Hunyuan-MT Technical Report

Tencent Hunyuan Team

In this report, we introduce our first translation models, **Hunyuan-MT-7B** and **Hunyuan-MT-Chimera-7B**. We propose a holistic training framework tailored for machine translation that combines general pre-training, MT-oriented pre-training, supervised fine-tuning, reinforcement learning, and weak-to-strong reinforcement learning. Building on this foundation, our models achieve SOTA performance among models of comparable size, enabling high-quality bidirectional translation across all covered languages. Furthermore, a distinctive focus of our work is *bidirectional translation between Mandarin and several ethnic minority languages*, where we apply targeted data curation and optimization to substantially improve performance in low-resource settings. Extensive experiments validate that **Hunyuan-MT-7B** and **Hunyuan-MT-Chimera-7B** outperform all SOTA baselines in translation between Mandarin and minority languages.
In the WMT2025 shared task (General Machine Translation), our translation model achieves superior performance, securing first place across 30 out of 31 language pairs. *The language pairs span both high-resource languages, including Chinese, English, and Japanese, as well as low-resource languages such as Czech, Marathi, Estonian, and Icelandic.*
GitHub: https://github.com/Tencent-Hunyuan/Hunyuan-MT

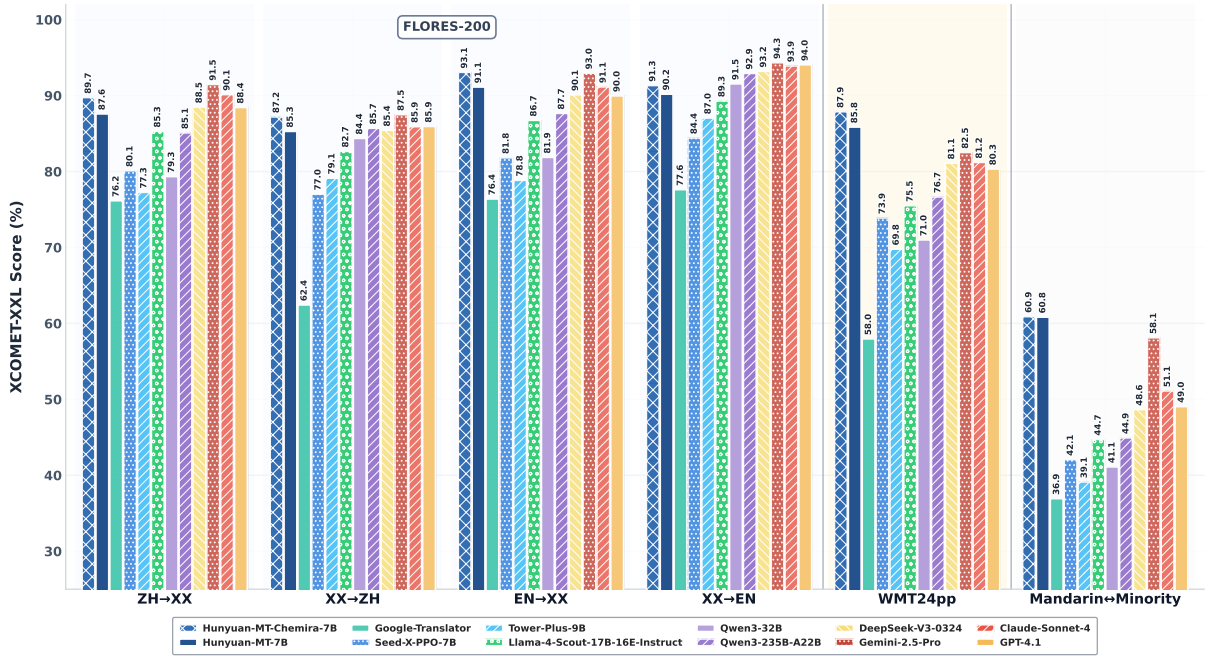

Figure 1 | Benchmark performance of Hunyuan-MT models and state-of-the-art baselines.

1. Introduction

Machine Translation (MT) has emerged as both a critically important practical application and one of the most formidable research challenges that the computational linguistics community has pursued over the past several decades (Bahdanau et al., 2015; Brown et al., 1990, 1993; Papineni et al., 2002; Sutskever et al., 2014; Vaswani et al., 2017; Wu et al., 2016). The recent advent and rapid advancement of Large Language Models (LLMs) have revolutionized the learning paradigm underlying MT systems, catalyzing a shift from traditional rule-based and statistical approaches toward sophisticated large-scale neural learning methodologies (Kocmi et al., 2024; Pang et al., 2025; Zhu et al., 2024). This continuous technological evolution of LLMs has dramatically pushed the boundaries of achievable translation quality to unprecedented levels, with state-of-the-art models such as GPT-4.1 (OpenAI, 2025), Gemini-2.5-Pro (DeepMind, 2025), and Claude-Sonnet-4 (Anthropic, 2025), demonstrating remarkable capabilities that exceed the performance of expert human translators across specific language pairs.

Nevertheless, significant challenges persist (Pang et al., 2025), particularly in the translation of non-literal language, such as internet neologisms, slang, and terminology, as well as place names. Furthermore, a prevailing bias within MT research favors high-resource language pairs, leaving translation for low-resource and minority languages critically under-resourced. The translation between China's minority languages and Mandarin constitutes a particularly acute manifestation of this neglect. Beyond its technical dimensions, facilitating high-quality translation in this context is pivotal for promoting social inclusion, preserving cultural heritage, and ensuring equitable access to essential services and information for minority communities (Hu et al., 2019; Lin and Jackson, 2021). Despite this pressing societal imperative, this specific domain represents a significant lacuna within the MT field.

Addressing these issues requires more than robust linguistic comprehension; it necessitates the ability to generate expressions that are both culturally resonant and idiomatically natural, thereby transcending literal word-for-word translation. Meanwhile, a notable performance disparity remains between proprietary and open-source models, a gap often attributed to the comparatively limited scale of open-source systems. This problem is compounded by a scarcity of well-defined methodologies for developing advanced LLM-based MT systems, which impedes the broader community's efforts to deploy and refine effective solutions (Cheng et al., 2025; Jiao et al., 2023; Kocmi et al., 2024; Pang et al., 2025).

Through extensive evaluations on representative MT benchmarks, Hunyuan-MT demonstrates superior performance, outperforming not only translation-specialized models of comparable size and prominent closed-source systems, such as Google-Translator, but also a range of larger LLMs, as detailed in Figure 1. Furthermore, it is noteworthy that our model demonstrates significant superiority over all state-of-the-art LLMs on the task of translation between China's ethnic minority languages and Mandarin Chinese (Minority<=>Mandarin Translation). In this technical report, we introduce Hunyuan-MT, the culmination of our ongoing efforts to develop more effective LLM-based multilingual translation models. Below, we show the main contributions of this technical report:

1. **Hunyuan-MT-7B**. We have developed and open-sourced a 7B parameter translation model, Hunyuan-MT-7B, that achieves state-of-the-art performance among models of a similar size. It is capable of mutual translation among 33 widely-used languages and notably supports translation between Chinese and its ethnic minority languages.
2. **Hunyuan-MT-Chimera-7B**. We present a first-of-its-kind, open-source weak-to-strong fusion model, Hunyuan-MT-Chimera-7B, for translation. This model is designed to integrate

multiple translation results from different systems at test time, generating a final output that surpasses the quality of any individual candidate.

3. **A Training Recipe**. We establish a holistic training framework for MT that systematically elevates model performance. This framework consists of five distinct sub-stages: general pre-training, MT-oriented pre-training, Supervised Fine-Tuning (SFT), Reinforcement Learning (RL), and weak-to-strong RL, which collectively lead to SOTA results.
4. **Mandarin<=>Minority Translation**. To the best of our knowledge, in this report, we are the first to specifically optimize the performance of bidirectional translation for Mandarin-Kazakh, Mandarin-Cantonese, Mandarin-Uyghur, Mandarin-Mongolian, and Mandarin-Tibetan on an LLM-based MT system. Through extensive experiments, we verify that our models achieve performance that is significantly superior to existing advanced LLMs.

In the subsequent sections, we first describe the Hunyuan-MT training methodology, report experimental results for the pre-trained and post-trained variants, and conclude with a synthesis of key insights and a discussion of prospective research directions.

