||| ByteDance | Seed

Seed-X: Building Strong Multilingual Translation LLM with 7B Parameters

ByteDance Seed

Full author list in Contributions

Abstract

Multilingual translation stands as a challenging task for large language models (LLMs) to handle intricate language patterns and stilted translations that arise in automated translations. In this paper, we introduce **Seed-X**, a family of open-source LLMs comprising instruct and reasoning models, pushing the limits of translation capability with 7B parameter size. The base model is pre-trained on a diverse, high-quality dataset encompassing both monolingual and bilingual content across 28 languages, harnessing the full potential of multilingual data. The instruct model is then finetuned to translate by Chain-of-Thought (CoT) reasoning and further enhanced through reinforcement learning (RL) to achieve better generalization across diverse language pairs. SEED-X achieves performance comparable to leading closed-source models, including Gemini-2.5 and GPT-4o, across 28 languages, and significantly outperforms larger open-source models in both automatic metrics and human evaluations. We share the best practices through our optimization process, and make the parameter public available for advancing translation research and applications.
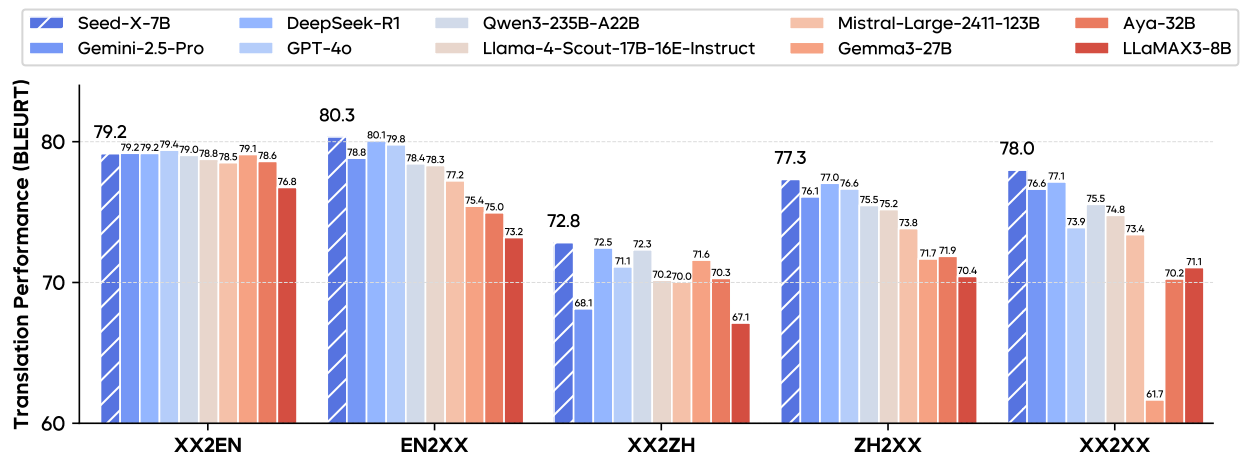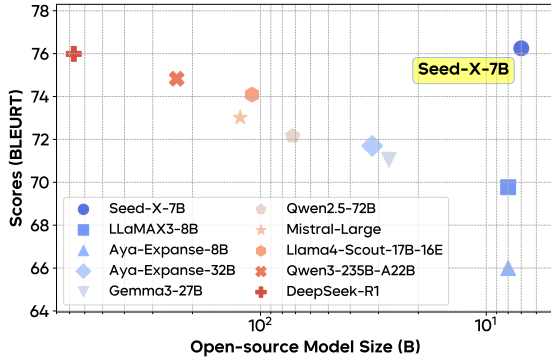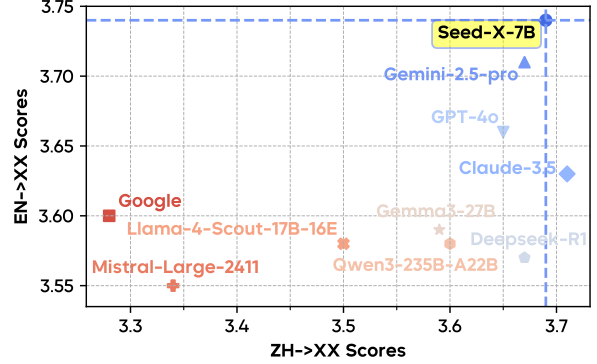
Figure 1 Benchmark performance on Flores-200 of SEED-X and its counterparts.

Figure 2 **Model size** versus average **multilingual translation performance** for different-scale open-source models. Scores averaged over directions and test sets (see Table 4).

Figure 3 **Human evaluation** (0-4) of models on **Seed-X-Challenge** for Chinese/English to 7 languages translation (detailed scores are presented in Appendix D).

1 Introduction

Machine translation has long been a high-demand practical goal and a significant research challenge pursued over the past few decades [1–4]. Recently, the development of large language models (LLMs) has revolutionized the learning paradigm of machine translation, shifting it towards large-scale learning approaches [5–8]. The continuous evolution of LLMs has consistently pushed the boundaries of translation quality, with state-of-the-art models like Claude-3.5 [9] even surpassing human experts on certain translation directions [10].

