



Hy-MT2: A Family of Fast, Efficient and Powerful Multilingual Translation Models in the Wild

Tencent Hunyuan Team

Abstract

Hy-MT2 is a family of **fast-thinking multilingual translation models** designed for complex real-world scenarios. It includes three model sizes: **1.8B, 7B, and 30B-A3B (MoE)**, all of which support translation among **33 languages** and effectively **follow translation instructions in multiple languages**. For on-device deployment, with AngelSlim 1.25-bit extreme quantization, the **1.8B** model requires only **440 MB** of storage and improves inference speed by **1.5x**. Multi-dimensional evaluations show that Hy-MT2 delivers outstanding performance across general, real-world business, domain-specific, and instruction-following translation tasks. The 7B and 30B models **outperform open-source models such as DeepSeek-V4-Pro and Kimi K2.6 in fast-thinking mode**, while the lightweight 1.8B model also **surpasses mainstream commercial APIs from providers such as Microsoft and Doubao overall**.

<https://huggingface.co/collections/tencent/hy-mt2>

<https://github.com/Tencent-Hunyuan/Hy-MT2>

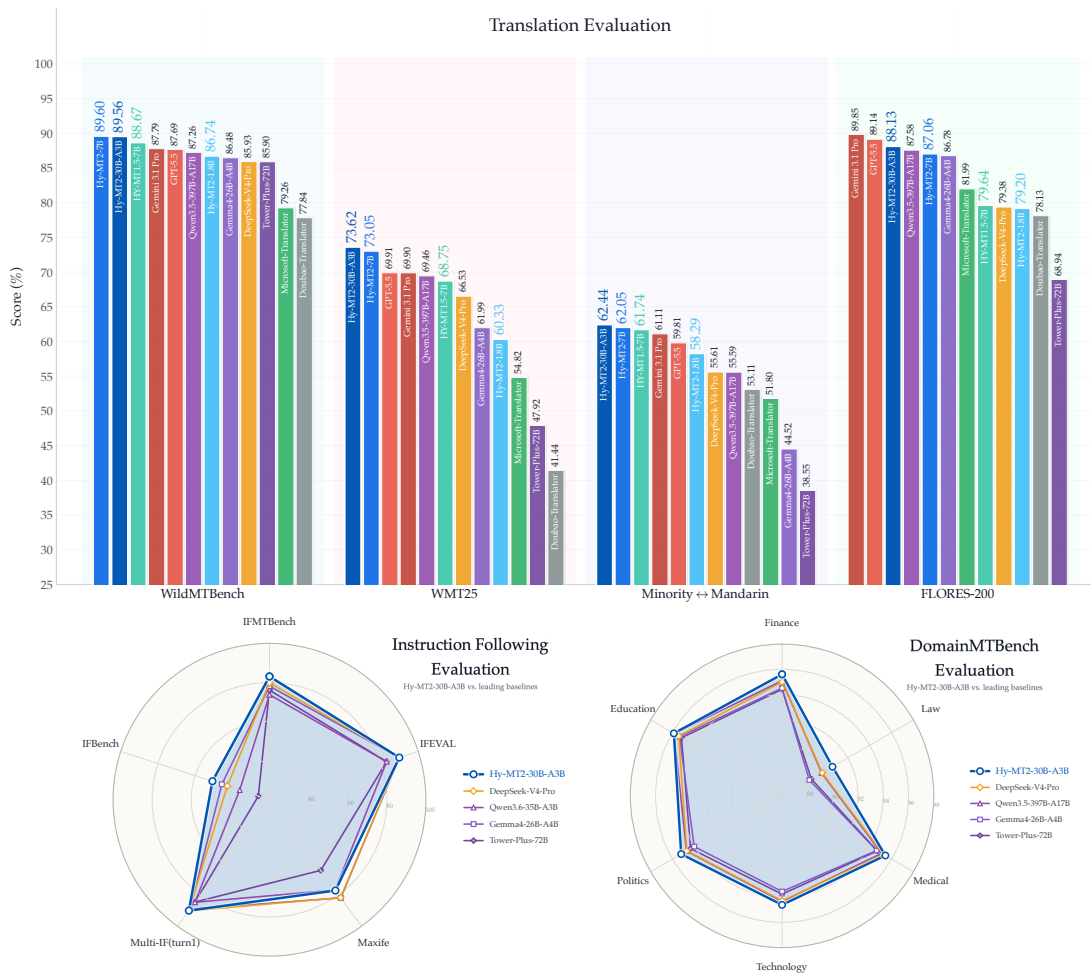


Figure 1: Benchmark performance of Hy-MT2 models and state-of-the-art baselines.

1 From Hy-MT1.5 to Hy-MT2

After its release, Hy-MT1.5 (Zheng et al., 2025) attracted broad attention from both the open-source community and real-world business applications. As the model was adopted in more practical translation scenarios, community and business feedback revealed that Hy-MT1.5 still had room for improvement in domain-specific translation, real-world scenario translation, translation instruction following, and efficient on-device deployment. Meanwhile, the substantial quality improvements achieved by Hy3-preview further motivated us to leverage it as a strong teacher model to improve the performance of the Hy translation model. To further address these limitations, we propose the Hy-MT2 model family.

First, domain-specific and real-world scenario translation remain challenging for Hy-MT1.5. Professional domains such as finance, law, and medicine contain a large number of domain-specific terms and established industry translations, placing higher requirements on translation accuracy and consistency. Real-world business scenarios, such as webpages, meetings, and social content, involve more diverse text formats and usage requirements. To address these challenges, Hy-MT2 strengthens translation capabilities for professional domains and real-world application scenarios, enabling the model to better adapt to translation needs across different domains, sources, and text forms.

Second, in practical use, users often impose additional constraints on translation outputs, such as keeping certain words untranslated, controlling the translation style, or producing outputs according to a specified template. In such scenarios, Hy-MT1.5 may ignore constraints or fail to satisfy the specified requirements. To this end, Hy-MT2 enhances multilingual translation instruction understanding and execution, enabling the model to reliably follow user requirements in different languages, including those related to style, format, and other translation constraints, as illustrated in Table 1.

In addition, community feedback indicates that Hy-MT1.5-7B still has a clear gap in translation quality compared with the strongest closed-source models, such as Gemini 3.1 Pro (DeepMind, 2025) and GPT-5.5 (OpenAI, 2026). Prior research and model practices suggest that scaling up model size generally helps improve understanding, expression, and instruction-following capabilities in complex translation scenarios. However, representative large-scale translation models, such as TransGemma-27B, mostly adopt dense architectures, leading to high inference costs and making them less suitable for practical service deployment. Therefore, Hy-MT2 introduces a mixture-of-experts architecture and releases Hy-MT2-30B-A3B to achieve a better balance between translation quality and inference efficiency.

Finally, real-world business deployment also exposed the limitations of Hy-MT1.5 in on-device efficiency. The 4-bit quantized version of Hy-MT1.5-1.8B still requires more than 1GB of storage, and its inference speed is insufficient for some low-latency translation scenarios. To address this issue, Hy-MT2 further explores ultra-low-bit quantization and implements 1.25-bit extreme quantization based on Hunyuan’s AngelSlim technology. This version requires only about 440MB of storage for deployment and achieves a $1.5\times$ inference speedup over the 4-bit quantized Hy-MT1.5 on Apple A15, significantly reducing on-device deployment costs while improving inference efficiency.

Overall, Hy-MT2 systematically addresses the limitations of Hy-MT1.5 in domain-specific translation, real-world scenario translation, translation instruction following, the performance gap with the strongest closed-source models, and efficient on-device deployment. It establishes a high-quality, efficient, and multi-capability multilingual translation model family that is better suited for real-world applications.

2 Methodology

This section introduces the overall methodology of Hy-MT2. Designed for multilingual machine translation, Hy-MT2 follows a staged pipeline consisting of **MT-oriented Mid-training** (Section 2.1), **Family-Centric Post-training** (FCPT; Section 2.2), and model **Quantization** (Section 2.3). Specifically, we start from a general Hy-series Pretraining Model and perform MT-oriented Mid-training to obtain a unified model with fundamental translation capabilities. The model is then further optimized through FCPT. As shown in Figure 2, FCPT consists of three key processes: **Reference-Guided On Policy Distillation** (RG-OPD; Section 2.2.1), **Family-specific RL Training** (Section 2.2.2), and **Cross-family On Policy Distillation** (Cross-family OPD; Section 2.2.3). The first two processes organize training around language families and construct multiple family-specific strong teachers; Cross-family OPD then transfers their capabilities into a unified student model while incorporating general instruction-following data to preserve the model’s instruction-following ability beyond translation.

2.1 MT-oriented Mid-training

In the **MT-oriented Mid-training** stage, we start from the Hy-series Pretraining Model and continue training it on approximately 1T tokens of large-scale multilingual translation-related data. This stage

Table 1: Instruction examples for Hy-MT2 translation tasks in Chinese and English.