2. Pre-training

In this section, we describe the details of our pre-training approach and present experimental results from evaluating the base models on standard benchmarks.

2.1. General Pre-training

During the general pre-training, data from Chinese/English, as well as minority languages, are co-trained. The minority language dataset comprises 1.3 trillion tokens, covering 112 non-Chinese/English languages and dialects from diverse sources. We implement a proprietary quality assessment model to grade multilingual data quality, which evaluates text across three dimensions: knowledge value, authenticity, and writing style. This model utilizes a three-tier scoring system (0, 1, 2) for each dimension. We employ weighted composite scores across all three quality dimensions and strategically prioritize specific dimensions in quality assessment weighting based on the characteristics of the data source. e.g., for book-type and professional website content, we prioritize texts achieving a knowledge-level score of 2. To ensure content diversity in multilingual training data, we established three tagging systems for data screening and ratio adjustment:

- Disciplinary Tagging System
  - Labels specialized data sources by academic discipline to balance subject distribution.
- Industry Tagging System (24 categories)
  - Guarantees cross-sector diversity.
- Content Theme Tagging System (24 categories)
  - Enables both diversity management and targeted filtering (e.g., gambling content, advertising).

This framework, comprising the aforementioned quality control model and taxonomic systems, ensures both content diversity and high-quality representation within the multilingual training corpora. Following the procedure described above, we obtain the Hunyuan-7b-Base[1] model.

[1] Hunyuan-7B-Base

Table 1 | **Comparison among Hunyuan-7B-Base and other strong open-source baselines. The highest and second-best scores are shown in bold and underlined, respectively.**

	Llama-3-8B-Base	Qwen2.5-7B-Base	Qwen3-8B-Base	Hunyuan-7B-Base
MMLU-Pro	35.36	45.00	56.73	**57.79**
SuperGPQA	20.54	26.34	**31.64**	30.47
BBH	57.70	70.40	78.40	**82.95**
GPQA	25.80	36.36	**44.44**	44.07
GSM8K	55.30	85.36	**89.84**	88.25
MATH	20.50	49.80	60.80	**74.85**
MultiPL-E	31.45	50.73	58.75	**60.41**
CRUX-O	36.80	48.50	**62.00**	60.75
IINCLUDE	44.94	53.98	59.40	**59.55**

General Evaluation. We conduct comprehensive evaluations of the base language model of the Hunyuan-MT series, focusing primarily on its performance in general knowledge, reasoning, mathematics, scientific knowledge, coding, and multilingual capabilities. Specifically, the evaluation benchmarks for pre-trained models include nine widely used benchmarks:

- General Tasks:
  - MMLU-Pro (Wang et al., 2024)
  - SuperGPQA (Team et al., 2025b)
  - BBH (Shi et al., 2023)
- Math & STEM Tasks:
  - GPQA (Rein et al., 2023)
  - GSM8K (Cobbe et al., 2021)
  - MATH (Hendrycks et al., 2021)
- Coding Tasks:
  - MultiPL-E (Cassano et al., 2023)
  - CRUX-O of CRUXEval (Gu et al., 2024)
- Multilingual Tasks:
  - INCLUDE (Romanou et al., 2025)

2.2. MT-oriented Pre-training

During the MT-oriented pre-training stage, we incorporate a curated mixture of monolingual and bilingual corpora. The monolingual data primarily comes from the mC4 (Raffel et al., 2019) and OSCAR (Ortiz Su'arez et al., 2019; Ortiz Su'arez et al., 2020) datasets. We subject this data to a rigorous cleaning pipeline that includes language identification with fastText [2], document-level deduplication via minLSH, and quality filtering with a KenLM-based model [3] to remove high-perplexity documents. For the bilingual data, we utilize publicly available parallel corpora, such as OPUS (Tiedemann, 2012) and ParaCrawl (Buck and Koehn, 2016), which we

[2] https://github.com/facebookresearch/fastText
[3] https://github.com/kpu/kenlm

filter using reference-free quality estimation metrics, including CometKiwi (Rei et al., 2022) , to ensure the selection of high-quality sentence pairs.

To determine the optimal data mixture ratio, we adopt a strategy inspired by RegMix (Liu et al., 2025). We first conduct experiments on a smaller-scale model to fit a function that maps sampling ratios to training loss. By simulating this function, we identify the mixture that minimizes the predicted loss, and we then use this ratio for the MT-oriented pre-training stage of our final translation model.

To mitigate catastrophic forgetting, we integrate a 20% replay of the original pre-training corpus. We also design the learning rate schedule to warm up to the peak learning rate of the initial pre-training phase and then decay to its minimum value.

Table 2 | **Comparison among Hunyuan-7B-Base† and Qwen3-8B-Base. Here, Hunyuan-7B-Base* denotes the Hunyuan-7B-Base model after MT-oriented Pre-training. The highest and second-best scores are shown in bold and underlined, respectively.**

	FLORES-200		**WMT24pp**		**Mandarin<=>Minority**	
	XCOMET-XXL	CometKiwi	XCOMET-XXL	CometKiwi	XCOMET-XXL	CometKiwi
Qwen3-8B-Base	<u>57.88</u>	<u>55.46</u>	<u>35.89</u>	<u>36.69</u>	<u>32.02</u>	<u>23.98</u>
Hunyuan-7B-Base*	**67.41**	**65.87**	**48.34**	**46.29**	**39.95**	**28.05**

MT-oriented Evaluation. To comprehensively evaluate the multilingual translation capabilities, we conducted extensive experiments using the following test sets:

- **FLORES-200**[4] (Team et al., 2022). We select 1,056 language pairs across 33 different languages (detailed in the Appendix) from the FLORES-200 dataset. These pairs are systematically categorized into five groups: English=>XX, XX=>English, Chinese=>XX, XX=>Chinese, and XX=>XX translations.
- **WMT24pp**[5] (Deutsch et al., 2025). We incorporate development sets from WMT-25, encompassing English-to-XX translations across 25 target languages. WMT24pp serves as the official development set recommended by WMT25. We select 29 language pairs that overlap with the general translation track of the WMT25 competition, with a primary focus on English-to-XX directions.
- **Mandarin<=>Minority Testset**. This test set encompasses translations between Chinese and five minority languages: Tibetan, Mongolian, Uyghur, Kazakh, and Cantonese.

We evaluate translations using two complementary approaches: automatic metrics and human evaluation. 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 but can be unreliable for certain translation phenomena. To address these limitations, 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.

3. Post-training

Following pre-training, we aim to equip the base model with robust multilingual machine translation capabilities through Supervised Fine-Tuning (SFT), Reinforcement Learning (RL),

[4] Flores-200
[5] WMT24pp

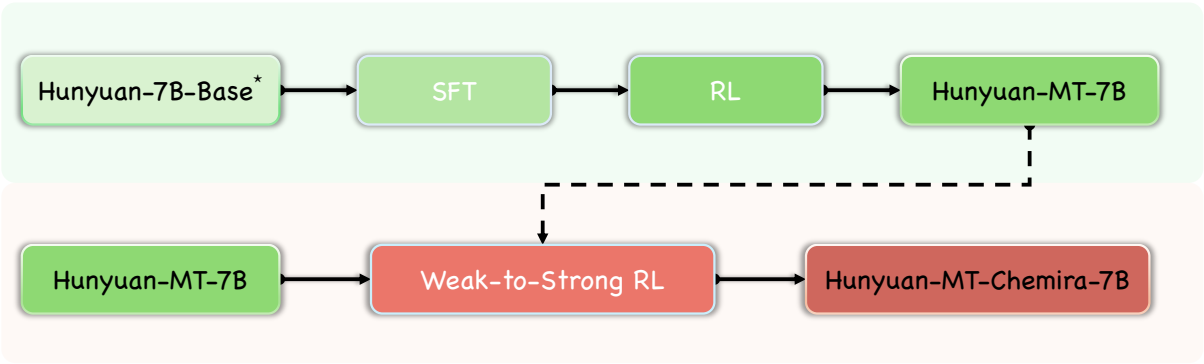

Figure 2 | Post-training pipeline of the Hunyuan-MT-7B and Hunyuan-MT-Chemira-7B models.

and Weak-to-Strong RL. The central challenge lies in optimizing performance for high-resource languages using model-generated and human-annotated data while ensuring effective generalization to low-resource languages.