However, the challenge has not been completely resolved, such as the translation of idioms, slang, and internet buzzwords [11]. This requires not only robust language comprehension but also the ability to produce culturally appropriate and natural expressions, going beyond mere word-for-word mapping between languages. Moreover, a persistent performance gap exists between proprietary and open-source models [10], particularly given the typically limited scale of open-source systems. The lack of clear methodologies for developing cutting-edge LLM-based translation systems continues to hinder community efforts to deploy and iterate effective translation solutions.

In this paper, we present SEED-X, a comprehensive family of 7B-parameter open-source translation models. The suite encompasses an instruct model and an RL model together with the corresponding reward model. The entire training process is designed to address the challenges of cross-lingual learning and enhance both computational efficiency and translation performance. First, the models undergo pre-training on a large-scale dataset comprising 28 languages from scratch, incorporating both monolingual and parallel corpora to ensure comprehensive linguistic coverage. Subsequently, the instruct model learns to implement chain-of-thought reasoning to facilitate the translation of high-resource languages and finally generalizes to a broader language spectrum by reinforcement learning.

We conduct extensive evaluations on various representative translation benchmarks and compare it with both open-source models and closed-source systems. SEED-X demonstrates superior performance compared to translation-specialized models of similar size and Google-Translator. It further outperforms numerous larger-scale LLM systems (Figure 2), including Gemma3-27B [12], Llama4-Scout-17B-16E [13] and Qwen3-235B-A22B [14].

Human evaluation in Figure 3 shows that its performance matches or even surpasses that of cutting-edge models including GPT-4o [15], Claude-3.5 [9], DeepSeek-R1 [16], and Gemini-2.5-Pro [17]. Overall, we set a higher-level baseline for machine translation research based on open-source 7B models by providing:

- **A family of Seed-X models** including the instruct model, RL model, and reward model.

- **A diverse and challenging translation testset** comprising English/Chinese⇒X idioms, proverbs, poetry, internet slang, and abbreviations across various domains, accompanied by annotated translation key

points.

- **A detailed training recipe** for developing translation-oriented LLMs, along with insights into multilingual knowledge transfer mechanisms.

2 Pre-training

This section describes the construction of our pre-training dataset, which incorporates both monolingual and bilingual corpora to facilitate effective multilingual modeling and cross-language semantic alignment.

2.1 Monolingual Data

The monolingual data comprises a diverse blend of languages with a total of 6 trillion tokens (see Table 1). Our data pre-processing pipeline is designed to optimize both data quality and linguistic diversity across language corpora while minimizing redundancy. We collect extensive monolingual data from various sources, implement document tagging, and balance the topic distribution across the corpus [18]. To maintain high data quality, we develop a document-level quality assessment model that categorizes documents into three tiers: high, medium, and low quality. High-quality documents are retained in their original form, medium-quality documents are enhanced through LLM-based paraphrasing, and low-quality content is excluded entirely. This quality control process is implemented iteratively to ensure optimal results.

Language	Prop. (%)	Tokens (T)
English	12.94	0.78
Chinese	10.54	0.63
Russian	8.32	0.49
French	7.36	0.44
Spanish	7.04	0.42
German	6.39	0.38

Table 1 The statistics on high-resource languages in monolingual data.

Besides, we do not spend too much effort on mathematics and code data, allocating the entire data share to multilingual data. We deliberately exclude STEM, coding, and reasoning-focused data, which consume training capacity without contributing significantly to translation quality. Thus, we can expect that the base model's capabilities in the field of code will be limited.

2.2 Bilingual Data

Bilingual data plays a crucial role in semantic alignment and can significantly improve translation capabilities. We build the multilingual bilingual data in a progressively weak-to-strong fashion using the boosting SEED-X model.

- **Collecting Seed Bilingual Data.** We create an initial seed bilingual dataset of 200B tokens from publicly available web data. The filtering process employs strict rules based on language identification confidence scores and bilingual alignment metrics from word alignment tools.

- **Building Early Translation Model.** We used the seed bilingual dataset to train an initial translation model, based on either an open-source model or our boosting pretraining SEED-X model. This early model is then fine-tuned for data augmentation, filtering, and rewriting.

- **Augmenting and Refining Bilingual Data.** To further expand the data volume and improve the data quality, we utilize the temporary translation model in the previous phase to construct a new version of bilingual data: (1) directly translating the existing monolingual data to different languages to construct pseudo parallel data, (2) rewriting and paraphrasing the first-round translation, and (3) evaluating and filtering the poor-quality pairs.

- **Repeating the Above Process.** Each iteration of this process progressively enhances the quality of the bilingual data, which is then used to train subsequent model versions. Concurrently, we replace the early translation model in step 2 with more advanced variants to generate successive versions of higher-quality parallel data. Thus, the data quality gradually improves as increasingly sophisticated models are employed.

Model	Exam		Understanding				Avg.
	INCLUDE	MMMLU	XCOPA	XWinograd	XStoryCloze	PAWS-X	
Llama3-8B [20]	55.27	50.60	72.03	80.25	65.45	60.85	64.08
Qwen2.5-7B [21]	61.29	**59.59**	72.73	79.94	65.76	62.88	67.03
Seed-X	**61.93**	53.25	**77.37**	**82.48**	**69.72**	**63.83**	**68.10**

Table 2 Performances of pretrained model on multilingual benchmarks.