Type	Chinese prompt	English prompt
Default Translation	<p>将以下文本翻译为{target_lang}，注意只需要输出翻译后的结果，不要额外解释：</p> <p>{source_text}</p>	<p>Translate the following text into {target_lang}. Note that you should only output the translated result without any additional explanation:</p> <p>{source_text}</p>
Terminology	<p>参考下面的翻译： {text} 翻译成 {text} {text} 翻译成 {text} {text} 翻译成 {text}</p> <p>将以下文本翻译为{target_lang}，注意只需要输出翻译后的结果，不要额外解释：</p> <p>{source_text}</p>	<p>Reference the following translations: {text} translates to {text} {text} translates to {text} {text} translates to {text}</p> <p>Translate the following text into {target_lang}. Note that you must ONLY output the translated result without any additional explanation:</p> <p>{source_text}</p>
Style	<p>请将以下文本翻译为{{target_lang}}。 注意翻译的风格要严格符合【{{target_style}}】</p> <p>{{source_text}}</p>	<p>Please translate the following text into {{target_lang}}. Note that the translation style must strictly conform to [{{target_style}}]:</p> <p>{{source_text}}</p>
Personalization	<p>【待翻译文本】 {source_text}</p> <p>【翻译任务】 1、{user_preferences} 2、{user_preferences} 3、..... 4、将【待翻译文本】翻译为{target_lang}。</p>	<p>[Source Text] {source_text}</p> <p>[Translation Tasks] 1. {user_preferences} 2. {user_preferences} 3. ... 4. Translate the [Source Text] into {target_lang}.</p>
Delimiters	<p>请将以下文本准确翻译为{{target_lang}}。 你必须在译文中保留等量的分隔符，绝对不可遗漏、转义或翻译该符号，并注意分隔符的位置。</p> <p>{{source_text}}</p>	<p>Please accurately translate the following text into {{target_lang}}. You must retain the exact same number of delimiters in the translation. Strictly do not omit, escape, or translate these symbols, and pay close attention to their placement.</p> <p>{{source_text}}</p>
Structured Data 1	<p># 任务目标 将下方{{source_text}}中的{{format.type}}格式数据翻译为{{target_lang}}。</p> <p># 严格约束 1. **结构锁定**: 绝对保持原有的{{format.type}}数据结构、缩进和层级完全不变。 2. **选择性翻译**: 仅翻译面向用户展示的可见文本内容。 3. **禁止修改**: **严禁**翻译或更改任何代码标签、键名(Key)、变量占位符(如'{{var}}'、'\${var}'、'%s'、'%d'等)或代码属性。</p> <p># 数据输入 {{source_text}}</p>	<p>### Task Translate the user-facing text within the following {{format.type}} data into {{target_lang}}.</p> <p>### Strict Rules 1. **Structure Preservation**: You MUST preserve the original {{format.type}} data structure, nesting, hierarchy, and indentation exactly as they are. 2. **Selective Translation**: Translate ONLY the visible, user-facing text content/values. 3. **Strict Non-Translation**: NEVER translate or alter code tags, keys, properties, object names, or variable placeholders. Leave them exactly in their original English/code form.</p> <p>### Source Data {{source_text}}</p>
Structured Data 2	<p>【背景信息】 {{background_text}}</p> <p>请结合背景信息将以下文本翻译为{{target_lang}}。</p> <p>【待翻译文本】 {source_text}</p>	<p>[Background Information] {{background_text}}</p> <p>Please translate the following text into {{target_lang}}, taking the provided background information into consideration.</p> <p>[Source Text] {source_text}</p>

Notes. This table shows representative instruction templates in Chinese and English. Additional multilingual instruction examples are provided in Section 3.6.

aims to strengthen the model’s translation capability and provide a unified foundation for the subsequent Family-Centric Post-training.

Specifically, the training data is organized along two dimensions:

- **Data format**: We use both multilingual monolingual corpus and parallel translation corpus to help the model capture linguistic characteristics across different languages and strengthen cross-lingual semantic mapping and source-target alignment.
- **Scenario coverage**: The data covers general translation, domain-specific translation, real-world scenarios, and instruction-following examples, improving translation quality, domain adaptation,

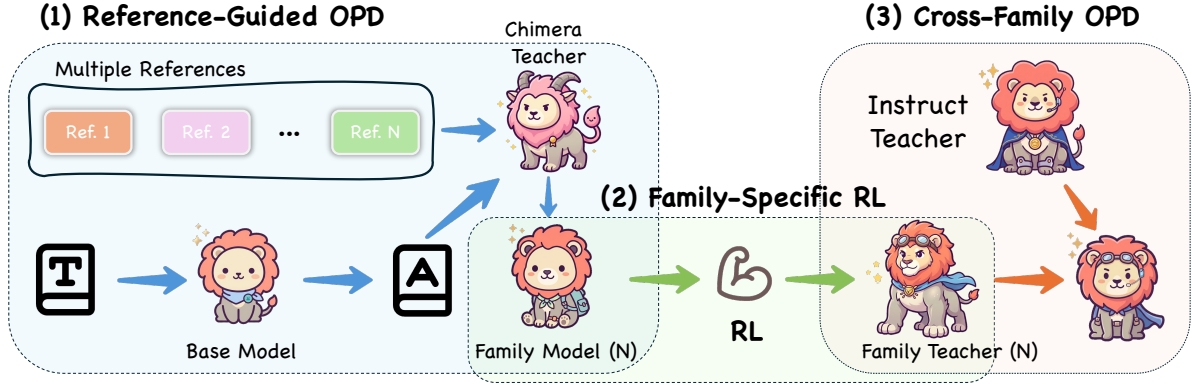


Figure 2: Family-Centric Post-training pipeline of Hy-MT2.

practical translation robustness, and the ability to follow translation-related instructions.

The output of this stage is an **MT-oriented Mid-trained Model**, which serves as the unified starting point for FCPT.

2.2 Family-Centric Post-training

Instead of directly mixing data from all language families, FCPT divides training into multiple family branches, covering diverse language groups, e.g., Western European, East Asian, and Middle Eastern right-to-left languages. Within each branch, we incorporate general translation data, domain-specific translation data, real-world business scenario data, and translation instruction-following data to construct a family-specific teacher. This family-centric design allows each teacher to learn under a more consistent language distribution, reducing interference across different language families.

2.2.1 Reference-Guided On Policy Distillation

Reference-Guided On Policy Distillation is the first stage of FCPT. In this stage, we perform On policy distillation separately on each family branch, aiming to obtain a family-specific translation policy that better captures the linguistic characteristics and translation preferences of the corresponding language family. The resulting model further serves as a stronger initialization for subsequent Family-specific RL Training.

The core of RG-OPD is the construction of a stronger Chimera Teacher. Unlike conventional distillation methods that rely on a single teacher model, Chimera¹. Teacher does not require training an additional large-scale translation-specialized teacher model. For each source sentence, it integrates candidate translations generated by multiple Hy-series reference models with the original dataset label. Although not all labels are manually annotated, they still serve as useful reference signals. By fusing these multi-source references, Chimera Teacher provides richer scoring signals, helps introduce greater diversity into the distillation process, and constructs a stronger supervision signal for On policy distillation.

Specifically, given a source sentence x and its reference set $\mathcal{R}(x)$, where $\mathcal{R}(x)$ consists of multiple candidate reference sources, the student model in the current family branch first generates a translation y based on its current policy π_θ . Chimera Teacher then evaluates the student output based on the multi-source reference set $\mathcal{R}(x)$, and produces a teacher policy or target distribution $\pi_T(\cdot | x, \mathcal{R}(x))$. The student model is optimized by minimizing the forward KL divergence from the teacher policy to the student policy. The training objective can be written as:

$$\mathcal{L}_{\text{RG-OPD}} = D_{\text{KL}}(\pi_T(\cdot | x, \mathcal{R}(x)) \| \pi_\theta(\cdot | x)). \quad (1)$$

Here, π_T denotes the distillation target distribution constructed by Chimera Teacher, and π_θ denotes the output policy of the current student model. We adopt forward KL divergence as the distillation objective, enabling the student model to learn the fused translation preference from Chimera Teacher in an online manner and gradually improve its translation policy for the corresponding language family.

¹The name *Chimera* is inspired by the mythological creature composed of multiple animals. In our setting, it refers to a teacher signal constructed by fusing multiple reference sources. We implement Chimera Teacher based on Hy3-Preview.

After **RG-OPD**, each family branch obtains a family-specific student model that has been adapted to the translation preferences and expression patterns of the corresponding language family. These models are then used as the initialization for subsequent **Family-specific RL Training**.

2.2.2 Family-specific RL Training

In **Family-specific RL Training**, each family branch is further optimized through Group Relative Policy Optimization (GRPO) (Shao et al., 2024), using the model obtained from RG-OPD as initialization. To provide more fine-grained and rigorous reward signals, we introduce a hybrid evaluation system combining a rule-based pre-filter with an LLM-based Multidimensional Quality Metrics (MQM) judge (Lommel et al., 2014; Freitag et al., 2021).

Rule-based Pre-filtering Before passing translations to the LLM evaluator, a rule-based filter is applied to intercept critical text degradation. Translations exhibiting severe repetition or mixed languages are immediately assigned a reward of 0. This ensures early penalization of degenerated outputs and avoids unnecessary LLM computation.

LLM Judge Evaluation System For translations that pass the pre-filter, the LLM-based judge evaluates the text based on a 5-dimensional error typology rather than assigning a holistic score. The dimensions are:

- **Terminology:** Identifies terminology errors, inconsistencies with terminological resources, or inconsistent usage throughout the text.
- **Accuracy:** Detects mistranslations, over-translation, under-translation, added or omitted content, unwarranted translations, missed translations, and instances of mixed languages.
- **Linguistic Conventions:** Checks for grammatical errors, punctuation errors, spelling mistakes, unintelligibility, discourse convention errors, and locale convention violations.
- **Style:** Assesses inconsistencies with external references, incorrect language register, obscure expressions, unnatural phrasing, and stylistic inconsistencies.
- **Instruction Following:** Evaluates adherence to task constraints, flagging wrong languages, unexecuted translation tasks, and failures to follow terminology, formatting, style, or context guidelines.

Scoring Rules The evaluation starts from a base score of 100. The LLM judge identifies translation errors and applies deductions according to their severity. Fatal errors, such as using the wrong language or failing to execute the translation task, directly result in an overall score of 0. Major errors incur a deduction of 10–20 points per instance, while minor errors incur a deduction of 2–5 points per instance. The overall score S_{overall} is obtained by subtracting all error deductions from the base score, with a lower bound of 0.

Length Penalty and Final Reward Calculation To prevent the model from exploiting the reward system by generating pathologically short, truncated, or excessively long and redundant sentences, we introduce a length penalty. Given a source sentence x , a ground-truth translation of length L_{gt} , and a model-generated translation y of length L_y , the length penalty P_{len} is computed as:

$$P_{\text{len}} = \min \left(0.5 \times \frac{|L_{gt} - L_y|}{L_{gt}}, 0.5 \right). \quad (2)$$

The final reward $r(x, y)$ is calculated by normalizing the overall MQM score to the range $[0, 1]$, subtracting the length penalty P_{len} , and clipping the result at 0.