Prompt Template for ZH<=>XX Translation.
把下面的文本翻译成<target_language>，不要额外解释。 <source_text>
Prompt Template for XX<=>XX Translation, excluding ZH<=>XX.
Translate the following segment into <target_language>, without additional explanation. <source_text>

Table 3 | Examples of prompt template.

3.1. Supervised Fine-Tuning (SFT)

Our Supervised Fine-Tuning (SFT) process is structured into two distinct stages.

Stage I focuses on enhancing the model's foundational translation capabilities and its adherence to translation instructions. This is achieved by training the model on an extensive parallel corpus comprising approximately 3 million pairs. This dataset is aggregated from five primary sources:

- The development set from the open-source Flores-200 benchmark, which covers mutual translations across 33 languages.
- Test sets from previous years of the WMT, predominantly featuring English-to-XX pairs.
- A human-annotated, curated dataset of Mandarin-to-minority and minority-to-Mandarin language pairs.
- A synthetic parallel corpus generated using DeepSeek-V3-0324 (DeepSeek-AI, 2024).
- A 20% component of the general-purpose and MT-oriented instruction-tuning dataset is curated to enhance the model's generalization capabilities for general and translation-related instructions, as shown in Table 3.

To improve the quality of the training data, we employ the reference-free quality estimation metrics CometKiwi and GEMBA (Kocmi and Federmann, 2023) to score the entire parallel

corpus. Training samples that fall below a predefined quality threshold are filtered out. For the Gemba scoring, the DeepSeek-V3-0324 model itself serves as the evaluator.

Stage II aims to refine the model's translation performance further using a smaller, higher-fidelity dataset of approximately 268,000 pairs. The training data for this stage undergoes a more rigorous selection process. Building on previous studies (Agarwal et al., 2024; Song et al., 2025b), we employ many-shot in-context learning to refine the training data further. Furthermore, any training samples exhibiting poor score consistency across multiple evaluation rounds are subjected to manual annotation and verification to ensure the data quality.

3.2. Reinforcement Learning (RL)

While large-scale RL has demonstrated significant effectiveness in enhancing reasoning capabilities for tasks with structured outputs, such as mathematical problem-solving and code generation (DeepSeek-AI, 2025; Luo et al., 2025; Song et al., 2025a; Team et al., 2025a), its application to MT presents a unique challenge. Unlike in structured domains, MT outputs are characterized by considerable semantic diversity, which makes them resistant to evaluation through explicit, rule-based evaluation.

To address this challenge, we adopt GRPO (Shao et al., 2024) as the RL algorithm and design a comprehensive reward function comprising the following components:

Quality-Aware Reward. To ensure translation quality during RL training, we employ two complementary reward signals. The first is XCOMET-XXL, a widely adopted metric in translation evaluation scenarios that demonstrates high correlation with human assessments. The second reward utilizes DeepSeek-V3-0324 for scoring, with prompts adapted from the GEMBA framework.

Terminology-Aware Reward. While XCOMET-based rewards primarily focus on overall semantic similarity between translated outputs and reference translations, they may inadequately capture critical information such as domain-specific terminology. To address this limitation, we incorporate the word alignment-based reward metric proposed in TAT-R1 Li et al. (2025). This reward mechanism extracts key information, including terminology, through word alignment tools, then computes the overlap ratio of these critical elements between the translation output and reference. Higher overlap ratios yield greater rewards, thereby enhancing the model's attention to terminology and other crucial information during training.

Repetition Penalty. We observe that models tend to generate repetitive outputs in later stages of reinforcement training, potentially leading to training collapse. To mitigate this issue, we implement a repetition detection mechanism that applies penalties when repetitive patterns are identified, thereby maintaining output diversity and training stability.

3.3. Weak-to-Strong RL

Recent studies demonstrate that increasing inference time significantly enhances model performance on mathematical and coding tasks (Muennighoff et al., 2025; Zhang et al., 2025). However, initial experiments incorporating Chain-of-Thought (CoT) (Wei et al., 2023) reasoning into translation tasks yield limited improvements in translation quality, as discussed comprehensively in Section 6. Consequently, in this report, we explore a new test-time scaling methodology from a novel perspective to enhance test-time performance for MT.

Specifically, to address the above issue, we propose a weak-to-strong RL approach that generates

Table 4 | Performances of different state-of-the-art models on Flores-200, WMT-24pp, and Mandarin<=>Minority translation. Specifically, we report the Chinese-centric (ZH=>XX and XX=>ZH), English-centric (EN=>XX and XX=>EN), XX=>XX, and Mand.<=>Min. performances of Hunyuan-MT-7B and prominent existing systems. Specifically, Mand.<=>Min. denotes Mandarin<=>Minority translation. Models with open-source weights are marked with †. 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					WMT24pp	Mand.<=>Min.
		ZH => XX	XX => ZH	EN => XX	XX => EN	XX => XX		
GPT4.1	XCOMET-XXL	0.8843	0.8593	0.8996	0.9405	0.8258	0.8032	0.4904
	CometKiwi	0.7859	0.7725	0.8702	0.8730	0.7424	0.7688	0.5020
Claude-Sonnet-4	XCOMET-XXL	0.9013	0.8590	0.9114	0.9390	0.8548	0.8120	0.5111
	CometKiwi	0.7883	0.7739	0.8742	0.8732	0.7668	0.7804	0.5033
Gemini-2.5-Pro	XCOMET-XXL	0.9146	0.8748	0.9295	0.9432	0.8773	0.8250	0.5811
	CometKiwi	0.7859	0.7828	0.8869	0.8720	0.7674	0.7876	0.5418
DeepSeek-V3-0324†	XCOMET-XXL	0.8848	0.8542	0.9010	0.9319	0.8082	0.8109	0.4865
	CometKiwi	0.7722	0.7708	0.8684	0.8703	0.7450	0.7840	0.4839
Google-Translator	XCOMET-XXL	0.7615	0.6243	0.7638	0.7761	0.6225	0.5796	0.3692
	CometKiwi	0.6647	0.5691	0.7242	0.7863	0.5947	0.5454	0.2891
Tower-Plus-9B†	XCOMET-XXL	0.7726	0.7912	0.7884	0.8704	0.6608	0.6977	0.3912
	CometKiwi	0.6633	0.7429	0.7576	0.8475	0.6141	0.6741	0.3466
Tower-Plus-72B†	XCOMET-XXL	0.7703	0.8235	0.7829	0.9002	0.7002	0.7276	0.3855
	CometKiwi	0.6795	0.7569	0.7603	0.8624	0.6553	0.7071	0.3540
Seed-X-PPO-7B†	XCOMET-XXL	0.8010	0.7702	0.8181	0.8442	0.6896	0.7388	0.4206
	CometKiwi	0.7089	0.7201	0.8118	0.8202	0.6436	0.7297	0.4861
GemmaX2-28-9B-v0.1†	XCOMET-XXL	0.8687	0.8280	0.8806	0.9108	0.8119	0.7173	0.4269
	CometKiwi	0.7604	0.7595	0.8576	0.8635	0.7371	0.7242	0.5178
Gemma-3-12B-IT†	XCOMET-XXL	0.8567	0.8249	0.8781	0.9189	0.8020	0.7527	0.4280
	CometKiwi	0.7603	0.7582	0.8498	0.8666	0.7320	0.7329	0.4186
Gemma-3-27B-IT†	XCOMET-XXL	0.8783	0.8441	0.9036	0.9331	0.8381	0.7742	0.4558
	CometKiwi	0.7667	0.7718	0.8630	0.8646	0.7471	0.7577	0.4844
Qwen3-8B†	XCOMET-XXL	0.7250	0.8056	0.7468	0.8825	0.6544	0.6532	0.3737
	CometKiwi	0.6605	0.7507	0.7257	0.8521	0.6285	0.6344	0.3166
Qwen3-14B†	XCOMET-XXL	0.7826	0.8318	0.8027	0.9049	0.7228	0.6983	0.3944
	CometKiwi	0.7116	0.7674	0.7877	0.8639	0.6887	0.6839	0.3568
Qwen3-32B†	XCOMET-XXL	0.7933	0.8436	0.8186	0.9154	0.7433	0.7099	0.4110
	CometKiwi	0.7139	0.7719	0.8001	0.8657	0.6965	0.6930	0.3841
Qwen3-235B-A22B†	XCOMET-XXL	0.8509	0.8569	0.8765	0.9292	0.8018	0.7665	0.4493
	CometKiwi	0.7551	0.7750	0.8475	0.8696	0.7313	0.7465	0.4456
Llama-3.1-8B-Instruct†	XCOMET-XXL	0.6385	0.5148	0.6848	0.6412	0.4408	0.5130	0.3016
	CometKiwi	0.6581	0.6185	0.7234	0.7746	0.5782	0.5990	0.2944
Llama-4-Scout-17B-16E-Instruct†	XCOMET-XXL	0.8529	0.8266	0.8673	0.8929	0.7788	0.7550	0.4472
	CometKiwi	0.7457	0.7417	0.8418	0.8523	0.7275	0.7422	0.4646
Hunyuan-MT-7B†	XCOMET-XXL	0.8758	0.8528	0.9112	0.9018	0.7829	0.8585	0.6082
	CometKiwi	0.7963	0.7863	0.8742	0.8477	0.7210	0.8061	0.4162
Hunyuan-MT-Chemira-7B†	XCOMET-XXL	0.8974	0.8719	0.9306	0.9132	0.8268	0.8787	0.6089
	CometKiwi	0.8066	0.7914	0.8849	0.8514	0.7424	0.8129	0.4417