2.3 Pre-training Stage

Given the monolingual and bilingual data introduced in Section 2.1 and 2.2, the next question is how to organize these datasets, which differ greatly in the volume and distribution of data between languages. Our core strategy focuses on first acquiring fundamental knowledge represented across major languages, then transferring this knowledge to develop multilingual capabilities. The training process follows three key stages:

- **S1: General Stage.** We train the model on large-scale monolingual data from predominant languages, as detailed in Table 1. The primary languages of the model are Chinese and English.

- **S2: Multilingual-dominant Stage.** We gradually increase the proportion of multilingual data, incorporating both monolingual data from underrepresented languages and bilingual corpora.

- **S3: Parallel-Only Stage.** We fine-tune the model using high-quality bilingual data that has undergone multiple rounds of rewriting and filtering to achieve optimal performance. Specifically, the parallel data is simply combined with language tags, without incorporating any additional instructions or prompts.

We build our 7B model based on the Mistral-7B architecture [19] and expand the vocabulary size to 65,269 tokens. To achieve optimal model performance, we develop scaling laws to predict optimal hyper-parameters, such as learning rate scheduler and batch size, specifically tailored for our three pre-training stages. More details about the comprehensive experimental settings and configurations can be found in Appendix B.

2.4 Pre-training Evaluation

We evaluate the multilingual capabilities of Seed-X after S2, with a focus on knowledge-based tasks rather than mathematical or logical reasoning assessments. Our evaluation framework encompasses two key benchmarks: (1) regional knowledge assessment using INCLUDE [22], general knowledge evaluated by MMMLU [23], and (2) comprehension tasks evaluated through XCOPA [24], XWinograd [25], XStoryCloze [26], and PAWS-X [27]. For MMMLU benchmarks, we sample only 10% of the original data to improve evaluation efficiency. We only report the average scores of languages we support. It's worth noting that S3 employs training data with a particular format that guides Seed-X towards becoming a specialized model. This specialization may have implications for the model's general capabilities, as discussed further in Section 5.

As shown in Table 2, Seed-X demonstrates strong performance in multilingual comprehension tasks, which serves as a fundamental foundation for translation capabilities. Seed-X exhibits strong performance in cultural and region-specific knowledge domains, yet demonstrates relatively lower proficiency in mathematical reasoning tasks within MMMLU.

3 Post-training

In the post-training, we aim to equip the base model with robust translation capabilities through supervised fine-tuning (SFT) and preference learning. The central challenge lies in optimizing performance for high-resource languages using human-annotated data while ensuring effective generalization to low-resource languages.

3.1 Supervised Fine-tuning

We curate our instruction-tuning datasets to include 236K instances to only support translation without general requirements. The dataset draws from two primary sources: the public FLORES devset [28] and

Prompts	Types
Translate the following text from `<src>` to `<trg>`:__ What does this sentence mean in `<trg>` from `<src>`:__ How do you translate this sentence into `<trg>` from `<src>`:__ Translate the following text to `<trg>`:__	Standard
Translate the following `<src>` sentence into `<trg>` and explain it in detail:__ First translate the `<src>` text into `<trg>` and then give the explanation:__ Translate the following sentence into `<trg>` and try to explain this translation. The input is:__	CoT

Table 3 A few examples of prompt templates.

manually annotated translation pairs encompassing both general domains and specific business scenarios. To ensure data quality, we employ G-DIG [29] to eliminate problematic examples that contribute most significantly to translation errors. Additionally, we implement rejection sampling [30] to further refine the instruction data, with particular emphasis on low-resource languages where manual annotations are scarce.

Chain-of-thought Data. As aforementioned, translation is a complex task that requires reasoning capabilities to comprehend both semantic nuances and cultural context of the source text. To capture this reasoning process, we engage professional linguists to document their thought process when handling challenging translation cases. Volunteers are required to annotate the following: (1) the overall sentence meaning, (2) interpretations and translations of challenging linguistic elements (including slang, internet buzzwords, and poetic expressions), (3) notable target language expressions and usage conventions, and common potential translation pitfalls and errors. We design specific CoT prompts to lead the model to translate with the reasoning process as seen in Table 3.

3.2 MT-oriented Preference Learning

Large-scale RL has demonstrated effectiveness in enhancing reasoning capabilities for code and mathematical tasks, where outputs follow structured formats [16]. However, MT presents a unique challenge as its outputs are semantically diverse and resist evaluation through explicit rule-based assessment. For this, we define two distinct types of rewards for different scenarios:

- **Human Preference-based Reward.** Leveraging human preference data, we train a reward model based on SEED-X-base to assign scalar scores to candidate translations. Given the substantial challenges in annotation, including the availability of qualified annotators and associated costs, we focus on preference data collection on high-resource language pairs with approximately 20k pairs.