After this stage, each family branch produces a strong expert, which is then used as a strong teacher in the subsequent **Cross-family OPD** stage.

2.2.3 Cross-family On Policy Distillation

Cross-family On Policy Distillation is the final training process of FCPT, aiming to transfer the language-family-specific translation capabilities learned by multiple family-specific strong teachers into a unified student model. Meanwhile, to improve the model’s instruction-following ability, we introduce general instruction-following data in this stage and use the Hy Instruct model as the corresponding instruction teacher to provide distillation signals. Therefore, Cross-family OPD can be viewed as a unified multi-teacher distillation process.

Specifically, given an input sample x , we denote its teacher policy as $\pi_T(\cdot | x)$ and the output policy of the unified student model as $\pi_\theta(\cdot | x)$. In this stage, we adopt reverse KL divergence as the distillation

objective, which is defined as:

$$\mathcal{L}_{\text{Cross-OPD}} = D_{\text{KL}}\left(\pi_{\theta}(\cdot | x) \parallel \pi_{T_{\tau(x)}}(\cdot | x)\right), \quad (3)$$

where $\pi_{\theta}(\cdot | x)$ denotes the output policy of the student model initialized from MT-oriented Mid-training, and $\pi_{T_{\tau(x)}}(\cdot | x)$ denotes the output policy of the selected teacher. For translation samples, $\tau(x)$ selects the corresponding family-specific strong teacher according to the language family of x ; for general instruction-following samples, $\tau(x)$ selects the Hy Instruct teacher.

After **Cross-family On Policy Distillation**, we obtain the final **Hy-MT2** series models. To facilitate efficient deployment, these models are further processed in the quantization stage.

2.3 Quantization

To accommodate deployment requirements under varying resource constraints, we perform model quantization on the obtained Hy-MT2 model series, offering a diverse suite of precision variants including FP16, 8-bit, 4-bit, 2-bit, and 1.25-bit.

For the 8-bit and 4-bit variants, we predominantly adopt a post-training quantization (PTQ) pipeline. Without retraining the model, this approach estimates the distribution of model weights or activations using a small set of calibration data, thereby reducing storage and computational overheads. Specifically, 8-bit quantization employs a higher-precision low-bit representation to minimize performance degradation, whereas 4-bit quantization further compresses weight representations and mitigates quantization errors through dedicated calibration strategies.

For the 2-bit version, we employ the ultra-low-bit quantization-aware training (QAT) scheme from AngelSlim Team (2026) framework. Compared with PTQ, 2-bit quantization imposes more stringent representational constraints. Consequently, it requires explicit simulation of low-bit quantization behaviors during the training process, allowing model weights to progressively adapt to low-precision representations. The 2-bit scheme utilizes Stretched Elastic Quantization (SEQ) Liu et al. (2026), quantizing weight into $-1.5, -0.5, +0.5, +1.5$. By optimizing the quantization mapping and scaling factors, this approach enhances the model’s stability and capability for performance recovery under 2-bit constraints.

For extreme compression scenarios, we further implement Sherry Huang et al. (2026), a 1.25-bit sparse ternary quantization method. Sherry quantizes model weights into a ternary space of $\{-1, 0, +1\}$ and introduces a 3:4 fine-grained sparsity pattern, where constraining every 4-weight block to contain exactly one zero and three sign-only weights. This structure enables packing four weights into 5 bits, thereby achieving a regularized 1.25-bit representation while maintaining a hardware alignment that is more favorable for Single Instruction Multiple Data (SIMD) computing patterns. Compared to conventional 2-bit packing or 1.67-bit irregular packing, Sherry achieves a superior balance between compression ratio and inference efficiency.

For 1.25-bit and 2-bit QAT, we adopt a distillation-based training strategy. Specifically, the low-precision model serves as the student, while the high-precision model acts as the teacher. The overall objective combines the standard language modeling (LM) loss with a KL-divergence-based distillation loss. For the KL component, inspired by Du et al. (2024), we incorporate both forward and backward KL divergences. Unlike prior approaches that use a fixed, manually selected weighting coefficient, we compute the teacher model’s confidence score for each token and dynamically adjust the relative weights of the forward and backward KL terms accordingly.

Ultimately, we obtain the quantized Hy-MT2 model series spanning FP16, 8-bit, 4-bit, 2-bit, and 1.25-bit precisions. Armed with these diverse precision variants, Hy-MT2 can be flexibly deployed across a wide range of scenarios, including high-accuracy service serving, low-resource device inference, and extreme edge-side compression.

3 Experiments

3.1 Benchmarks

To comprehensively evaluate the translation capabilities of Hy-MT2, we construct an evaluation suite from four perspectives: general translation, real-world business scenario translation, domain-specific translation, and translation instruction following.

General translation evaluation. We use FLORES-200 (Team et al., 2022), WMT25 (Kocmi et al., 2025), and the Mandarin \leftrightarrow Minority Testset to evaluate general translation capabilities. FLORES-200 covers 1,056 translation directions across 33 languages. WMT25 adopts the human evaluation sets from WMT25

and covers 12 translation directions. The Mandarin \leftrightarrow Minority Testset focuses on bidirectional translation between Mandarin Chinese and minority languages.

Real-world business scenario evaluation. We construct WildMTBench to assess model performance on practical business inputs. The dataset covers six types of scenarios, including webpages, meetings, books, social content, news, and documents, with 2,000 samples in total. It focuses on evaluating model robustness and adaptability to diverse text forms, real-world input distributions, and complex application requirements.

Domain-specific translation evaluation. We construct DomainMTBench to assess translation quality in professional domains. The dataset covers six domains: finance, law, politics, technology, medicine, and education. The data are collected from open-source corpora and processed through cleaning, filtering, domain classification, and human translation annotation, resulting in 24,000 samples. This benchmark focuses on evaluating the model’s ability to handle domain terminology, specialized expressions, and established industry translations.

Instruction-following evaluation. We construct IFMTBench to evaluate translation-specific instruction following in multilingual settings. It contains 7,344 high-quality human-aligned samples with instructions in Chinese, German, Japanese, French, English, Spanish, and Korean, covering industrial translation constraints such as terminology, format, and style. It includes 4,506 single-constraint and 2,838 multi-constraint samples, evaluating both basic constraint execution and robustness under complex instruction combinations. In addition, we use IFBench (Pyatkin et al., 2025), IFEval (Zhou et al., 2023), MaXIFE (Liu et al., 2025), and Multi-IF (He et al., 2024) to assess general instruction-following capabilities. IFEval is a verifiable instruction-following benchmark for large language models, containing around 500 prompts. IFBench focuses on model generalization to diverse and unseen verifiable constraints. MaXIFE evaluates multilingual and cross-lingual instruction following across 23 languages. Multi-IF focuses on instruction following in multi-turn and multilingual interactions.

3.2 General Translation Evaluation

In general translation evaluation, we use three metrics: XCOMET-XXL, CometKiwi, and GEMBA. XCOMET-XXL is a reference-based automatic evaluation metric, CometKiwi is a reference-free evaluation metric, and GEMBA is an LLM-based evaluation metric. The results are reported in Table 2.

Overall, Hy-MT2 achieves substantial improvements over Hy-MT1.5 across general translation benchmarks. On the XX \leftrightarrow XX setting of FLORES-200, Hy-MT2-1.8B, Hy-MT2-7B, and Hy-MT2-30B-A3B reach 79.77, 86.89, and 87.47, corresponding to 89.9%, 97.9%, and 98.6% of Gemini 3.1 Pro^T, respectively. In particular, Hy-MT2-7B and Hy-MT2-30B-A3B outperform strong baselines such as DeepSeek-V4-Pro, Kimi K2.6, Qwen3.5-397B-A17B, and Gemma4-26B-A4B in this setting. Compared with Hy-MT1.5-7B, Hy-MT2-7B improves the XCOMET-XXL score from 80.98 to 86.89, showing a clear gain in overall multilingual translation performance.

On WMT25, Hy-MT2-7B and Hy-MT2-30B-A3B also show strong performance. Hy-MT2-7B achieves 63.86 / 71.21 / 82.24, while Hy-MT2-30B-A3B achieves 62.89 / 71.08 / 84.34. Compared with Hy-MT1.5-7B, whose scores are 61.59 / 68.85 / 75.91, Hy-MT2-7B improves on all three metrics, with a particularly large gain on GEMBA. Hy-MT2-30B-A3B further achieves the best GEMBA score among all compared systems, surpassing Gemini 3.1 Pro^T and GPT-5.5, indicating stronger overall translation quality and readability in challenging WMT settings.

On Mandarin \leftrightarrow Minority translation, Hy-MT2-7B and Hy-MT2-30B-A3B achieve XCOMET-XXL scores of 62.05 and 62.44, respectively, outperforming both Gemini 3.1 Pro and Hy-MT1.5-7B. This suggests that Hy-MT2 preserves strong performance in Mandarin-minority language translation and further improves translation quality in low-resource language scenarios.

For the lightweight setting, Hy-MT2-1.8B shows consistent improvements over Hy-MT1.5-1.8B and remains highly competitive against larger open-source models and commercial translation systems. Despite its small size, it outperforms Tower-Plus-72B and achieves competitive results against Microsoft Translator and Doubao Translator. On WMT25, Hy-MT2-1.8B surpasses both commercial systems across all three metrics, demonstrating a strong efficiency-quality trade-off.

3.3 Domain-Specific and In-the-Wild Translation Evaluation

Domain-specific translation. In domain-specific translation evaluation, we use DomainMTBench to assess model performance across multiple professional domains. We report XCOMET and GEMBA scores, where XCOMET measures reference-based translation quality and GEMBA provides an LLM-based assessment of overall translation quality. The results are shown in Table 3.

Table 2: Performance comparison on general translation benchmarks.