multiple translation outputs and employs a fusion model based on Hunyuan-MT-7B to aggregate these outputs through GRPO. The reward function comprises three primary elements: XCOMET-XXL scoring, DeepSeek-V3-0324 scoring, and a repetition penalty term. This multi-faceted

Prompt Template for Hunyuan-MT-Chimera-7B.

Analyze the following multiple <target_language> translations of the <source_language> segment surrounded in triple backticks and generate a single refined <target_language> translation. Only output the refined translation, do not explain.

The <source_language> segment:
```<source_text>```

The multiple <target_language> translations:
1. ```<translated_text1>```
2. ```<translated_text2>```
3. ```<translated_text3>```
4. ```<translated_text4>```
5. ```<translated_text5>```
6. ```<translated_text6>```

Table 5 | Prompt template of Hunyuan-MT-Chimera-7B.

reward mechanism ensures comprehensive evaluation of translation quality while mitigating redundancy in generated outputs. This methodology culminates in the development of our Hunyuan-MT-7B-Chimera model. Specifically, the prompt template is shown in Table 5.

During test-time inference, Hunyuan-MT-7B-Chimera accepts multiple translation candidates as input and synthesizes their respective strengths to produce a superior unified translation output. This aggregation approach leverages the complementary advantages of diverse translation hypotheses to achieve enhanced translation quality.

3.4. Main Results

As presented in Table 4, the experimental results demonstrate that Hunyuan-MT-7B and Hunyuan-MT-Chimera-7B consistently outperform most baselines across both XCOMET-XXL and CometKiwi metrics, indicating significant and robust improvements. We provide a detailed analysis of the Hunyuan-MT series models' performance across multiple dimensions below.

On the WMT24pp benchmark, Hunyuan-MT-7B achieves an XCOMET-XXL score of 0.8585, surpassing all baseline models, including ultra-large models such as Gemini-2.5-Pro (0.8250) and Claude-Sonnet-4 (0.8120). A particularly notable finding is the models' performance on Mandarin<=>Minority language pairs. Both Hunyuan-MT-7B (0.6082) and Hunyuan-MT-Chimera-7B (0.6089) achieve substantially higher scores than all competing systems, with the nearest competitor being Gemini-2.5-Pro at 0.5811. This represents an improvement of approximately 4.7% over the best ultra-large model and demonstrates improvements of 55-110% over translation-specialized models.

When compared to dedicated translation models (marked in purple), the Hunyuan-MT series demonstrates comprehensive superiority. It outperforms Google-Translator by 15-65% across different metrics and exceeds the Tower-Plus series (including the 72B variant) by 10-58% despite utilizing significantly fewer parameters. The models also surpass recent translation-optimized models, such as the Gemma-3 series, while maintaining superior generalization.

Hunyuan-MT-Chimera-7B exhibits systematic improvements across all metrics compared to the base model, achieving an average XCOMET-XXL improvement of 2.3% across all Flores-
```

200 directions. The most significant gains are observed in ZH=>XX (2.5% improvement) and XX=>XX (5.6% improvement). The model maintains strong performance in minority languages while enhancing the capabilities of mainstream languages. These improvements validate the effectiveness of the proposed approach in further optimizing translation quality.

4. Discussion

4.1. Case Study

Table 6, Table 7, and Table 8 illustrate representative translation cases, comparing the outputs of Hunyuan-MT-7B and Hunyuan-MT-Chemira-7B on key testing points. These translation cases demonstrate the model's sophisticated contextual understanding and proficiency in managing complex translations.

The cases presented in Table 5 demonstrate Hunyuan-MT-7B's superior capability in handling culturally-specific content and context-dependent translations across diverse scenarios. In Chinese social media contexts, Hunyuan-MT-7B correctly interprets "小红薯" as the platform "REDnote" and understands "砍一刀" as Pinduoduo's price-cutting mechanism. In contrast, Google-Translator produces literal but incorrect translations ("sweet potatoes" and "cuts"). For English slang expressions, Hunyuan-MT-7B accurately captures the idiomatic meanings (e.g., translating "You are killing me" as expressing amusement rather than literal harm), whereas Google-Translator fails to recognize the non-literal usage. The model also demonstrates stronger performance in domain-specific terminology translation, correctly rendering medical terms such as "blood disorders" and "uric acid kidney stones," and successfully translating complete addresses between languages, whereas Google-Translator leaves them unchanged. These examples illustrate that Hunyuan-MT-7B possesses a deeper understanding of linguistic nuances, cultural context, and domain knowledge, enabling it to produce more accurate and natural translations compared to conventional translation systems.

Table 6 | Comparative case studies across different scenarios. Red text highlights translation errors in Google-Translator outputs.

Scenario 1: Chinese Social	
Example #1	都知道小红薯在国外疯魔了，数据来告诉你有多逆天！
Testing Points	"小红薯": REDnote.
Huanyuan-MT-7B	Everyone knows that **REDnote** has become incredibly popular abroad, and the data is proof of just how remarkable its success is!
Google-Translator	Everyone knows that **sweet potatoes** are popular abroad. Here are the statistics to tell you how popular they are!
Example #2	没有一个外国人可以逃脱拼多多砍一刀
Testing Points	"砍一刀": Its core mechanism revolves around social sharing to reduce product prices.
Huanyuan-MT-7B	No foreigner can escape the **price cuts** offered by Pinduoduo.
Google-Translator	No foreigner can escape Pinduoduo's **cuts**
Scenario 2: English Slang	
Example #1	Oh!You are killing me!

Testing Points	"You are killing me": it is an informal and exaggerated English expression used to emphasize intense feelings like extreme amusement, frustration, or being overwhelmed.
Huanyuan-MT-7B	哦！你真要把我笑死了！
Google-Translator	噢！你要杀死我了！
Example #2	If you need advice on vintage vinyl, talk to Jim; he's the OG collector around here.
Testing Points	"OG": s a slang term originally short for "Original Gangster," now mainly used to describe a senior, authoritative, or pioneering person/thing in a field (with no negative connotation, often a term of recognition).
Huanyuan-MT-7B	如果你需要关于复古黑胶唱片方面的建议，可以找吉姆聊聊——他是这里最资深的收藏家。
Google-Translator	如果您需要有关老式黑胶唱片的建议，请与Jim 联系；他是这里的 **OG** 收藏家。
	Scenario 3: Terminology
Example #1	已知有血液疾病及尿酸性肾结石的患者不推荐使用本品，二岁以下儿童不得服用。
Testing Points	"血液疾病": blood disorder, "尿酸性肾结石": uric acid kidney stone.
Huanyuan-MT-7B	This product is not recommended for patients with known **blood disorders** or **uric acid kidney stones**, and it should not be taken by children under the age of two.
Google-Translator	Patients with known **blood diseases** and **uricidal** kidney stones are not recommended for use, and children under two years of age are not allowed to take it.
	Scenario 4: Place Name
Example #1	Unit 602, HUANGPU Complex, 654 Huangpu Drive, Huangpu District, Guangzhou, Guangdong Province, 510000, China
Huanyuan-MT-7B	中国广东省广州市黄埔区黄埔大道**654**号黄埔大厦**602**单元，邮编**510000**
Google-Translator	**Unit 602, HUANGPU Complex, 654 Huangpu Drive, Huangpu District, Guangzhou, Guangdong Province, 510000, China**