- **Dual-based Reward without Reference Answer.** When the preference data are unavailable, we follow DuPO (DuPO's article will be published soon.) to employ the dual-based reward to estimate the quality of translation without the need for annotated preference data or reference translations. Specifically, we perform $A \Rightarrow B \Rightarrow \widetilde{A}$ and measure the similarity between the A and $\widetilde{A}$ to score the reward of $A \Rightarrow B$.

It is noted that we avoid using automatic evaluation metrics (e.g., BLEU [31], BLEURT [32] or COMET [33–35]) as rewards. These metrics partially reflect translation quality, but are difficult to entirely equivalent to human judgment, especially in assessing subtle linguistic features and contextual appropriateness.

RL Algorithm. We optimize the model parameters using the Proximal Policy Optimization (PPO) algorithm [36]. We employ a large batch size combined with multiple rollouts per query for training efficiency and initialize the critic model with the reward model to ensure training stability.

4 Translation Evaluation

4.1 Settings

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

- **FLORES and WMT Testsets.** From FLORES-200 [4], we selected 756 translation pairs across 28 languages (listed in Appendix A), which were systematically categorized into five groups: English⇒XX, XX⇒English, Chinese⇒XX, XX⇒Chinese, and XX⇒XX translations. Additionally, we incorporated dev sets from WMT-25[1], which include English-to-X translations spanning 25 target languages.

- **Seed-X-Challenge Testset.** Unlike existing datasets that mainly focus on simple news translation, we aim to further challenge and evaluate the boundaries of LLMs' translation capabilities. To this end, we develop a challenging testset containing diverse linguistic elements from internet slang and classical literature to idioms and proper nouns. The dataset spans multiple domains (healthcare, science, tourism, education, e-commerce, entertainment, etc.) and features both formal and colloquial styles. Each entry contains source texts in Chinese or English, with expert annotations highlighting potential translation difficulties. We test translation from Chinese/English to seven languages: Spanish, German, French, Russian, Arabic, Portuguese and Italian.

Evaluation. We evaluate translations through two approaches: automatic metrics and human assessment. For automatic evaluation, we employ neural-based metrics XCOMET-XL [33] and BLEURT [32], which generally correlate well with human judgements but may be unreliable in detecting certain translation phenomena. Therefore, we conduct human evaluation where multilingual experts score translations on a 0-4 scale, focusing on previously annotated error-prone points and considering accuracy, fluency, and idiomaticity.

4.2 Main Results

Table 4 demonstrates the competitive performance of our SEED-X on FLORES-200 and WMT-25 benchmarks. For comparison, we categorize baseline models into **three tiers**: **(1) ultra-large models** (e.g., Gemini-2.5, Claude-3.5, GPT-4o and so on), **(2) medium to small-sized models** such as Gemma-3-27B and Aya-32B, and **(3) translation-specialized models** like LLaMAX3-8B and TowerInstruct-13B. It is worth noting that we only evaluate the highest-performing variant from each model series for comparison. Google Translator is the only exception due to its undisclosed model architecture and serving infrastructure.

Super large models dominate the top performance rankings in translation quality, maintaining a significant margin over other models. Their superiority is particularly pronounced in low-resource translation pairs, extending beyond the Chinese-English centered directions. The performance hierarchy among the three tiers is: **Tier 1 ≫ Tier 2 ≈ Tier 3**. It also shows that previous specialized translation models, despite their task-specific design, remain constrained by their relatively smaller model sizes, leading to suboptimal performance.

According to automatic evaluation metrics, SEED-X, despite its 7B parameter scale, performs comparably to Tier 1 models and substantially outperforms Tier 2 and Tier 3 models. This shifts the performance hierarchy to **Tier 1 ≈ Seed-X ≫ Tier 2 ≈ Tier 3**, demonstrating a significant advancement in translation model capabilities.

Human Evaluation. Figure 3 further demonstrates the effectiveness of SEED-X across 7 widely-used translation language pairs, where professional linguists evaluated translations between Chinese/English and Spanish/German/French/Russian/Portuguese/Italian/Arabic. SEED-X demonstrates superior performance in two aspects: it leads EN⇒XX evaluations, outperforming even ultra-large models, while achieving the second-best results in ZH⇒XX, significantly ahead of mid-tier competitors. Although closed-source large language models typically outperform open-source ones, Seed-X has successfully challenged this dominance. Notably, while Google's system achieves high scores in automatic metrics, it performs relatively poorly in human evaluations (showcased in Appendix E). This discrepancy clearly illustrates the limitations of automatic evaluation metrics and validates our motivation to avoid using them directly as rewards in RL.