Models	FLORES-200			WMT25	Mand. \Leftrightarrow Min.
	ZH \Leftrightarrow XX	EN \Leftrightarrow XX	XX \Leftrightarrow XX		
Gemini 3.1 Pro [†]	90.30 / 78.96 / 92.14	94.42 / 88.38 / 92.68	88.74 / 77.60 / 90.97	57.58 / 69.06 / 82.23	61.11 / 53.50 / 79.70
GPT-5.5 [†]	89.94 / 78.95 / 91.98	94.16 / 88.33 / 92.76	88.44 / 77.36 / 90.93	56.68 / 69.31 / 83.29	60.87 / 55.18 / 79.68
GPT-5.5	89.60 / 78.96 / 91.65	93.98 / 88.17 / 92.54	87.92 / 77.20 / 90.37	56.41 / 69.15 / 83.41	59.81 / 53.32 / 75.16
DeepSeek-V4-Pro [†]	88.86 / 78.13 / 90.97	93.22 / 87.83 / 91.73	86.81 / 76.66 / 89.86	54.83 / 68.25 / 81.99	56.31 / 51.24 / 69.48
DeepSeek-V4-Pro	88.60 / 78.11 / 90.15	93.02 / 87.71 / 91.39	83.29 / 67.64 / 75.46	53.97 / 67.58 / 79.09	55.61 / 51.47 / 66.02
Kimi K2.6 [†]	88.49 / 77.97 / 90.84	92.95 / 87.72 / 91.28	86.17 / 76.72 / 89.44	54.48 / 68.64 / 81.68	55.98 / 54.76 / 68.73
Kimi K2.6	87.68 / 77.55 / 89.85	91.96 / 86.95 / 90.31	84.05 / 75.84 / 87.43	49.76 / 66.08 / 76.84	54.29 / 52.52 / 64.00
GLM5.1 [†]	88.78 / 78.28 / 91.38	93.26 / 87.82 / 91.84	86.57 / 76.84 / 89.92	54.23 / 68.39 / 80.71	57.51 / 54.46 / 71.72
GLM5.1	87.09 / 77.60 / 89.44	91.59 / 86.73 / 89.92	83.83 / 75.40 / 87.14	49.71 / 65.41 / 73.70	56.00 / 50.68 / 67.70
Qwen3.5-397B-A17B [†]	87.79 / 77.75 / 90.58	92.38 / 87.19 / 90.65	86.06 / 76.52 / 89.28	54.95 / 68.63 / 81.37	55.97 / 51.89 / 70.21
Qwen3.5-397B-A17B	88.50 / 78.64 / 90.66	93.07 / 87.73 / 91.49	86.29 / 76.39 / 88.87	55.79 / 68.80 / 83.14	55.59 / 54.44 / 67.45
Qwen3.6-35B-A3B [†]	87.71 / 77.93 / 90.32	92.32 / 87.55 / 90.49	84.84 / 75.95 / 88.21	50.75 / 65.24 / 73.42	54.07 / 53.37 / 64.44
Gemma4-31B [†]	89.30 / 78.86 / 91.16	93.79 / 88.27 / 91.90	87.77 / 77.28 / 90.11	54.48 / 67.12 / 79.59	51.99 / 51.49 / 64.17
Gemma4-26B-A4B [†]	88.68 / 78.68 / 90.74	93.31 / 88.09 / 91.53	86.80 / 77.14 / 89.58	52.13 / 65.40 / 77.73	47.83 / 49.19 / 56.73
Gemma4-E4B [†]	83.49 / 75.94 / 83.96	89.97 / 85.90 / 87.67	81.29 / 74.10 / 83.74	39.70 / 56.48 / 62.34	41.16 / 40.64 / 42.25
Gemma4-E2B [†]	83.21 / 76.18 / 84.21	88.94 / 85.38 / 86.39	79.69 / 73.78 / 82.54	37.07 / 53.51 / 57.21	40.44 / 40.39 / 42.67
Qwen3.6-35B-A3B	86.32 / 78.00 / 87.34	90.87 / 86.52 / 88.77	82.11 / 74.81 / 84.92	51.11 / 66.70 / 75.28	52.12 / 51.74 / 58.28
Gemma4-31B	88.40 / 78.68 / 89.89	93.30 / 87.96 / 91.03	86.84 / 76.16 / 88.67	52.49 / 65.66 / 75.75	48.98 / 50.05 / 56.75
Gemma4-26B-A4B	87.61 / 78.31 / 88.97	92.80 / 87.71 / 90.65	85.60 / 76.58 / 87.97	49.62 / 64.22 / 74.36	44.52 / 46.14 / 47.70
Gemma4-E4B	83.36 / 75.92 / 83.52	89.54 / 85.61 / 86.95	79.96 / 72.91 / 80.97	38.36 / 55.45 / 59.95	41.17 / 40.23 / 43.19
Gemma4-E2B	79.78 / 73.19 / 76.46	87.20 / 84.00 / 83.63	75.97 / 71.00 / 76.64	35.60 / 51.83 / 54.35	39.07 / 33.83 / 34.62
Tower-Plus-72B	79.69 / 71.82 / 77.86	84.16 / 81.13 / 78.82	70.02 / 65.53 / 67.85	41.00 / 55.54 / 54.85	38.55 / 35.40 / 26.70
translategemma-27b-it	- / - / -	- / - / -	- / - / -	58.02 / 66.08 / 81.24	- / - / -
Microsoft-Translator	82.95 / 72.85 / 84.79	89.31 / 85.48 / 87.34	80.13 / 72.42 / 83.85	42.01 / 60.21 / 67.63	51.80 / 41.77 / 55.27
Doubao-Translator	80.92 / 71.57 / 82.14	86.77 / 83.49 / 84.46	76.54 / 69.92 / 79.72	33.13 / 54.52 / 49.75	53.11* / 40.16* / 66.81*
iFLYTEK-Translator	83.00 / 73.96 / -	87.18 / 83.65 / -	76.53 / 69.90 / -	35.10 / 56.15 / -	59.04 / 44.67 / -
Hy-MT1.5-1.8B	83.61 / 76.55 / 76.79	89.42 / 84.11 / 81.35	78.40 / 71.82 / 75.12	53.08 / 61.95 / 63.58	58.06 / 40.84 / 61.66
Hy-MT1.5-7B	86.90 / 79.24 / 84.35	90.93 / 86.50 / 85.66	80.98 / 73.36 / 78.30	61.59 / 68.85 / 75.91	61.74 / 44.55 / 67.26
Hy-MT2-1.8B	84.26 / 76.43 / 82.71	90.00 / 85.46 / 84.05	79.77 / 73.41 / 78.64	50.30 / 64.59 / 70.36	58.29 / 39.22 / 62.36
Hy-MT2-7B	89.45 / 78.97 / 88.89	93.52 / 88.07 / 89.34	86.89 / 76.03 / 87.23	63.86 / 71.21 / 82.24	62.05 / 43.60 / 68.93
Hy-MT2-30B-A3B	89.83 / 79.03 / 90.26	93.85 / 88.23 / 90.89	87.47 / 76.34 / 88.79	62.89 / 71.08 / 84.34	62.44 / 42.37 / 69.43

Notes. Each cell reports XCOMET-XXL / CometKiwi / GEMBA scores, and all scores are multiplied by 100. [†] denotes thinking mode. In FLORES-200, XX \Leftrightarrow XX denotes the average performance over all evaluated translation directions, including ZH \Leftrightarrow XX and EN \Leftrightarrow XX. Mand. \Leftrightarrow Min. denotes Mandarin \Leftrightarrow Minority translation. Values marked with * are computed only on supported language pairs. Values replaced by - indicate that the model does not support the corresponding test set. Baselines are grouped into large-scale models and all Think-mode models, medium to small-sized general models in non-Think mode, and translation-specialized models. Our models are shown in bold. The best results among large-scale models and all Think-mode models in each column are highlighted in blue background, while the best results among non-thinking-mode, small-to-medium-sized models are highlighted in light orange background.

Overall, Hy-MT2 shows strong performance on DomainMTBench and consistently improves over Hy-MT1.5 in GEMBA. On the average score, Hy-MT2-1.8B improves the GEMBA score from 88.82 to 91.08 compared with Hy-MT1.5-1.8B, while Hy-MT2-7B improves from 92.04 to 92.79 compared with Hy-MT1.5-7B. Hy-MT2-30B-A3B further achieves 95.04 / 93.73 on the average score, obtaining the best XCOMET result among all compared systems and the best GEMBA result among open-source and translation-specialized models. Across individual domains, Hy-MT2-30B-A3B achieves the best XCOMET scores in finance, law, and politics, with 97.08, 89.15, and 96.63, respectively. It also remains highly competitive in medical, technology, and education. These results indicate that the larger MoE model effectively strengthens domain-specific translation ability while maintaining strong performance across diverse professional domains. Hy-MT2-7B also demonstrates strong domain translation quality, improving over Hy-MT1.5-7B on the average score from 94.82 / 92.04 to 94.92 / 92.79, with clear gains on GEMBA across multiple domains. For the lightweight model, Hy-MT2-1.8B achieves a notable GEMBA improvement over Hy-MT1.5-1.8B, increasing the average score from 88.82 to 91.08. It also surpasses commercial systems such as Microsoft Translator and Doubao Translator on the average score, especially in GEMBA.

In-the-wild translation. On WildMTBench, Hy-MT2 also shows clear advantages in real-world translation scenarios. Hy-MT2-7B achieves 90.28 / 88.93, outperforming Hy-MT1.5-7B on both XCOMET and GEMBA. Hy-MT2-30B-A3B further reaches 89.87 / 89.25, obtaining the best GEMBA score among all compared systems and surpassing Gemini 3.1 Pro in LLM-based evaluation. For the lightweight setting, Hy-MT2-1.8B improves substantially over Hy-MT1.5-1.8B in GEMBA, increasing from 80.84 to 86.04, while maintaining a similar XCOMET score. It also clearly outperforms commercial translation systems such as Microsoft Translator and Doubao Translator on WildMTBench.

Overall, these results show that Hy-MT2 improves not only standard domain-specific translation, but also robustness and usability in real-world business scenarios. The consistent gains over Hy-MT1.5, especially in GEMBA, suggest that Hy-MT2 produces more natural and reliable translations under both

Table 3: Performance comparison on DomainMTBench and WildMTBench.