The cases in Scenarios 5 and 6 further demonstrate Hunyuan-MT-7B's superiority in handling less commonly used and ethnic minority languages. For European languages (Italian, German) and Asian languages (Korean, Persian), Hunyuan-MT-7B consistently produces more accurate and natural translations, correctly interpreting context-specific terms and avoiding literal translation errors. Most notably, in the translation of ethnic minority languages (Kazakh, Tibetan, and Cantonese), Hunyuan-MT-7B demonstrates exceptional performance, accurately translating complete sentences where Google-Translator either produces nonsensical output (Kazakh) or fails to recognize the language entirely (Cantonese). These results underscore the model's robust multilingual capabilities, particularly its unique strength in low-resource language translation, which aligns with the quantitative findings on the Mandarin<=>Minority benchmark.

The cases in Scenario 7 showcase Hunyuan-MT-Chemira-7B's enhanced capabilities compared to other models. In the first example, while other models fail to recognize "d2" as Diablo II or misinterpret "make a game" as game development, Hunyuan-MT-Chemira-7B correctly identifies

the gaming context and trading terminology. For handling informal language, it appropriately translates the intensifier "fucking" without resorting to literal profanity, demonstrating better pragmatic understanding. In the sports context, it accurately interprets "三分" as "three points" rather than mistranslating it as "winning three matches." Finally, it shows superior contextual awareness by translating "穿过" as "sped through" rather than the problematic "drove through" (which implies running into the crowd). These examples illustrate how the Chemira enhancement facilitates a more nuanced understanding of context, colloquialisms, and domain-specific terminology, yielding translations that are both more accurate and natural.

Table 7 | Comparative case studies across different scenarios. Red text highlights translation errors in Google-Translator outputs.

Scenario 5: Less Commonly Used Languages	
Example #1	Questo avviene anche in Norvegia, Svezia e Nuova Zelanda, ma in generale è un fatto piuttosto singolare (ad es. nei Paesi Bassi il rapporto è di uno a quaranta).
Testing Points	Italian => English
Huanyuan-MT-7B	This is also the case in Norway, Sweden, and New Zealand. However, overall it's a rather **unusual phenomenon**; for example, in the Netherlands the **ratio** is one to forty.
Google-Translator	This also occurs in Norway, Sweden and New Zealand, but in general it is a rather **singular fact** (e.g. in the Netherlands the **relationship** is one by forty).
Example #2	Mehrere Geiseln wurden gerettet und es gibt bisher mindestens sechs bestätigte Tote.
Testing Points	German => English
Huanyuan-MT-7B	Several hostages have been **rescued**, and there are at least six confirmed deaths so far.
Google-Translator	Several hostages have been **saved** and there are at least six confirmed deaths so far.
Input #3	무리는 1~3마리의 동족 성인 수컷과 무려 30마리에 이르는 암컷과 새끼로 구성됩니다.
Testing Points	Korean => Chinese
Huanyuan-MT-7B	这样的群体通常由1到3只同类的成年雄性，以及多达30只雌性和幼崽组成。
Google-Translator	一个狼群由一至三只同种成年雄性和最多30 只雌性及其幼崽组成。
Input #4	اگر از قطب شمال یا جنوب در زمستان بازدید کنید، شاهد شب قطبی خواهید بود، یعنی خورشید از خط افق بالاتر نمی‌آید.
Testing Points	Persian => Chinese
Huanyuan-MT-7B	如果您在冬季前往北极或南极，将会目睹极夜现象，也就是说太阳始终无法升出地平线。
Google-Translator	如果您在冬季参观北极或南极，您会看到一个极地夜晚，这意味着太阳不会高于地平线。
Scenario 6: Ethnic Minority Languages	
Example #1	كۆيەۋىڭسەنىڭبۇلايستەۋىڭقورسىلمايمادەپالاڭدايسىڭبا؟
Testing Points	Kazakh => Chinese
Huanyuan-MT-7B	你担心你的丈夫不同意你这样做吗?

Google-Translator	不欢迎**Kieuingulets**吗?
Example #2	ཁོ་ཚོར་བརྩི་འཇོག་བྱས་ན་གཞི་ནས་ཁོ་ཚོ་བསྒྱུར་བཀོད་བྱེད་ཐུབ
Testing Points	Tibetan => Chinese
Huanyuan-MT-7B	只有尊重他们，才能真正改变他们。
Google-Translator	只有尊重它们，才能进行改革。
Example #3	黄尸癫狗又出黎周围吠人了!
Testing Points	Cantonese => Chinese
Huanyuan-MT-7B	黄尸疯狗又出来到处吠人了！
Google-Translator	黄尸癫狗又出黎周围吠人了!

Table 8 | Case studies of the Hunyuan-MT-Chemira-7B model.

Scenario 7: Hunyuan-MT-Chemira-7B Case Study	
Input #1	I recently returned to d2 after several year, now I'm windering: Where do you guys sell/buy your stuff? Do you just make a game "O xxx N yyy" and hope for the best? Or are there a website that's more efficient?
Testing Points	"d2": Diablo II, "make a game "O xxx N yyy"": it means creating a game room where the room name uses "O" (short for "Offer," standing for the item xxx you want to sell) and "N" (short for "Need," standing for the item yyy you want in exchange) to inform other players of your trading needs.
Model #1	我最近几年后又回到了**d2**，现在我在想: 你们一般在哪里买卖装备？你们只是开个游戏房间"O xxx N yyy"，然后碰运气吗？还是有更高效的网站?
Model #2	我最近在几年后重新回到了**d2**，现在我想知道: 你们都在哪里买卖东西？你们是直接创建一个"O xxx N yyy"的游戏然后碰运气吗？还是有更高效的网站?
Model #3	我最近时隔几年又回来玩**d2**了，想问一下: 各位都是在哪里交易物品的？是自己建个名叫"出**xxx**换**yyy**"的房间然后随缘等，还是有更高效的网站?
Model #4	时隔多年重回暗黑2，现在我想知道: 你们都在哪买卖装备？是直接建个"**O**物品换**N**符文"的房间碰运气吗？还是有更高效的交易网站?
Model #5	我最近在离开几年后重新回到**d2**，现在我很疑惑: 你们都是在哪里买卖东西的？你们只是做一款游戏叫"O xxx N yyy"，然后听天由命吗？还是有更高效的网站?
Model #6	我最近在几年后回到了**d2**，现在我在想: 你们这些人在哪里卖/买你们的东西？你们只是做一个游戏"O xxx N yyy"然后希望是最好的吗？还是有一个更有效率的网站?
Hunyuan-MT-Chemira-7B	时隔多年，我又重新开始玩《暗黑破坏神**2**》（**Diablo II**），现在我想知道: 大家都是在哪里进行物品交易的？是通过自己创建名为**"O xxx N yyy"**的游戏房间来交易，然后听天由命吗？还是有更高效的交易网站?