4.3 Ablation

The Effect of Quality and Diversity of Monolingual Data. This section explores our monolingual data cleaning strategy through a toy experiment on a 1.3B-parameter model, evaluated using two methods: perplexity-based evaluation [41] on the translated ARC dataset [42] ("ARC-mm") and generation-based evaluation on the translation dataset with 5-shot as prompts ("Trans."). As shown in Table 5a, both the filtering and

---

[1]https://www2.statmt.org/wmt25/

Models	Metrics	FLORES-200					WMT-25	Avg.
		XX⇒EN	EN⇒XX	XX⇒ZH	ZH⇒XX	XX⇒XX	EN⇒XX	
TowerInstruct-13B[†] [8]	BLEURT	76.18	68.58	66.08	51.03	56.38	56.35	62.43
	COMET	93.15	79.65	78.81	83.44	81.20	66.03	80.38
LLaMAX3-8B[†] [37]	BLEURT	76.75	73.19	67.12	70.43	71.06	60.07	69.77
	COMET	95.13	86.22	78.09	83.85	83.42	70.89	82.93
Google-Translator	BLEURT	80.13	81.07	73.81	74.84	76.93	71.07	76.31
	COMET	96.98	94.76	85.67	89.64	89.96	84.12	90.19
InternLM3-8B-Instruct[†] [38]	BLEURT	70.37	56.91	62.00	50.64	40.30	47.54	54.63
	COMET	85.40	63.21	71.74	59.77	58.75	53.56	65.41
Aya-32B[†] [39]	BLEURT	78.60	74.95	70.29	71.86	70.25	64.25	71.70
	COMET	96.04	87.36	82.34	84.49	82.51	75.37	84.69
Gemma3-27B[†] [12]	BLEURT	79.09	75.42	71.59	71.67	61.66	66.86	71.05
	COMET	96.67	92.71	84.00	89.30	85.11	81.90	88.28
Qwen-2.5-72B-Instruct[†] [21]	BLEURT	78.79	75.15	71.14	70.07	72.59	65.23	72.16
	COMET	96.64	88.65	83.79	84.84	84.72	76.86	85.92
Mistral-Large-2411-123B[†] [40]	BLEURT	78.51	77.22	70.02	73.83	73.40	65.16	73.02
	COMET	96.28	91.18	82.23	88.02	86.31	78.11	87.02
Llama-4-Scout-17B-16E[†] [13]	BLEURT	78.75	78.31	70.16	75.18	74.77	67.37	74.09
	COMET	96.25	92.45	83.11	89.32	87.80	80.44	88.23
Qwen3-235B-A22B[†] [14]	BLEURT	79.03	78.43	72.32	75.47	75.55	68.19	74.83
	COMET	96.75	92.66	85.38	89.55	89.00	81.45	89.13
Gemini-2.5-Pro [17]	BLEURT	79.18	78.83	68.13	76.08	76.63	69.23	74.68
	COMET	96.52	93.16	79.79	90.59	89.80	83.17	88.84
GPT-4o [15]	BLEURT	79.40	79.78	71.11	76.64	73.90	70.08	75.15
	COMET	97.08	94.20	83.04	91.56	90.08	83.47	89.90
DeepSeek-R1[†] [16]	BLEURT	79.17	80.06	72.45	77.05	77.14	71.06	76.16
	COMET	96.89	94.53	85.27	91.43	90.77	84.82	90.62
Claude-3.5-Sonnet [9]	BLEURT	79.66	80.53	72.65	77.76	78.14	71.06	76.63
	COMET	97.17	94.88	85.47	92.36	91.71	84.82	91.07
Our Systems								
Seed-X-Instruct[†]	BLEURT	77.77	79.46	72.16	75.51	76.38	68.90	75.03
	COMET	95.52	93.51	85.54	89.18	88.86	81.69	89.05
Seed-X-DuPO[†]	BLEURT	78.57	80.13	72.57	77.05	77.35	69.32	75.83
	COMET	96.55	94.03	85.76	91.02	90.43	81.99	89.96
Seed-X-PPO[†]	BLEURT	79.15	80.34	72.83	77.32	77.98	69.88	76.25
	COMET	96.98	94.60	87.24	91.80	91.34	82.89	90.81

Table 4 Performances of different models on FLORES-200 and WMT-25. We report the English-centric (EN⇒XX and XX⇒EN), Chinese-centric (ZH⇒XX and XX⇒ZH) and the overall XX⇒XX performances of Seed-X and prominent existing systems. Systems with open-source weights are marked with †. Baseline models are categorized into three groups: (1) ultra-large general models , (2) medium to small-sized general models , and (3) translation-specialized models .

paraphrasing operations significantly reduce the training loss and improve the model performance on the multilingual ARC datasets, which also greatly enhance the LLMs' cross-lingual generalization ability. To measure diversity, we quantify it using the number of unique tags, as shown in Table 5b. The model's

Model	Loss	Arc-mm	Trans.
Raw Data	3.23	25.17	6.97
+ Filter	2.93	30.18	8.19
+ Paraphrasing	**2.81**	**32.98**	**9.41**

(a) Impact of quality improvements in pre-training data.

Model	Loss	Arc-mm	Trans.	# Tag (M)
Baseline	2.81	**32.98**	**9.41**	**1.55**
2/3 Diversity	2.73	29.71	8.55	1.04
1/3 Diversity	**2.51**	26.51	7.67	0.51

(b) Effect of diversity in pre-training data.

Table 5 Performance improvements achieved through monolingual data quality enhancement and diversity selection strategies. We observe significant gains in translation ability as the quality of monolingual data improves.