Models	DomainMTBench								WildMTBench	
	Finance	Law	Medical	Technology	Politics	Education	Avg.			
Gemini 3.1 Pro ^T	96.30 / 94.73	88.48 / 93.25	95.64 / 95.44	94.94 / 95.23	96.16 / 94.58	97.24 / 95.11	94.50 / 94.64	86.62 / 88.96		
GPT-5.5 ^T	96.07 / 94.69	88.13 / 93.08	95.40 / 95.35	94.58 / 94.89	95.93 / 94.47	97.06 / 95.08	94.23 / 94.51	86.66 / 88.72		
GPT-5.5	95.93 / 94.56	88.26 / 92.48	95.37 / 95.24	94.49 / 94.78	95.83 / 94.31	97.01 / 94.81	94.20 / 94.28	86.51 / 88.00		
DeepSeek-V4-Pro ^T	95.76 / 94.47	87.65 / 92.22	95.33 / 95.21	94.40 / 94.71	95.55 / 94.02	96.96 / 94.77	93.96 / 94.13	85.77 / 88.23		
DeepSeek-V4-Pro	95.81 / 94.20	87.55 / 91.76	95.42 / 95.08	94.33 / 94.26	95.61 / 93.78	96.72 / 94.16	93.96 / 93.80	86.05 / 85.81		
Kimi K2.6 ^T	95.91 / 94.31	87.81 / 92.51	95.34 / 94.96	94.42 / 94.37	95.68 / 93.96	96.86 / 94.36	94.04 / 94.01	86.39 / 87.91		
Kimi K2.6	95.30 / 93.74	87.30 / 91.07	94.95 / 94.42	93.98 / 93.68	95.18 / 93.43	96.66 / 93.75	93.58 / 93.27	85.88 / 87.09		
GLM5.1 ^T	95.77 / 94.55	87.72 / 92.54	95.08 / 95.21	94.40 / 94.87	95.60 / 94.29	96.82 / 94.86	93.92 / 94.30	86.10 / 88.29		
GLM5.1	95.31 / 94.06	86.84 / 91.76	94.95 / 94.85	93.96 / 94.18	95.22 / 93.71	96.54 / 94.38	93.49 / 93.72	85.47 / 86.83		
Qwen3.5-397B-A17B ^T	94.21 / 92.76	86.11 / 91.01	93.12 / 92.86	92.34 / 92.41	94.11 / 92.48	94.07 / 91.85	92.10 / 92.20	82.56 / 83.54		
Qwen3.5-397B-A17B	95.88 / 94.32	87.59 / 91.63	95.24 / 94.94	94.38 / 94.35	95.64 / 93.80	97.13 / 94.60	93.98 / 93.82	86.97 / 87.56		
Qwen3.6-35B-A3B ^T	96.03 / 94.31	87.58 / 91.66	95.35 / 94.87	94.53 / 94.23	95.76 / 93.76	96.91 / 94.33	94.06 / 93.77	87.24 / 87.75		
Gemma4-31B ^T	96.07 / 94.32	87.66 / 92.13	95.37 / 94.89	94.15 / 94.27	95.89 / 93.80	97.12 / 94.55	94.07 / 93.90	86.87 / 87.33		
Gemma4-26B-A4B ^T	95.78 / 94.30	87.44 / 91.81	95.19 / 94.83	94.11 / 94.13	95.50 / 93.68	96.92 / 94.29	93.84 / 93.76	86.30 / 86.03		
Gemma4-E4B ^T	93.66 / 91.60	83.79 / 87.03	93.61 / 92.19	92.27 / 91.02	93.10 / 90.43	95.19 / 91.86	91.57 / 90.53	84.11 / 84.64		
Gemma4-E2B ^T	92.77 / 91.25	82.69 / 86.77	93.12 / 91.67	91.94 / 90.56	92.41 / 89.77	94.94 / 91.25	90.90 / 90.05	83.12 / 82.85		
Qwen3.6-35B-A3B	95.52 / 93.53	87.00 / 90.13	94.88 / 94.26	94.13 / 93.55	95.29 / 93.27	96.57 / 93.57	93.59 / 92.96	86.49 / 86.63		
Gemma4-31B	95.56 / 93.81	87.10 / 91.28	95.12 / 94.44	94.02 / 93.97	95.28 / 93.14	96.77 / 94.17	93.65 / 93.34	86.35 / 86.55		
Gemma4-26B-A4B	95.39 / 93.65	86.63 / 90.35	95.02 / 94.25	93.69 / 93.44	94.97 / 93.06	96.69 / 94.03	93.39 / 93.01	86.26 / 86.69		
Gemma4-E4B	93.45 / 91.42	83.46 / 86.46	93.51 / 92.10	91.95 / 90.40	93.13 / 90.48	95.13 / 91.52	91.41 / 90.27	83.45 / 84.18		
Gemma4-E2B	92.07 / 89.59	81.52 / 84.51	92.53 / 89.73	91.03 / 88.26	91.61 / 87.82	94.30 / 90.45	90.09 / 88.11	82.01 / 81.58		
Tower-Plus-72B	95.23 / 93.60	86.54 / 90.80	95.04 / 94.35	93.91 / 93.62	95.41 / 93.29	96.53 / 93.83	93.46 / 93.14	85.13 / 86.66		
TranslateGemma-27B-IT	96.33 / 92.38	87.56 / 88.88	95.81 / 93.84	94.66 / 92.29	96.00 / 92.01	97.26 / 93.20	94.30 / 91.94	88.44 / 85.65		
Microsoft-Translator	92.51 / 90.34	81.76 / 85.17	92.47 / 90.64	91.19 / 87.96	92.77 / 90.25	94.96 / 90.88	90.49 / 89.01	79.19 / 79.32		
Doubao-Translator	93.39 / 91.80	83.18 / 87.56	91.63 / 90.53	91.78 / 89.64	93.77 / 91.97	94.69 / 91.38	91.02 / 90.36	78.07 / 77.61		
Hy-MT1.5-1.8B	95.55 / 90.01	86.03 / 84.90	95.26 / 90.27	94.37 / 88.59	95.23 / 89.59	96.61 / 90.89	93.52 / 88.82	87.41 / 80.84		
Hy-MT1.5-7B	96.66 / 92.82	88.37 / 89.52	96.28 / 93.40	95.24 / 91.52	96.51 / 92.55	97.57 / 93.09	94.82 / 92.04	90.21 / 87.13		
Hy-MT2-1.8B	95.36 / 91.69	86.63 / 88.89	94.97 / 92.10	93.97 / 91.25	95.01 / 91.18	96.42 / 92.46	93.41 / 91.08	87.43 / 86.04		
Hy-MT2-7B	96.79 / 93.14	89.02 / 91.14	96.23 / 93.76	95.15 / 92.80	96.47 / 93.02	97.50 / 93.43	94.92 / 92.79	90.28 / 88.93		
Hy-MT2-30B-A3B	97.08 / 94.14	89.15 / 92.04	96.22 / 94.63	95.16 / 94.08	96.63 / 93.79	97.54 / 94.23	95.04 / 93.73	89.87 / 89.25		

Notes. Each cell reports XCOMET / GEMBA scores, and all scores are multiplied by 100. Avg. denotes the overall average score on DomainMTBench. ^T denotes thinking mode; for models with both modes, the row without ^T denotes non-thinking mode. Baselines are grouped into **large-scale models** and **all Think-mode models**, **medium to small-sized general models in non-Think mode**, and **translation-specialized models**. Our models are shown in bold. The best results among large-scale models and all Think-mode models in each column are highlighted in **blue background**, while the best results among non-thinking-mode, small-to-medium-sized models are highlighted in **light orange background**.

professional-domain and in-the-wild settings.

3.4 Instruction-Following Evaluation

We evaluate instruction-following ability on both general instruction-following benchmarks and our translation-specific IFMTBench. The results are reported in Table 4.

Hy-MT2 shows strong translation-specific instruction-following ability on IFMTBench. Hy-MT2-7B achieves 89.73 / 72.67 / 83.14 on Simple, Complex, and Total, respectively, outperforming Gemma4-E4B and other medium-sized open-source baselines. Hy-MT2-30B-A3B further improves to 90.20 / 75.94 / 84.69, achieving the best overall IFMTBench score among small-to-medium-sized models. Compared with Qwen3.6-35B-A3B and Gemma4-26B-A4B, Hy-MT2-30B-A3B obtains consistent gains, especially on Complex instructions, showing stronger capability in handling multi-constraint translation requests.

The performance of Hy-MT2 is also competitive with much larger general-purpose models. On IFMTBench, Hy-MT2-30B-A3B approaches Kimi K2.6, GPT-5.5, and Gemini 3.1 Pro in Total score, and even surpasses several ultra-large models on Simple instructions. This indicates that targeted optimization for translation instruction following can effectively improve constraint understanding and execution.

On general instruction-following benchmarks, Hy-MT2-30B-A3B maintains solid performance, reaching 89.80 on IFEval and 77.39 overall on MaXIFE, outperforming Qwen3.6-35B-A3B on both metrics. However, its Multi-IF scores are lower in later turns than some comparable baselines, suggesting that the main advantage of Hy-MT2 lies in translation-specific instruction following rather than general multi-turn instruction following.

3.5 Quantization Experiment

We evaluate the model size and performance of various quantized Hy-MT2 models across general translation, domain-specific translation, and instruction-following benchmarks, as shown in Table 5. Overall, quantization substantially reduces deployment cost while preserving strong translation quality. For Hy-MT2-1.8B and Hy-MT2-7B, FP8 achieves performance very close to BF16 across most benchmarks,

Table 4: Performance comparison on instruction-following benchmarks.