Input #2	Now I obviously can't get out and shut it on the main road, so I pull down a few side streets and ensure that no one fucking knows I've quickly jumped out and shut it.I jump back in and drive home like nothing's happened.
Testing Points	"no one fucking knows": it means that absolutely no one is aware of, with the word "fucking" used here as a strong informal intensifier to emphasize the degree of secrecy—stressing that the action is completely unknown to anyone.
Model #1	显然我不能在主路上下车去关门，所以我驶进几条小路，确保没人他妈的知道我迅速下车把门关上了。我跳回车里，像什么都没发生一样开车回家。
Model #2	我显然不能在主路上下车去关门，所以我拐进了几条小巷，确保没有人知道我快速跳下车关了门。我跳回车里，像什么都没发生一样开车回家。
Model #3	我显然不能在主路上停车关门，于是我拐进几条小路，确保他妈的没人知道我赶紧跳下车把门关上。我再跳回车里，若无其事地开车回家。
Model #4	现在我显然没法在大马路上停车关门，所以我拐进几条小巷，确保没人他妈发现我迅速跳下车关好门。我跳回车里，若无其事地开回家。
Model #5	现在我显然不能在主干道上停下来关门，所以我拐进几条小街，确保没人他妈的知道我迅速跳下车把门关上。我跳回车上，像什么都没发生一样开车回家。
Model #6	显然我不能在主路上下车去关门，所以我驶进几条小路，确保没人他妈的知道我迅速下车把门关上了。我跳回车里，像什么都没发生一样开车回家。
Hunyuan-MT-Chemira-7B	显然我不能在主路上停车、下车去关门；于是我把车拐进几条小巷，确保没有人看到我迅速下车把门关上。之后我再跳回车里，若无其事地开车回家。
Input #3	希望我们能得到三分，不过那将是一场棘手的比赛，因为他们在上一场比赛中取得了不错的成绩，但是有这么多人支持我们，我相信我们一定可以继续前进，取得好成绩。
Testing Points	"三分": three points, not for winning three of the matches
Model #1	Hopefully, we can **win all three of those matches**. It's going to be a tough competition, though, because they performed well in their last match. But with so many people supporting us, I believe we can continue to do well and achieve good results.
Model #2	Hopefully, we can manage to **win three of these matches**. It's going to be a tough competition, though, because they performed well in their last match. But with so many people supporting us, I believe we can keep going and achieve good results.
Model #3	Hopefully, we can manage to **win three of these matches**. It's going to be a tough challenge, though, because they performed well in their last match. But with so many people supporting us, I'm confident that we can continue to move forward and **do well**.
Model #4	Hopefully, we can secure three points. It's going to be a tough match, because they performed well in their last game. However, with so many people supporting us, I believe we can continue to move forward and **do well**.
Model #5	Hopefully, we can secure three points. It's going to be a tough match, though; they performed well in their last game. But with so many people supporting us, I'm confident that we can continue to move forward and do well.
Model #6	Hopefully, we can secure three points. It's going to be a tough match, though; they performed well in their last game. But with so many people supporting us, I believe we can continue to move forward and **do well**.
Hunyuan-MT-Chemira-7B	I hope we can secure those **three points**, but it's going to be a tough match, as they performed well in their last game. However, with so many people supporting us, I'm confident that we can continue to move forward and **achieve good results**.
Input #4	一辆1948 年的福特水星汽车穿过佐治亚州门罗小镇的一群围观者，朝着小小的摩尔滩桥隆隆奔行。

Model #1	A 1948 Ford Mercury **drove through** a crowd of onlookers in the town of Monroe, Georgia, and then made its way **across** the small Moorestown Bridge.
Model #2	A 1948 Ford Mercury **drove through** a crowd of onlookers in the town of Monroe, Georgia, and then made its way **across** the small Moore's Bay Bridge.
Model #3	A 1948 Ford Mercury **convertible drove through** a crowd of onlookers in Monroe, Georgia, and then made its way **across** the small Moore's Bend Bridge.
Model #4	A 1948 Ford Mercury **drove through** a crowd of onlookers in the town of Monroe, Georgia, and then made its way **across** the small Moorestown Bridge.
Model #5	A 1948 Ford Mercury **driven by a man named George White** sped through a crowd in Monroe, Georgia, and then over the small Moore's Bay Bridge.
Model #6	A 1948 Ford Mercury **driven by a man named Steve Myers** sped through a crowd in Monroe, Georgia, and over the small bridge that crosses Moore Creek.
Hunyuan-MT-Chemira-7B	A 1948 Ford Mercury **sped through** a crowd of onlookers in the small town of Monroe, Georgia, **on its way towards** the tiny Moorestown Bridge.

4.2. The Role of CoT in MT

In the translation task, we also experimented with COT-based training methods. We found that if reward calculation is performed solely on the model's translation outputs, it ultimately fails to elicit an effective COT process. The model tends to generate only meaningless statements in the COT, such as "I need to translate the English text into Chinese and ensure the translation accurately conveys the original meaning." The translation performance shows no improvement compared to training without COT. Additionally, we explored simultaneously calculating rewards for both the translation results and the COT process, as demonstrated in our work on TAT-R1 (Li et al., 2025). We found that this approach enables the model to generate meaningful COT processes, resulting in improved translation performance.

4.3. Human Evaluation

Hunyuan-MT Challenge Testset, which encompasses a diverse range of scenarios, including social interactions, emails, food ordering, shopping, and navigation inquiries. This test set involves bidirectional translation between Chinese and English. Existing open-source test sets predominantly focus on the news domain. To comprehensively evaluate the translation capabilities, we construct a challenging test set that covers multiple domains, including news, medicine, government, literature, law, natural sciences, arts, computing, and the internet.

The experimental results reveal distinct performance tiers among the six evaluated translation models. The top-performing cluster, comprising Gemini-2.5-Pro (3.223), DeepSeek-V3-0324 (3.219), and Huanyuan-MT-7B (3.189), demonstrates marginal performance differences of less than 0.034 points, suggesting a convergence of state-of-the-art translation capabilities. Notably, Gemini-2.5-Pro exhibits exceptional bidirectional balance with nearly identical scores for ZH=>EN (3.225) and EN=>ZH (3.222) translations.

At the same time, most other models display a systematic bias favoring Chinese-to-English translation over the reverse direction, a phenomenon potentially attributable to the greater complexity of generating grammatically correct Chinese text. The significant performance gap between these leading models and Google-Translator (2.344), which underperforms by approximately 27%, underscores the superiority of modern transformer-based architectures

Table 9 | Human Evaluation. Here, we report the ZH=>EN and EN=>ZH translation results.

Models	ZH⇒EN	EN⇒ZH	Avg.
Gemini-2.5-Pro	3.225	3.222	3.223
DeepSeek-V3	3.253	3.203	3.219
Qwen3-32B	3.137	3.073	3.094
Google-Translator	2.841	2.101	2.344
Seed-X-PPO-7B	3.139	3.033	3.068
Huanyuan-MT-7B	3.258	3.155	3.189

over traditional translation systems. Particularly noteworthy is the competitive performance of Huanyuan-MT-7B, which, despite its relatively modest 7B parameters, achieves results comparable to larger models, suggesting that task-specific optimization can effectively compensate for model scale in specialized translation tasks.

5. Conclusion

In this paper, we present Hunyuan-MT, a family of open-source LLMs specifically designed for machine translation, supporting bidirectional translation across 33 languages. We provide a comprehensive description of our training pipeline, encompassing pre-training, instruction tuning, and reinforcement learning, while sharing key insights and best practices derived from our iterative optimization process. Notably, despite having only 7B parameters, Hunyuan-MT achieves translation quality comparable to—and in some cases exceeding—that of state-of-the-art LLMs and leading commercial translation systems, as demonstrated by both automatic metrics and human evaluation. By publicly releasing Hunyuan-MT's model weights, we aim to provide the research community with an accessible, high-performance foundation model to facilitate advances in machine translation research and applications.

6. Authors

6.1. Core Contributors

Mao Zheng, Zheng Li, Bingxin Qu, Mingyang Song, Yang Du, Mingrui Sun, Di Wang

6.2. Contributors

Tao Chen, Jiaqi Zhu, Xingwu Sun, Yufei Wang, Can Xu, Chen Li, Kai Wang, Decheng Wu

References

R. Agarwal, A. Singh, L. Zhang, B. Bohnet, L. Rosias, S. C. Y. Chan, B. Zhang, A. Anand, Z. Abbas, A. Nova, J. D. Co-Reyes, E. Chu, F. M. P. Behbahani, A. Faust, and H. Larochelle. Many-shot in-context learning. In A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/8cb564df771e9eacbfe9d72bd46a24a9-Abstract-Conference.html.

Anthropic. Introducing claude 4. https://www.anthropic.com/news/claude-4, 2025. Accessed: 2025-08-26.

D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In Y. Bengio and Y. LeCun, editors, *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.

P. F. Brown, J. Cocke, S. D. Pietra, V. J. D. Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin. A statistical approach to machine translation. *Comput. Linguistics*, 16(2):79–85, 1990.

P. F. Brown, S. D. Pietra, V. J. D. Pietra, and R. L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Comput. Linguistics*, 19(2):263–311, 1993.