(a) Enc-Dec Model

(b) Enc-Dec Model

(c) LLM
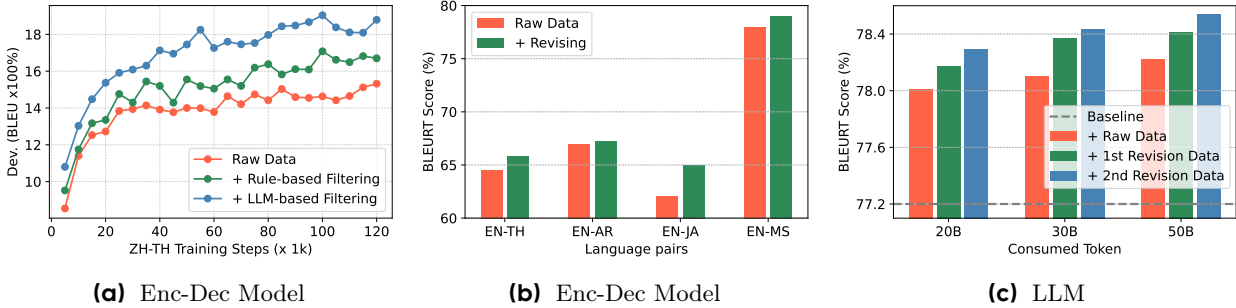
Figure 4 Performance gains of revising and filtering parallel data on encoder-decoder models and LLMs.

reasoning and translation abilities significantly decline as the number of unique tags decreases, highlighting the importance of maintaining diversity in the pre-training dataset.

The Effect of Filtering and Revising Parallel Data. Parallel corpora serve as the most direct supervision signals for translation tasks. Therefore, we design several rounds of filtering and rewriting to progressively improve the quality of parallel corpora. The base model of both filter and rewriter is updated each round to better detect and correct translation errors. To evaluate our approach, we first experiment with Transformer models (13-layer encoder, 3-layer decoder) for each language pair. Figure 4a shows our LLM-based filtering outperforms rule-based methods in identifying noisy pairs. Figure 4b demonstrates that our revision method further improves parallel data quality. To assess the impact on LLM performance, we train models on different scales of revised data. Results in Figure 4c confirm the effectiveness of this multi-round strategy in improving LLM translation performance.

The Effect of Prompt Design. As a multitasking process, multilingual translation relies heavily on effective prompt design to enhance the model's generalization capabilities. Our findings, as shown in Table 6, demonstrate that omitting the source language in prompts does not impact performance, while making the model more robust when handling mixed-language inputs. Besides, improving the diversity of prompts (as shown in Table 3) and introducing CoT both greatly enhance the translation quality.

The Effect of Delimiters. Motivated by [43, 44], we use language tags `<LANG_CODE>` (`<EN>`, `<ZH>`, `<JA>`, etc.) as prefixes and concatenate bilingual data for continued training. Table 7 demonstrates that the incorporation of language tags, similar to language names, in organizing bilingual data can significantly enhance translation quality when compared to other delimiters. Such language-aware delimiters contains much more information than `<SEP>`, explicitly helping the model distinguish between languages. However, combining bilingual data according to the natural language of the prompt (denoted as `<NL>`) is formally closer to monolingual data, it introduces a large amount of low-information prompt text.

SFT	EN⇒XX	XX⇒EN
BASELINE	75.84	75.66
w/ NO [SRC] LANG.	75.82	75.89
w/ COT PROMPTS	77.48	77.58
w/ DIVERSED PROMPTS	77.17	77.5
w/ ALL	**77.88**	**77.74**

Table 6 Ablation of different prompts in SFT.

Delimiter	EN⇒XX	XX⇒EN
<SEP>	53.39	74.97
<NL>	74.81	75.47
<LANGUAGE>	77.66	77.75
<LANG CODE>	**78.31**	**77.86**

Table 7 Comparison of different delimiters used for concatenating parallel data. `<NL>` denotes using translation instruction written in natural language.

Mono.	CoT	BLEURT		Factual Acc.
		ZH⇒EN	EN⇒ZH	
✗	✗	62.22	64.00	-
✗	✓	63.22	64.70	59.1
✓	✗	63.55	65.88	-
✓	✓	**64.59**	**66.62**	**67.7**

Table 8 Translation quality on in-house ZH⇔EN testsets with different settings. "w/o Mono." means the model is purely trained on concatenated parallel data during pretraining, "CoT" stands for the model is trained with a chain of task: translate-then-explain. We ask annotators to check whether the model makes factual errors or misunderstands the semantics of the given sentence based on generated explanations as factual accuracy.

5 Empirical Insights

5.1 Monolingual Data Shapes Core Language Capabilities

LLMs have fundamentally transformed machine translation, primarily through their advanced capabilities derived from massive monolingual data. While previous studies [8, 45] have demonstrated MT capabilities through direct finetuning of open-source LLMs, the specific impact of pretraining remains unclear. In this section, we perform preliminary experiments on ZH⇔EN translation with a 1.3B model. We lead the model to generate explanations together with translations to see how pretraining on large-scale monolingual data enhances LLM's understanding of the given sentence and further improves translation quality.

While monolingual data's primary contribution to translation tasks is often assumed to be improved fluency in generation, we also investigate several additional potential hypotheses as below:

- ☒ **Unlocking Natural Language Instruction Understanding**. Results show that models trained solely on parallel data can effectively follow diverse instructions and generate explanations after translation-oriented fine-tuning. This demonstrates that **the ability to follow natural language instructions is not exclusively acquired from monolingual data**.