Models	IFBench	IFEval	MaXIFE			Multi-IF	IFMTBench		
			Loose	Strict	Overall		Simple	Complex	Total
Gemini 3.1 Pro ^T	71.33	96.30	90.58	87.62	89.10	95.02 / 89.57 / 84.89	91.95	84.53	89.08
GPT-5.5 ^T	67.33	93.90	89.96	84.58	87.27	94.90 / 89.33 / 84.46	89.25	84.44	87.39
GPT-5.5	43.33	91.68	87.04	81.04	84.04	93.25 / 87.21 / 82.02	86.97	83.74	85.72
DeepSeek-V4-Pro ^T	76.00	91.31	88.53	84.14	86.34	93.62 / 87.02 / 81.69	86.97	83.74	85.72
DeepSeek-V4-Pro	42.67	89.83	84.98	78.77	81.88	90.36 / 83.45 / 76.60	81.38	78.23	80.16
Kimi K2.6 ^T	68.67	95.56	89.19	85.80	87.50	93.87 / 87.15 / 82.29	83.17	80.81	82.26
Kimi K2.6	38.00	90.02	86.40	80.49	83.45	90.60 / 84.93 / 77.48	88.07	79.92	84.92
GLM5.1 ^T	74.00	94.09	88.59	85.61	87.10	93.94 / 86.47 / 82.77	88.81	82.14	86.23
GLM5.1	50.00	90.57	86.40	80.49	83.45	92.24 / 85.25 / 79.67	85.53	78.83	82.94
Qwen3.5-397B-A17B ^T	65.67	89.83	89.88	86.80	88.34	90.81 / 84.17 / 79.56	85.05	72.08	80.04
Qwen3.5-397B-A17B	48.33	88.54	86.26	80.46	83.36	90.50 / 81.62 / 75.86	81.71	77.61	80.13
Gemma4-E2B	26.00	80.44	77.32	68.79	73.06	80.79 / 70.59 / 63.81	62.93	49.10	57.59
Hy-MT2-1.8B	35.33	80.22	61.29	51.05	56.17	70.50 / 49.79 / 35.75	76.76	57.61	69.36
Gemma4-E4B	32.00	85.76	81.58	74.83	78.21	86.81 / 78.04 / 71.73	74.67	66.98	71.70
Hy-MT2-7B	35.33	86.14	76.77	68.73	72.75	79.53 / 66.50 / 54.35	89.73	72.67	83.14
Qwen3.6-35B-A3B	36.00	83.00	80.72	73.39	77.06	84.84 / 78.12 / 71.48	77.79	69.70	74.66
Gemma4-26B-A4B	45.60	89.80	84.87	79.25	82.06	89.17 / 81.89 / 75.29	83.02	73.18	79.22
Hy-MT2-30B-A3B	50.67	89.80	80.46	74.31	77.39	90.10 / 72.73 / 66.66	90.20	75.94	84.69

Notes. All scores are reported as percentages. For Multi-IF, each cell reports turn1 / turn2 / turn3 scores. ^T denotes thinking mode; for models with both modes, the row without ^T denotes non-thinking mode. IFMTBench is our translation instruction-following benchmark, and Simple, Complex, and Total correspond to single-constraint, multi-constraint, and overall scores, respectively. Baselines are grouped into **ultra-large general models** and **medium to small-sized general models**. The best results within the ultra-large models are highlighted in blue background, while the best results within each corresponding parameter-scale group of small-to-medium models are highlighted in pink background.

Table 5: Performance comparison on various quantized model.

Model	FLORES-200			WMT25	Mand	DMTB	WMTB	IFBench	IFEVAL	IFMTB
	ZH ↔ XX	EN ↔ XX	XX ↔ XX							
Hy-MT2-1.8B-BF16	83.49	87.02	79.21	60.33	60.32	92.25	86.74	35.33	80.22	69.36
Hy-MT2-1.8B-FP8	83.11	86.66	78.63	59.51	59.73	92.15	86.57	35.00	80.59	67.06
Hy-MT2-1.8B-Q4_K_M	82.22	85.87	77.19	57.46	57.23	91.85	86.09	33.33	78.93	63.47
Hy-MT2-1.8B-2bit	80.86	84.74	76.31	57.96	58.96	89.67	83.33	34.00	77.63	58.99
Hy-MT2-7B-BF16	89.17	91.43	87.06	73.05	65.49	93.86	89.61	35.33	86.14	83.14
Hy-MT2-7B-FP8	88.92	91.45	86.92	72.59	65.50	93.84	89.49	34.00	86.51	82.38
Hy-MT2-7B-Q4_K_M	88.96	91.46	86.90	72.30	65.57	93.81	89.41	37.67	85.58	75.11
Hy-MT2-30B-A3B-BF16	90.04	92.37	88.13	73.62	65.93	94.39	89.56	50.67	89.80	84.69
Hy-MT2-30B-A3B-FP8	89.97	91.87	88.01	73.38	65.91	94.35	89.61	50.33	88.70	83.47

Notes. All scores are reported as percentages. DMTB, WMTB, and IFMTB denote DomainMTBench, WildMTBench, and IFMTBench, respectively. Mand denotes Mandarin-minority language bidirectional translation. IFMTB is our translation instruction-following benchmark.

indicating that low-precision inference can be applied with minimal quality degradation. Q4_K_M also maintains competitive performance, especially on FLORES-200 and domain benchmarks, though larger drops appear on instruction-following tasks such as IFMTBench. The 2-bit Hy-MT2-1.8B shows a more noticeable decline, suggesting that extreme quantization introduces a stronger trade-off between efficiency and accuracy. For the MoE-based Hy-MT2-30B-A3B, FP8 remains highly stable compared with BF16, demonstrating that larger architectures are more robust to quantization. These results show that Hy-MT2 provides flexible deployment options, from highly efficient compact models to high-quality large models, while maintaining strong multilingual translation capability.

3.6 Case Study on Translation Instruction Following

Figures 3 to 5 illustrate representative examples of Hy-MT2 handling translation-specific instructions. These cases encompass common real-world requirements, including style control, delimiter preservation, legal redaction, idiom explanation, subtitle translation with cultural preservation, legal-document anonymization, and language-learning-oriented annotation.

Type	Prompt	Response
	<p>参考下面的术语进行翻译： Cloud Native 翻译成云端原生化，Load Balancer 翻译成流量分发器，Container 翻译成隔离沙箱，Endpoint 翻译成服务触点，Traffic 翻译成数据洪流，Secure Protocol 翻译成高保密协定</p> <p>将以下文本翻译为中文，注意只需要输出翻译后的结果，不要额外解释：</p> <p>In a Cloud Native architecture, the Load Balancer routes the Traffic to each Container using a Secure Protocol via the designated Endpoint.</p>	<p>在云端原生化架构中，流量分发器会通过指定的服务触点，利用高保密协定将数据洪流路由到各个隔离沙箱。</p>
Terminology	<p>Consulta las siguientes traducciones: Junta Directiva se traduce como 管理委员会 Accionista se traduce como 出资人 Fusión se traduce como 资产重组 Activo se traduce como 核心资产 Pasivo se traduce como 遗留债务 Acuerdo se traduce como 终局协议 Voto se traduce como 表决权 Borrador se traduce como 征求意见稿</p> <p>Traduce el siguiente texto al chino. Atención: proporciona únicamente el resultado de la traducción, sin explicaciones adicionales:</p> <p>El Borrador del Acuerdo de Fusión fue revisado por la Junta Directiva. Cada Accionista utilizó su Voto para evaluar el Activo y el Pasivo de la empresa.</p>	<p>该资产重组终局协议征求意见稿已经过管理委员会的审议。各位出资人均行使各自的表决权，对公司的核心资产及遗留债务进行了评估。</p>
Style control	<p>请将以下这则现代软件的更新日志翻译为中文。</p> <p>使用【中国古代文言文】风格进行翻译。用词需古雅精炼，句式尽量对仗，仿佛是一份呈交给古代帝王的奏折。</p> <p>源文本： We just fixed a critical bug where the application crashes immediately when the user clicks the "Save" button. Please update your software to the latest version right now so you don't lose your important data. Veuillez traduire l'avis d'annulation de vol suivant en chinois. Utilisez le style d'un service client de compagnie aérienne haut de gamme, faisant preuve d'une profonde empathie. Le ton doit être extrêmement sincère et empreint d'excuses. Attention : fournissez uniquement le résultat de la traduction, sans explications supplémentaires : Le vol AF123 est annulé à cause de la météo. Vous serez remboursés d'ici 7 jours. Appelez le centre d'aide pour toute question.</p>	<p>启禀陛下：</p> <p>近日弊处应用程序有一要务疏漏，凡有臣工误触“存盘”之钮，则程式即刻崩溃，致事功功败垂成。伏望陛下即刻敕令群臣将软件更新至最新款式，庶几能保全珍贵资料，免遭遗失之虞。</p> <p>很遗憾地通知您，由于恶劣天气影响，原定起飞的 AF123 次航班不得不被迫取消。我们深知这一变动定会给您带来诸多不便与困扰，对此我们深表歉意。</p> <p>请您放心，我们承诺将在 7 个工作日内为您办理全额退款手续。倘若您有任何疑问或需要协助，随时欢迎致电我们的客服中心，我们将竭诚为您提供服务。</p>
Structured output	<p>请将以下 CSV 文本翻译为中文。</p> <pre>id,event_name,description,status 101,System_Boot,"Normal boot, all systems go.,"SUCCESS 102,Data_Sync,"Error: Connection lost. Retrying in 5 seconds.,"FAILED 103,User_Update,"Smith, John updated his profile.,"PENDING</pre> <p>【任务指令】 请解析以下【待翻译文本】，并严格输出为 JSON 格式。JSON 必须包含以下字段： - "sentence": 原始中文文本。 - "reading": 带有拼音标注的文本，格式要求：逐字标注拼音，拼音用方括号包裹（例如：“测试”转换为“测 [cè] 试 [shì]”）。 - "pinyin": 全句的完整标准拼音。 - "translation": 英文翻译。 注意：只需输出 JSON 结构，禁止输出任何解释性文字或代码块标记。</p> <p>【待翻译文本】 混元翻译模型真棒，速度又快翻译又准。</p>	<pre>id,event_name,description,status 101,系统启动,"正常启动,所有系统运行正常。",成功 102,数据同步,"错误: 连接中断。5 秒后重试。",失败 103,用户信息更新,"约翰·史密斯更新了他的个人资料。",待处理</pre> <p>["sentence": "混元翻译模型真棒，速度又快翻译又准。", "reading": "混 [hùn] 元 [yuán] 翻 [fān] 译 [yì] 模 [mó] 型 [xíng] 真 [zhēn] 棒 [bàng], 速 [sù] 度 [dù] 又 [yòu] 快 [kuài] 翻 [fān] 译 [yì] 又 [yòu] 准 [zhǔn]。", "pinyin": "hùn yuán fān yì mó xíng zhēn bàng, sù dù yòu kuài fān yì yòu zhǔn。", "translation": "The Hunyuan translation model is wonderful; it's fast and highly accurate."]</p>
Context awareness	<p>背景：两人在排查服务器故障。 请翻译：That was a disaster. We need to do another run.</p> <p>背景：两人在剧院舞台上排练话剧。 请翻译：That was a disaster. We need to do another run.</p>	<p>真是糟透了。我们得重新排查一遍。</p> <p>这次排练彻底搞砸了。我们得再从从头排一遍。</p>
Legal redaction	<p>帮我把这句英文公关稿翻译成正儿八经的中文商务话术：“We are currently implementing a strategic organizational restructuring to streamline operations and empower our remaining workforce with more dynamic and expansive roles.”</p> <p>翻完之后，你再化身一个被压榨的打工人，用一句 50 字左右的话疯狂吐槽一下</p>	<p>目前，我正着手推进组织架构的战略调整，旨在优化业务流程，为现有员工创造更具活力且空间更为广阔的职业发展平台。</p> <p>老天，这所谓的“结构调整”说白了就是裁员换岗啊！留下来的倒霉蛋得扛下成倍的工作量，结果工资照样一分不涨，纯粹是被榨干啊！</p>
Idiom explanation	<p>【任务指令】 请将以下【待翻译文本】翻译为地道的中文，并解释其中的俚语。同时请告诉我，对于没有海外背景的中国读者来说，这个俚语是否有理解门槛。</p> <p>【待翻译文本】 I'm a bit busy today, but let's touch base tomorrow morning to finalize the meeting agenda.</p>	<p>【俚语解释】 “Touch base”是一个常用的英语俚语，字面意思是“接触基地”，实际含义是“联系一下”或“互通消息”。它通常用于工作场合，表示双方需要通过沟通来确认信息或推进事务。</p> <p>【理解门槛】 对于没有海外背景的中国读者来说，这个俚语的理解门槛较低。虽然“touch base”并非中文里现成的固定表达，但其含义与中文的“联系一下”“碰个头”等说法高度契合，读者稍加联想就能明白其意思。因此，它不会造成太大的理解困难。</p>