C. Buck and P. Koehn. Findings of the wmt 2016 bilingual document alignment shared task. In *Proceedings of the First Conference on Machine Translation*, pages 554–563, Berlin, Germany, August 2016. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W/W16/W16-2347.

F. Cassano, J. Gouwar, D. Nguyen, S. Nguyen, L. Phipps-Costin, D. Pinckney, M. Yee, Y. Zi, C. J. Anderson, M. Q. Feldman, A. Guha, M. Greenberg, and A. Jangda. Multipl-e: A scalable and polyglot approach to benchmarking neural code generation. *IEEE Trans. Software Eng.*, 49(7): 3675–3691, 2023. doi: 10.1109/TSE.2023.3267446. URL https://doi.org/10.1109/TSE.2023.3267446.

S. Cheng, Y. Bao, Q. Cao, L. Huang, L. Kang, Z. Liu, Y. Lu, W. Zhu, J. Chen, Z. Huang, T. Li, Y. Li, H. Lin, S. Liu, N. Peng, S. She, L. Xu, N. Xu, S. Yang, R. Yu, Y. Yu, L. Zou, H. Li, L. Lu, Y. Wang, and Y. Wu. Seed-x: Building strong multilingual translation llm with 7b parameters, 2025. URL https://arxiv.org/abs/2507.13618.

K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, and J. Schulman. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168, 2021. URL https://arxiv.org/abs/2110.14168.

DeepMind. We're expanding our gemini 2.5 family of models. https://blog.google/products/gemini/gemini-2-5-model-family-expands/, 2025. Accessed: 2025-08-26.

DeepSeek-AI. Deepseek-v3 technical report, 2024. URL https://arxiv.org/abs/2412.19437.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

D. Deutsch, E. Briakou, I. Caswell, M. Finkelstein, R. Galor, J. Juraska, G. Kovacs, A. Lui, R. Rei, J. Riesa, S. Rijhwani, P. Riley, E. Salesky, F. Trabelsi, S. Winkler, B. Zhang, and M. Freitag. WMT24++: Expanding the Language Coverage of WMT24 to 55 Languages & Dialects, 2025. URL https://arxiv.org/abs/2502.12404.

A. Gu, B. Rozière, H. J. Leather, A. Solar-Lezama, G. Synnaeve, and S. Wang. Cruxeval: A benchmark for code reasoning, understanding and execution. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=Ffpg52swvg.

N. M. Guerreiro, R. Rei, D. van Stigt, L. Coheur, P. Colombo, and A. F. T. Martins. xcomet: Transparent machine translation evaluation through fine-grained error detection, 2023. URL https://arxiv.org/abs/2310.10482.

D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

B. Hu, A. Han, Z. Zhang, S. Huang, and Q. Ju. Tencent minority-mandarin translation system. In S. Huang and K. Knight, editors, *Machine Translation*, pages 93–104, Singapore, 2019. Springer Singapore. ISBN 978-981-15-1721-1.

W. Jiao, W. Wang, J. Huang, X. Wang, and Z. Tu. Is chatgpt A good translator? A preliminary study. *CoRR*, abs/2301.08745, 2023. doi: 10.48550/ARXIV.2301.08745. URL https://doi.org/10.48550/arXiv.2301.08745.

T. Kocmi and C. Federmann. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland, June 2023. European Association for Machine Translation. URL https://aclanthology.org/2023.eamt-1.19.

T. Kocmi, E. Avramidis, R. Bawden, O. Bojar, A. Dvorkovich, C. Federmann, M. Fishel, M. Freitag, T. Gowda, R. Grundkiewicz, B. Haddow, M. Karpinska, P. Koehn, B. Marie, C. Monz, K. Murray, M. Nagata, M. Popel, M. Popovic, M. Shmatova, S. Steingrímsson, and V. Zouhar. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In B. Haddow, T. Kocmi, P. Koehn, and C. Monz, editors, *Proceedings of the Ninth Conference on Machine Translation, WMT 2024, Miami, FL, USA, November 15-16, 2024*, pages 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.

Z. Li, M. Zheng, M. Song, and W. Yang. Tat-r1: Terminology-aware translation with reinforcement learning and word alignment, 2025. URL https://arxiv.org/abs/2505.21172.

C. Lin and L. Jackson. Assimilation over protection: rethinking mandarin language assimilation in china. *Multicultural Education Review*, 13(4):338–361, 2021. doi: 10.1080/2005615X.2021.2006117.

Q. Liu, X. Zheng, N. Muennighoff, G. Zeng, L. Dou, T. Pang, J. Jiang, and M. Lin. Regmix: Data mixture as regression for language model pre-training, 2025. URL https://arxiv.org/abs/2407.01492.

M. Luo, S. Tan, R. Huang, A. Patel, A. Ariyak, Q. Wu, X. Shi, R. Xin, C. Cai, M. Weber, C. Zhang, L. E. Li, R. A. Popa, and I. Stoica. Deepcoder: A fully open-source 14b coder at o3-mini level. https://pretty-radio-b75.notion.site/DeepCoder-A-Fully-Open-Source-14B-Coder-at-O3-mini-Level-1cf81902c14680b3bee5eb349a512a51, 2025. Notion Blog.

N. Muennighoff, Z. Yang, W. Shi, X. L. Li, L. Fei-Fei, H. Hajishirzi, L. Zettlemoyer, P. Liang, E. J. Candès, and T. Hashimoto. s1: Simple test-time scaling. *CoRR*, abs/2501.19393, 2025. doi: 10.48550/ARXIV.2501.19393. URL https://doi.org/10.48550/arXiv.2501.19393.

OpenAI. Introducing gpt-4.1 in the api. https://openai.com/index/gpt-4-1/, 2025. Accessed: 2025-08-26.

P. J. Ortiz Su'arez, B. Sagot, and L. Romary. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim, 2019. Leibniz-Institut f"ur Deutsche Sprache. doi: 10.14618/ids-pub-9021. URL http://nbn-resolving.de/urn:nbn:de:bsz:mh39-90215.

P. J. Ortiz Su'arez, L. Romary, and B. Sagot. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online, July 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.acl-main.156.

J. Pang, F. Ye, D. F. Wong, D. Yu, S. Shi, Z. Tu, and L. 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.

K. Papineni, S. Roukos, T. Ward, and W. 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*, pages 311–318. ACL, 2002. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv e-prints*, 2019.

R. Rei, M. Treviso, N. M. Guerreiro, C. Zerva, A. C. Farinha, C. Maroti, J. G. C. de Souza, T. Glushkova, D. M. Alves, A. Lavie, L. Coheur, and A. F. T. Martins. Cometkiwi: Ist-unbabel 2022 submission for the quality estimation shared task, 2022. URL https://arxiv.org/abs/2209.06243.

D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael, and S. R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. *CoRR*, abs/2311.12022, 2023. doi: 10.48550/ARXIV.2311.12022. URL https://doi.org/10.48550/arXiv.2311.12022.

A. Romanou, N. Foroutan, A. Sotnikova, Z. Chen, S. H. Nelaturu, S. Singh, R. Maheshwary, M. Altomare, M. A. Haggag, I. Schlag, M. Fadaee, S. Hooker, A. Bosselut, S. A, A. Amayuelas, A. H. Amirudin, V. Aryabumi, D. Boiko, M. Chang, J. Chim, G. Cohen, A. K. Dalmia, A. Diress, S. Duwal, D. Dzenhaliou, D. F. E. Florez, F. Farestam, J. M. Imperial, S. B. Islam, P. Isotalo, M. Jabbarishiviari, B. F. Karlsson, E. Khalilov, C. Klamm, F. Koto, D. Krzeminski, G. A. de Melo,

S. Montariol, Y. Nan, J. Niklaus, J. Novikova, J. S. Obando-Ceron, D. Paul, E. Ploeger, J. Purbey, S. Rajwal, S. S. Ravi, S. Rydell, R. Santhosh, D. Sharma, M. P. Skenduli, A. S. Moakhar, B. S. Moakhar, R. Tamir, A. K. Tarun, A. T. Wasi, T. O. Weerasinghe, S. Yilmaz, and M. Zhang. INCLUDE: evaluating multilingual language understanding with regional knowledge. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=k3gCieTXeY.

Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, M. Zhang, Y. K. Li, Y. Wu, and D. 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.

F. Shi, M. Suzgun, M. Freitag, X. Wang, S. Srivats, S. Vosoughi, H. W. Chung, Y. Tay, S. Ruder, D. Zhou, D. Das, and J. Wei. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=fR3wGCk-IXp.

M. Song, M. Zheng, Z. Li, W. Yang, X. Luo, Y. Pan, and F. Zhang. Fastcurl: Curriculum reinforcement learning with stage-wise context scaling for efficient training r1-like reasoning models, 2025a. URL https://arxiv.org/abs/2503.17287.

M. Song, M. Zheng, and X. Luo. Can many-shot in-context learning help llms as evaluators? A preliminary empirical study. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, and S. Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pages 8232–8241. Association for Computational Linguistics, 2025b. URL https://aclanthology.org/2025.coling-main.548/.

I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112, 2014. URL https://proceedings.neurips.cc/paper/2014/hash/a14ac55a4f27472c5d894ec1c3c743d2-Abstract.html.

K. Team, Y. Bai, Y. Bao, G. Chen, J. Chen, N. Chen, R. Chen, Y. Chen, Y. Chen, Y. Chen, Z. Chen, J. Cui, H. Ding, M. Dong, A. Du, C. Du, D. Du, Y. Du, Y. Fan, Y. Feng, K. Fu, B. Gao, H. Gao, P. Gao, T. Gao, X. Gu, L. Guan, H. Guo, J. Guo, H. Hu, X. Hao, T. He, W. He, W. He, C. Hong, Y. Hu, Z. Hu, W. Huang, Z. Huang, Z. Huang, T. Jiang, Z. Jiang, X. Jin, Y. Kang, G. Lai, C. Li, F. Li, H. Li, M. Li, W. Li, Y. Li, Y. Li, Z. Li, Z. Li, H. Lin, X. Lin, Z. Lin, C. Liu, C. Liu, H. Liu, J. Liu, J. Liu, L. Liu, S. Liu, T. Y. Liu, T. Liu, W. Liu, Y. Liu, Y. Liu, Y. Liu, Y. Liu, Z. Liu, E. Lu, L. Lu, S. Ma, X. Ma, Y. Ma, S. Mao, J. Mei, X. Men, Y. Miao, S. Pan, Y. Peng, R. Qin, B. Qu, Z. Shang, L. Shi, S. Shi, F. Song, J. Su, Z. Su, X. Sun, F. Sung, H. Tang, J. Tao, Q. Teng, C. Wang, D. Wang, F. Wang, H. Wang, J. Wang, J. Wang, J. Wang, S. Wang, S. Wang, Y. Wang, Y. Wang, Y. Wang, Y. Wang, Y. Wang, Z. Wang, Z. Wang, Z. Wang, C. Wei, Q. Wei, W. Wu, X. Wu, Y. Wu, C. Xiao, X. Xie, W. Xiong, B. Xu, J. Xu, J. Xu, L. H. Xu, L. Xu, S. Xu, W. Xu, X. Xu, Y. Xu, Z. Xu, J. Yan, Y. Yan, X. Yang, Y. Yang, Z. Yang, Z. Yang, Z. Yang, H. Yao, X. Yao, W. Ye, Z. Ye, B. Yin, L. Yu, E. Yuan, H. Yuan, M. Yuan, H. Zhan, D. Zhang, H. Zhang, W. Zhang, X. Zhang, Y. Zhang, Y. Zhang, Y. Zhang, Y. Zhang, Y. Zhang, Z. Zhang, H. Zhao, Y. Zhao, H. Zheng, S. Zheng, J. Zhou, X. Zhou, Z. Zhou, Z. Zhu, W. Zhuang, and X. Zu. Kimi k2: Open agentic intelligence, 2025a. URL https://arxiv.org/abs/2507.20534.

M. Team, X. Du, Y. Yao, K. Ma, B. Wang, T. Zheng, K. Zhu, M. Liu, Y. Liang, X. Jin, Z. Wei, C. Zheng, K. Deng, S. Jia, S. Jiang, Y. Liao, R. Li, Q. Li, S. Li, Y. Li, Y. Li, D. Ma, Y. Ni, H. Que, Q. Wang, Z. Wen, S. Wu, T. Xing, M. Xu, Z. Yang, Z. M. Wang, J. Zhou, Y. Bai, X. Bu, C. Cai, L. Chen, Y. Chen, C. Cheng, T. Cheng, K. Ding, S. Huang, Y. Huang, Y. Li, Y. Li, Z. Li, T. Liang, C. Lin, H. Lin, Y. Ma, T. Pang, Z. Peng, Z. Peng, Q. Qi, S. Qiu, X. Qu, S. Quan, Y. Tan, Z. Wang, C. Wang, H. Wang, Y. Wang, Y. Wang, J. Xu, K. Yang, R. Yuan, Y. Yue, T. Zhan, C. Zhang, J. Zhang, X. Zhang, X. Zhang, Y. Zhang, Y. Zhao, X. Zheng, C. Zhong, Y. Gao, Z. Li, D. Liu, Q. Liu, T. Liu, S. Ni, J. Peng, Y. Qin, W. Su, G. Wang, S. Wang, J. Yang, M. Yang, M. Cao, X. Yue, Z. Zhang, W. Zhou, J. Liu, Q. Lin, W. Huang, and G. Zhang. Supergpqa: Scaling LLM evaluation across 285 graduate disciplines. *CoRR*, abs/2502.14739, 2025b. doi: 10.48550/ARXIV.2502.14739. URL https://doi.org/10.48550/arXiv.2502.14739.

N. Team, M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, A. Sun, S. Wang, G. Wenzek, A. Youngblood, B. Akula, L. Barrault, G. M. Gonzalez, P. Hansanti, J. Hoffman, S. Jarrett, K. R. Sadagopan, D. Rowe, S. Spruit, C. Tran, P. Andrews, N. F. Ayan, S. Bhosale, S. Edunov, A. Fan, C. Gao, V. Goswami, F. Guzmán, P. Koehn, A. Mourachko, C. Ropers, S. Saleem, H. Schwenk, and J. Wang. No language left behind: Scaling human-centered machine translation, 2022. URL https://arxiv.org/abs/2207.04672.

J. Tiedemann. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218, 2012.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Y. Wang, X. Ma, G. Zhang, Y. Ni, A. Chandra, S. Guo, W. Ren, A. Arulraj, X. He, Z. Jiang, T. Li, M. Ku, K. Wang, A. Zhuang, R. Fan, X. Yue, and W. Chen. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/ad236edc564f3e3156e1b2feafb99a24-Abstract-Datasets_and_Benchmarks_Track.html.

J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL https://arxiv.org/abs/2201.11903.

Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

Q. Zhang, F. Lyu, Z. Sun, L. Wang, W. Zhang, W. Hua, H. Wu, Z. Guo, Y. Wang, N. Muennighoff, I. King, X. Liu, and C. Ma. A survey on test-time scaling in large language models: What, how, where, and how well?, 2025. URL https://arxiv.org/abs/2503.24235.

W. Zhu, H. Liu, Q. Dong, J. Xu, S. Huang, L. Kong, J. Chen, and L. Li. Multilingual machine translation with large language models: Empirical results and analysis. In K. Duh, H. Gómez-Adorno, and S. Bethard, editors, *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 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.

7. Appendix

7.1. Supported Languages

Languages	Abbr.	Languages	Abbr.	Languages	Abbr.
Chinese	zh	Malay	ms	Marathi	mr
English	en	Indonesian	id	Hebrew	he
French	fr	Filipino	tl	Bengali	bn
Portuguese	pt	Hindi	hi	Tamil	ta
Spanish	es	Traditional Chinese	zh-Hant	Ukrainian	uk
Japanese	ja	Polish	pl	Tibetan	bo
Turkish	tr	Czech	cs	Kazakh	kk
Russian	ru	Dutch	nl	Mongolian	mn
Arabic	ar	Khmer	km	Uyghur	ug
Korean	ko	Burmese	my	Cantonese	yue
Thai	th	Persian	fa		
Italian	it	Gujarati	gu		
German	de	Urdu	ur		
Vietnamese	vi	Telugu	te		

Table 10 | Supported languages.