- ☑ **Enriching Factual Knowledge Base**. We ask annotators to evaluate the model's factual accuracy and semantic understanding based on its generated explanations. Results show that 200B monolingual data improves the factual accuracy from 59.1% to 67.7%. Table 9 shows that "model w/ Mono." is able to handle complex contexts with typos ("**feveryone**" in the first case) and translate the terminology correctly with additional knowledge ("**Lark 4.1**" in the second case). This improvement suggests that large-scale monolingual data significantly enhances LLMs' ability to comprehend source sentences.

- ☑ **Improving Reasoning Capabilities**. Results in Table 8 show that even the "model w/ Mono." unexpectedly benefits from CoT patterns. However, the reasoning ability derived from parallel data remains limited, primarily due to the absence of code and mathematical data.

It is noted that parallel data also develops the above abilities that contributes to translation, but acquiring large-scale parallel data remains challenging. Therefore, a practical alternative is to leverage large amounts of accessible monolingual data for pretraining and then transfer the acquired knowledge to MT tasks.

Prompt	Model w/o Mono.	Model w/ Mono.
Translate the sentence to Chinese and try to explain: Thank you **feveryone** for helping us grow spiritually!	感谢**Youfeone** 帮助我们在精神上成长！(This sentence is a polite way to thank **someone named Youfeone** for helping the speaker grow spiritually.) ✗	感谢你们帮助我们在精神上成长！(This sentence is a thank you message to someone who has helped the speaker to grow spiritually.) ✓
Translate the sentence to Chinese and explain it: From **Lark 4.1** onwards, you can now add up to 10 group chat administrators.	从**4.1**开始，您现在可以添加最多10名群聊管理员。(This sentence is related to adding group chat administrators to a group chat. **It starts with the date 4.1** and describes the action that can be done from that date.) ✗	从**Lark 4.1**开始，您现在可以添加最多10个群聊管理员。(**Lark is a popular messaging and collaboration app in China. "4.1" refers to the release of version 4.1.** "group chat administrators" are administrators who manage and configure group chats within a group.) ✓

Table 9 EN⇒ZH translation examples. The **bold** words in the prompt are translated in different ways by these two models. The colored parts are the corresponding translations and explanations. ✗ denotes the incorrect interpretation of the input sentence judged by annotators and vice versa.

5.2 Parallel Data Quality and Usage Matter

Prior research has proposed various "optimal" approaches for leveraging bilingual data in LLM-based machine translation, but our experiments demonstrate their limitations. Xu et al. cautioned against using excessive parallel data during LLM fine-tuning to avoid washing out the pre-existing knowledge. Alves et al. suggested that simply combining monolingual and parallel data would maximize performance. However, our work presents an alternative training strategy using parallel data and the key points are three-folds:
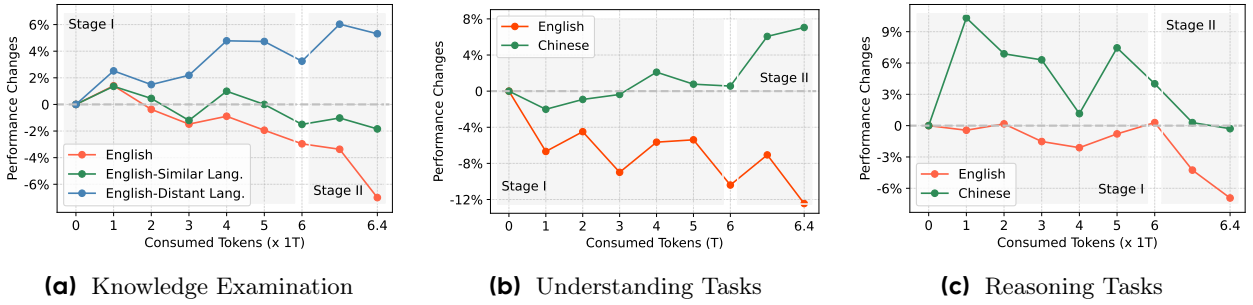
- **Simple word-mapping sentence pairs are detrimental to translation quality**. Current open-source bilingual datasets suffer from significant quality issues, notably domain limitations, inadequate semantic alignment, and unnatural expressions. These deficiencies may have influenced the findings reported by Xu et al.. To address this, we spend much efforts in iterative filtering and refinement of the collected parallel data (detailed in Section 2.2), which is empirically validated in Table 4c.

- **Continue training on pure parallel data maximizes translation effectiveness**. While this approach is generally avoided due to the risk of catastrophic forgetting, our experiments demonstrate its effectiveness when combined with high-quality parallel data and optimized learning rate scheduling. Furthermore, our investigation (Section 5.4) reveals intriguing insights into how parallel data influences multilingual transfer capabilities throughout this phase.

- **Over-training on pure parallel data may lead to overfitting**. Given the distributional shift during this process, we carefully monitor the trend of translation performance and stop training in time to prevent overfitting. While such degradation is often imperceptible through automated metrics, it becomes readily apparent during human evaluation.