Figure 3: Case study of Hy-MT2 on translation instruction following (Part 1).

The examples demonstrate that Hy-MT2 can handle both explicit translation constraints and compound user requirements. In the style-control case, the model rewrites a modern software update log into classical Chinese with a formal, memorial-like tone. For delimiter preservation, it accurately maintains special markers (e.g., “###”) while translating surrounding content, showing precise formatting awareness.

Hy-MT2 also excels in more complex scenarios. In legal redaction and anonymization tasks, it translates while applying required redactions, paragraph restructuring, and numbering rules. In idiom-explanation and language-learning cases, it provides not only translations but also furigana, romaji output, and

Type	Prompt	Response
	<p>【任務指令】 以下の日本語の会話を英語の字幕に翻訳してください。その際、以下の制約を厳格に遵守すること:</p> <p>1. 文化的要素の維持: 会話内の敬称・敬語の接尾辞 (-san, -kun, senpai など) は翻訳せず、そのまま英語に含めること。 2. フォーマット要件: 標準的なSRTフォーマットに従い、すべてのセリフにシーケンス番号とタイムコード (00:00:00.000 → 00:00:00.000) を必ず含めて出力すること。</p> <p>【翻訳対象テキスト】 たけし: 「先輩、これ本当に大丈夫なんですか?」 ゆうこ先輩: 「だからさっきから言ってるじゃん。心配しすぎだよ、たけし君。」 たけし: 「いや、でも校則違反ですよ、絶対にバレますって!」 ゆうこ先輩: 「バレなきゃ違反じゃないの! はい、行くよ!」</p>	<p>1 00:00:00.000 → 00:00:03.000 Takeshi: "Senpai, is this really alright to do?"</p> <p>2 00:00:03.000 → 00:00:06.000 Yuuko-senpai: "Like I've been saying. You're worrying way too much, Takeshi-kun."</p> <p>3 00:00:06.000 → 00:00:10.000 Takeshi: "But this violates the school rules! We're definitely going to get caught!"</p> <p>4 00:00:10.000 → 00:00:13.000 Yuuko-senpai: "If we don't get caught, it's not a violation! Alright, let's go!"</p>
		<p>Traducción al español: Todos se rieron cuando empezó a entrenar su propio modelo MoE en una Mac, pero la verdad: dejémoslo trabajar tranquilo.</p>
Allusion or cultural adaptation analysis	<p>[Task] Translate this tweet into Spanish and explain the slang. Tell me if any reference would be lost on a Latin American reader.</p> <p>[Source Text] Everyone laughed when he started training his own MoE model on a Mac, but honestly, just let him cook.</p>	<p>Explicación del slang: - "Let him cook": Es una expresión muy popular en inglés (sobre todo en contextos digitales y de gaming). Significa "deja que haga lo que quiera/que trabaje en paz", es decir, no interfieras ni critiques; permítele desarrollar su idea sin interrupciones. En español no hay una traducción única, pero se puede expresar como "dejémoslo trabajar tranquilo" o "no le pongamos trabas".</p> <p>¿Se perdería alguna referencia para un lector latinoamericano? En general, la frase es fácil de entender en cualquier país de Latinoamérica. Sin embargo, el término "MoE model" (Modelo de Expertos Múltiples) podría no ser conocido por todos, ya que es un concepto técnico especializado en IA. Si el público objetivo no tiene familiaridad con temas de machine learning, podría ser útil añadir una breve aclaración como: "un tipo de modelo de inteligencia artificial". Aparte de eso, no hay referencias culturales que se pierdan.</p>
Data Anonymization	<p>【Anweisung】 Übersetzen Sie den folgenden Text in ein deutsches juristisches Dokument.</p> <p>【Regeln】 1. Erzwingene Schwärzung: Ersetzen Sie alle Firmennamen [Firmenname], Daten [Datum] und Geldbeträge [Geldbetrag] exakt durch '[REDACTED]'. 2. Restrukturierung der Klauseln: Jeder Satz des Ausgangstextes muss als eigenständiger Absatz formatiert werden. Jeder Absatz muss zwingend mit einer fortlaufenden Nummerierung am Anfang versehen werden (z. B. '\$ 1.', '\$ 2.'...).</p> <p>Hinweis: Geben Sie NUR den reinen Übersetzungstext aus. Jegliche Erklärungen oder Kommentare sind absolut untersagt.</p> <p>【Zu übersetzender Text】 The non-disclosure agreement between TechCorp Inc. and Global Data Solutions shall officially enter into force on May 15, 2026. In the event of a material breach of confidentiality, the breaching party shall be liable for a penalty fee of \$500,000. This agreement is governed by the laws of the State of California.</p>	<p>\$ 1. Dieser Geheimhaltungsvertrag zwischen [REDACTED] und [REDACTED] tritt am [REDACTED] offiziell in Kraft.</p> <p>\$ 2. Im Falle eines wesentlichen Verstoßes gegen die Vertraulichkeitsverpflichtungen haftet die verletzte Partei für die Zahlung einer Vertragsstrafe in Höhe von [REDACTED].</p> <p>\$ 3. Für diesen Vertrag gilt ausschließlich das Recht des Bundesstaates Kalifornien.</p>
Blocklist settings	<p>【任務】 将以下【待翻译文本】翻译为俄语。</p> <p>【规则】 1. 核心词屏蔽: 在生成的俄语译文中, 绝对禁止出现以下三个最基础的高频词汇: - 禁止出现 'хорошо' (好) 及其所有变体 - 禁止出现 'делать' (做/进行) 及其所有变体 - 禁止出现 'работа' (工作/运行) 及其所有变体 你必须使用更高级、更小众的俄语同义词来优雅地表达原意。</p> <p>【待翻译文本】 算法表现很不错; 模型现在正完美地运行。</p>	<p>Алгоритм демонстрирует превосходные результаты; модель в настоящее время функционирует безупречно.</p>
Language study	<p>我想学点地道的西语。帮我把手发我的朋友翻译成西班牙语: '说实话, 这段感情早就名存实亡了, 不是谁对谁错, 就是单纯的走散了。祝你前程似锦。'</p> <p>翻完之后, 不要就这么结束了。你要化身一个严厉但懂幽默的西班牙语外教, 从你刚才的译文里挑出 3 个最地道、最难的表达 (比如俚语或者特殊语法现象) 给我解释一下。最后, 根据这 3 个知识点, 马上给我出 3 道选择题考考我, 并在末尾等我回答。</p>	<p>La verdad es que esta relación ya hacía tiempo que había perdido todo sentido. No se trata de quién tiene la razón o el error; simplemente, cada uno siguió su propio camino. Te deseo todo lo mejor en tu futuro.</p> <p>好了, 现在我来给你分析一下刚才那段译文里的三个地道难点:</p> <p>1. "había perdido todo sentido" 这里用了过去完成时 "había perdido", 用来描述在过去某个时间点之前就已经发生并持续存在的状态。这种用法非常常见于描述某种状况的消逝或变化, 是西语里很地道的表达。</p> <p>2. "no se trata de quién tiene la razón o el error" 短语 "no se trata de" 意为 "并不是说……", 用来把话题从某点引开, 强调重点不在于此。它后面常接名词或完整句子, 是口语和书面语里都很常用的结构。</p> <p>3. "cada uno siguió su propio camino" "Cada uno" 表示 "每个人各自", "seguir su propio camino" 则是地道的习语, 字面是 "走自己的路", 实际用来指人们各自走上了不同的方向或选择了不同的生活轨迹。</p> <p>下面是我为你准备 of 3 道选择题, 请仔细思考后给出答案吧:</p> <p>1. 在句子 "La verdad es que esta relación ya había perdido todo sentido." 中, 使用过去完成时 "había perdido" 的主要目的是什么? A. 强调动作刚刚完成 B. 描述在某一过去时间点之前已经发生的状态 C. 表示动作正在进行</p> <p>2. 短语 "no se trata de" 在句中起到什么作用? A. 引出结论 or 总结 B. 把话题从当前点转移开, 强调重点不在此处 C. 表示因果关系</p> <p>3. 习语 "seguir su propio camino" 最贴切的汉语意思是? A. 沿着别人指的路走 B. 走捷径以节省时间 C. 各自走上不同的道路或选择不同的方向</p> <p>期待你的回答哦!</p>

Figure 4: Case study of Hy-MT2 on translation instruction following (Part 2).

explanatory notes, meeting diverse instructional requirements.

