grained error detection, 2023. URL <https://arxiv.org/abs/2310.10482>.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondrej Bojar, Anton Dvorkovich, Christian Fellermann, 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 Popovic, Mariya Shmatova, Steinthór Steinþrímsson, and Vilém Zouhar. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz (eds.), *Proceedings of the Ninth Conference on Machine Translation, WMT 2024, Miami, FL, USA, November 15-16, 2024*, pp. 1–46. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.WMT-1.1. URL <https://doi.org/10.18653/v1/2024.wmt-1.1>.

---

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 Lakouagna, Jessica Lundin, Christof Monz, Kenton Murray, Masaaki Nagata, Stefano Perrella, Lorenzo Proietti, Martin Popel, Maja Popović, Parker Riley, Mariya Shmatova, Steinthór Steinþrímsson, Lisa Yankovskaya, and Vilém Zouhar. Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz (eds.), *Proceedings of the Tenth Conference on Machine Translation*, pp. 355–413, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-341-8. doi: 10.18653/v1/2025.wmt-1.22. URL <https://aclanthology.org/2025.wmt-1.22/>.

Uday Kulkarni, Abhishek S Hosamani, Abhishek S Masur, Shashank Hegde, Ganesh R Vernekar, and K Siri Chandana. A survey on quantization methods for optimization of deep neural networks. In *2022 International Conference on Automation, Computing and Renewable Systems (ICACRS)*, pp. 827–834, 2022. doi: 10.1109/ICACRS55517.2022.10028742.

Alon Lavie, Greg Hanneman, Sweta Agrawal, Diptesh Kanodia, Chi-Kiu Lo, Vilém Zouhar, Frederic Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Deoghare, Archchana Sindhuwan, Jiayi Wang, David Ifeoluwa Adelani, Brian Thompson, Tom Kocmi, Markus Freitag, and Daniel Deutsch. Findings of the WMT25 shared task on automated translation evaluation systems: Linguistic diversity is challenging and references still help. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz (eds.), *Proceedings of the Tenth Conference on Machine Translation*, pp. 436–483, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-341-8. doi: 10.18653/v1/2025.wmt-1.24. URL <https://aclanthology.org/2025.wmt-1.24/>.

Kevin Lu and Thinking Machines Lab. On-policy distillation. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20251026. <https://thinkingmachines.ai/blog/on-policy-distillation>.

Microsoft, :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, Mei Gao, Min Gao, Amit Garg, Abhishek Goswami, Junheng Hao, Amr Hendy, Yuxuan Hu, Xin Jin, Mahmoud Khademi, Dongwoo Kim, Young Jin Kim, Gina Lee, Jinyu Li, Yunsheng Li, Chen Liang, Xihui Lin, Zeqi Lin, Mengchen Liu, Yang Liu, Gilsinia Lopez, Chong Luo, Piyush Madan, Vadim Mazalov, Arindam Mitra, Ali Mousavi, Anh Nguyen, Jing Pan, Daniel Perez-Becker, Jacob Platin, Thomas Portet, Kai Qiu, Bo Ren, Liliang Ren, Sambuddha Roy, Ning Shang, Yelong Shen, Saksham Singhal, Subhajit Som, Xia Song, Tetyana Sych, Praneetha Vaddamanu, Shuohang Wang, Yiming Wang, Zhenghao Wang, Haibin Wu, Haoran Xu, Weijian Xu, Yifan Yang, Ziyi Yang, Donghan Yu, Ishmam Zabir, Jianwen Zhang, Li Lyra Zhang, Yunan Zhang, and Xiren Zhou. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras, 2025. URL <https://arxiv.org/abs/2503.01743>.

Jianhui Pang, Fanghua Ye, Derek Fai Wong, Dian Yu, Shuming Shi, Zhaopeng Tu, and Longyue Wang. Salute the classic: Revisiting challenges of machine translation in the age of large language models. *Trans. Assoc. Comput. Linguistics*, 13:73–95, 2025. doi: 10.1162/TACL\_A\_00730. URL [https://doi.org/10.1162/tacl\\_a\\_00730](https://doi.org/10.1162/tacl_a_00730).

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pp. 311–318. ACL, 2002. doi: 10.3115/1073083.1073135. URL <https://aclanthology.org/P02-1040/>.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C. Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte M. Alves, Alon Lavie, Luisa Coheur, and André F. T. Martins. Cometkiwi: Ist-unbabel 2022 submission for the quality estimation shared task, 2022. URL <https://arxiv.org/abs/2209.06243>.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, João Alves, Pedro Teixeirinha, Amin Farajian, and André F. T. Martins. Tower+: Bridging generality and translation specialization in multilingual llms, 2025. URL <https://arxiv.org/abs/2506.17080>.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300, 2024. doi: 10.48550/ARXIV.2402.03300. URL <https://doi.org/10.48550/arXiv.2402.03300>.

---

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pp. 3104–3112, 2014. URL <https://proceedings.neurips.cc/paper/2014/hash/a14ac55a4f27472c5d894ec1c3c743d2-Abstract.html>.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No language left behind: Scaling human-centered machine translation, 2022. URL <https://arxiv.org/abs/2207.04672>.

Qwen Team. Qwen3 technical report, 2025. URL <https://arxiv.org/abs/2505.09388>.

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

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google’s neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

Mao Zheng, Zheng Li, Bingxin Qu, Mingyang Song, Yang Du, Mingrui Sun, and Di Wang. Hunyuan-mt technical report, 2025. URL <https://arxiv.org/abs/2509.05209>.

Wenhai Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. Multilingual machine translation with large language models: Empirical results and analysis. In Kevin Duh, Helena Gómez-Adorno, and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pp. 2765–2781. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-NAACL.176. URL <https://doi.org/10.18653/v1/2024.findings-naacl.176>.