Moreover, in subtitle translation, Hy-MT2 preserves culturally specific elements such as honorifics while following SRT formatting rules, reflecting its ability to manage multilingual instructions and maintain cultural fidelity.

Type	Prompt	Response
Style constraints and conflict evaluation	帮我把这句餐饮品牌的英文升级公告翻译成 优雅、高逼格的中文营销话术 : "We are reimagining our culinary offerings to focus on artisanal, premium ingredients, resulting in a more curated and mindful dining experience." 翻完之后, 你再化身一个刚在这家店吃完饭、没吃饱的老主顾, 用一句 50 字左右的话疯狂吐槽一下。	我们正全面重塑菜单体系, 力邀顶级匠人甄选珍稀上等食材, 只为您呈现一场极具格调且匠心独运的极致用餐体验。 点了一堆所谓“精粹美食”, 结果量少得可怜, 根本填不饱肚子! 这哪是高端体验, 纯粹是拿我当冤大头来宰啊!
Social media-style copywriting	我在网上看到个老外狂喷一家餐厅: "Absolutely disgusting! The steak was as hard as a brick, the waiter completely ignored us for 40 minutes, and to top it off, they charged me \$5 for a glass of tap water! A total scam." 你先帮我接地气地翻译一下他到底在骂啥。然后帮我整点活, 把这个差评的核心槽点改成一首 七言绝句 , 表面上要显得很有文化很文雅, 记得要押韵啊。	这家老外把餐厅骂得狗血喷头: 简直恶心到极点! 牛排硬得跟石头似的, 服务员整整 40 分钟压根不搭理我们, 最离谱的是, 他们居然收我 5 美元就给一杯自来水! 完全是坑人骗局! 青石如铁难下咽, 侍者漠然久置前。 一杯清水价五贯, 欺客敛财实可叹。
Delimiter preservation	请将以下文本准确翻译为中文。 你必须在译文中 保留等量的分隔符 , 绝对不可遗漏、转义或翻译该符号, 并注意分隔符的位置。 For the party, we need to buy @@ some apples, @@ two bananas, @@ and a bottle of milk.	为了聚会, 我们需要买@@一些苹果, @@两根香蕉, @@还有一瓶牛奶。
Multilingual translation instruction	Переведите следующий текст на русский. Обратите внимание: выведите тот же результат перевода, без каких-либо дополнительных об'яснений: The new algorithm significantly improves the translation quality and reduces the processing latency.	Новый алгоритм значительно повышает качество перевода и сокращает время обработки.
Emotional translation	[Aufgabe] Übersetzen Sie den folgenden [Ausgangstext] ins Englische. [Regeln] 1. Großschreibung: Verwenden Sie ausnahmslos GROSSBUCHSTABEN für alle englischen Buchstaben. Es sind keine Kleinbuchstaben erlaubt. 2. Interpunktion: Jeder einzelne Satz muss mit genau drei aufeinanderfolgenden Ausrufezeichen (!!!) enden. Nicht mehr und nicht weniger. 3. Ausgabeformat: Geben Sie NUR die reine Übersetzung aus. Keine Erklärungen oder zusätzlichen Text. [Ausgangstext] Drücken Sie die Entkalkungstaste für drei Sekunden. Warten Sie, bis die rote Leuchte erlischt.	PRESS THE DESCALING BUTTON FOR THREE SECONDS!!! WAIT UNTIL THE RED LIGHT GOES OUT!!!

Figure 5: Case study of Hy-MT2 on translation instruction following (Part 3).

Overall, these examples confirm that Hy-MT2 reliably executes diverse translation instructions, accommodating constraints on language, style, format, cultural context, and auxiliary explanatory needs.

4 Conclusion

In this paper, we present Hy-MT2, a multilingual machine translation model family designed for real-world translation scenarios. Hy-MT2 covers both dense and mixture-of-experts architectures, including **Hy-MT2-1.8B**, **Hy-MT2-7B**, and **Hy-MT2-30B-A3B**, all supporting translation among 33 languages. Compared with Hy-MT1.5, Hy-MT2 provides systematic improvements in domain-specific translation, real-world scenario translation, translation instruction following, model scaling, and efficient on-device deployment. Hy-MT2-7B and Hy-MT2-30B-A3B outperform strong open-source translation baselines such as DeepSeek-V4-Pro and Kimi K2.6, and achieve performance close to or even surpassing leading closed-source models such as Gemini 3.1 Pro on multiple benchmarks. The lightweight Hy-MT2-1.8B also demonstrates strong small-model translation capability, outperforming several commercial translation APIs. To support diverse deployment scenarios, Hy-MT2 is released in multiple precision formats, including **1.25-bit**, **2-bit**, **4-bit**, **8-bit**, and **FP16**. Among them, the 1.25-bit and 2-bit versions are built on Hunyuan self-developed quantization techniques, significantly reducing model resource consumption while improving inference efficiency. Overall, Hy-MT2 provides a high-quality, efficient, and multi-capability multilingual translation model family for real-world applications.

5 Contributions

5.1 Core Contributors

Mao Zheng, Zheng Li, Tao Chen, Bo Lv, Mingrui Sun, Mingyang Song, Jinlong Song, Hong Huang, Decheng Wu, Hai Wang, Yifan Song, Yanfeng Chen, Guanwei Zhang

5.2 Contributors

Guanghua Yu, Yi Su, Hong Liu, Jinxiang Ou, Keyao Wang, Weile Chen, Haozhao Kuang, Kai Wang, Nuo Chen, Zihao Zheng, Chenhao Wang, Bin Xing, Chengcheng Xu, Tinghao Yu, Binghong Wu, Long Xu, Jiacheng Shi, Yunhao Wang, Baifang Chen, Lei Zhang, Qi Yang, Zhao Wu, Jiacheng Li, Lan Jiang, Lanrui Wang, Kai Zhang, ShuaiPeng Li, Zhongzhi Chen, Weixuan Sun, Jiaqi Zhu, An Wang, Wei Li, Jun Xia, Weidong Han, Wutian Yang, Litong Hui, Luoguo Jia, Jiajia Wu, Hongchuan Zeng, Zheng Zhang, Xinpeng Zhou, Tianxiang Fei

References

- DeepMind. Introducing gemini 3. <https://blog.google/products/gemini/gemini-3-collection/>, 2025. Accessed: 2025-12-29.
- Dayou Du, Yijia Zhang, Shijie Cao, Jiaqi Guo, Ting Cao, Xiaowen Chu, and Ningyi Xu. Bitdistiller: Unleashing the potential of sub-4-bit llms via self-distillation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 102–116, 2024.
- Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In Loic Barrault, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz (eds.), *Proceedings of the Sixth Conference on Machine Translation*, pp. 733–774, Online, November 2021. Association for Computational Linguistics. URL <https://aclanthology.org/2021.wmt-1.73/>.
- Yun He, Di Jin, Chaoqi Wang, Chloe Bi, Karishma Mandyam, Hejia Zhang, Chen Zhu, Ning Li, Tengyu Xu, Hongjiang Lv, et al. Multi-if: Benchmarking llms on multi-turn and multilingual instructions following. *arXiv preprint arXiv:2410.15553*, 2024.
- Hong Huang, Decheng Wu, Qiangqiang Hu, Guanghua Yu, Jinhai Yang, Jianchen Zhu, Xue Liu, and Dapeng Wu. Sherry: Hardware-efficient 1.25-bit ternary quantization via fine-grained sparsification. *arXiv preprint arXiv:2601.07892*, 2026.
- 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, Steinhórfur Steingrímsson, 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/>.
- Yile Liu, Ziwei Ma, Xiu Jiang, Jinglu Hu, ChangJing ChangJing, and Liang Li. MaXIFE: Multilingual and cross-lingual instruction following evaluation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14252–14332, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. URL <https://aclanthology.org/2025.acl-long.698/>.
- Zechun Liu, Changsheng Zhao, Hanxian Huang, Sijia Chen, Jing Zhang, Jiawei Zhao, Scott Roy, Lisa Jin, Yunyang Xiong, Yangyang Shi, et al. Paretoq: Improving scaling laws in extremely low-bit llm quantization. *Advances in Neural Information Processing Systems*, 38:91311–91336, 2026.
- Arlé Lommel, Aljoscha Burchardt, Maja Popović, Kim Harris, Eleftherios Avramidis, and Hans Uszkoreit. Using a new analytic measure for the annotation and analysis of MT errors on real data. In Mauro Cettolo, Marcello Federico, Lucia Specia, and Andy Way (eds.), *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*, pp. 165–172, Dubrovnik, Croatia, June 16-18 2014. European Association for Machine Translation. URL <https://aclanthology.org/2014.eamt-1.38/>.
- OpenAI. Openai gpt-5 system card, 2026. URL <https://arxiv.org/abs/2601.03267>.
- Valentina Pyatkin, Saumya Malik, Victoria Graf, Hamish Ivison, Shengyi Huang, Pradeep Dasigi, Nathan Lambert, and Hannaneh Hajishirzi. Generalizing verifiable instruction following, 2025.
- 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>.
- Hunyuan AI Infra Team. Angelslim: A more accessible, comprehensive, and efficient toolkit for large model compression. *arXiv preprint arXiv:2602.21233*, 2026.

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>.

Mao Zheng, Zheng Li, Tao Chen, Mingyang Song, and Di Wang. Hy-mt1.5 technical report, 2025. URL <https://arxiv.org/abs/2512.24092>.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.