5.3 Translation Entails Thinking

In existing open-source translation-oriented LLMs [8, 37], the application of CoT reasoning to translation tasks remains largely unexplored. We propose that the model should systematically consider three key aspects: comprehending the source text's context, mastering domain-specific terminology and vocabulary, and formulating idiomatic expressions in the target language. We ask linguists to annotate the CoT as the above process, with an example illustrated in Table 10. The phrase "做斗争" in this context significantly deviates from its literal meaning, thus requiring an appropriate adjustment in translation. Results presented in Table 6 demonstrate that CoT significantly contributes to improving translation accuracy.

Prompt	Translate the following Chinese sentence into English and explain it in detail: 每次化妆都在跟自己极不对称的脸做斗争，今天拍视频的截图给大家对比下。
CoT	**(Summary the meaning of the given sentence)** This is the description of having to deal with facial asymmetry every time the speaker puts on makeup, and intending to let everyone see the comparison by sharing screenshots of the video she took. **(Explain unclear or difficult words)** "做斗争" refers to the speaker uses makeup techniques to adjust asymmetrical face. "做斗争" is a metaphor. It figuratively expresses that in the process of make-up, the make-up artist needs to put in efforts to deal with the asymmetry of the face in order to achieve a relatively satisfactory make-up effect. **(Give the proper translation)** "做斗争" in this context does not convey a literal struggle or confrontation, but rather refers to "making efforts to deal with" something. Therefore, it should be translated as "try to use makeup techniques to adjust asymmetrical face".
Translation	Every time I put on makeup, I'm trying to use makeup techniques to adjust my asymmetrical face. Today, I've taken screenshots from a video to show you all a comparison.

Table 10 A Chinese-to-English translation case study with annotated Chain-of-Thought process.

(a) Knowledge Examination   **(b)** Understanding Tasks   **(c)** Reasoning Tasks

Figure 5 Performance changes on benchmarks in different stages of pretraining (**Stage I**: mixed corpora with an increased ratio of parallel data, **Stage II**: parallel data only). English-distant languages include Japanese, Thai, Russian, Malay, Indonesian, and Arabic, while English-similar languages refer to German, Spanish, Portuguese, and French.

5.4   Mono–Parallel Balance Promotes Language Transfer

5.4.1   Progressive Training Strategy with Language Prioritization

Prior work [46] identifies the existence of a core language mechanism, whereby the model comprehends queries by converting multilingual inputs into English for task processing. This finding motivates our strategic organization of data ratios based on the distinction between core and side languages. We regard English as the primary language, which should constitute the dominant portion of the LLM's knowledge storage. As described in section 2.3, we continue with multilingual-dominant and parallel-only stages. Compared with direct mixing, this strategy can prevent the secondary languages from being overlooked due to data scarcity.

5.4.2   Knowledge Transfer from Primary to Secondary Languages

This section investigates how generic knowledge can be leveraged to enhance translation capabilities, and examines the transfer of knowledge from primary to secondary languages under our proposed training strategy. We analyze the model performance starting from the completion of general pretraining through subsequent stages where translation capabilities are progressively enhanced through increased exposure to parallel data as **Stage I** and **Stage II**. We evaluate three fundamental capabilities: knowledge retention, language understanding, and reasoning ability.[2] Based on the observations in Table 5, we can conclude that:

---

[2]We refer to OpenCompass [47] to select datasets. For the task of knowledge examination, we choose ARC-easy/challenge and their multilingual version. For understanding tasks, we evaluate the model on OpenBookQA, TNEWS, AFQMC, C3, and iFLYTEK to report average accuracy for English and Chinese. For reasoning tasks, we select HellaSwag/COPA for English and CMNLI for Chinese.

Multi-parallel	BLEURT	
Data	EN⇒XX	XX⇒EN
✓	71.63	75.44
✗	**79.46**	**77.77**

Table 11 Model performance comparison with and without multi-parallel data during instruction tuning.

- **Parallel corpora allows knowledge transfer from English-similar to English-distant languages**. As shown in Figure 5a, English-distant languages demonstrate consistently improving accuracy in knowledge examination with the number of tokens consumed. Importantly, since our parallel data consists of multiple revisions of the original content, no new factual knowledge is introduced in later training stages. Thus, we attribute this growth to knowledge transfer from core languages.

- **Continue training on pure parallel data significantly enhances LLMs' understanding by establishing cross-lingual semantic equivalence**. As seen in Figure 5b, the performance of Chinese shows modest improvement during stage I but exhibits substantial gains in stage II. These significant improvements can be primarily attributed to the predominant proportion of Chinese-English parallel data in Stage II.

- **Certain performance trade-offs remain inevitable**. First, core language capabilities experience varying degrees of degradation during both two stages, attributed to the incorporation of additional languages. Second, English reasoning performance deteriorates significantly in stage II, primarily because the segmented nature of parallel corpora disrupts the coherence of monolingual long-form texts. With a limited model size, we need to sacrifice some capabilities that have a relatively small impact on translation. In the future, we will explore how to mitigate these losses under distribution shifts during continual training.

5.5 Overfitting Risks in Multilingual Instruction Data

Acquiring high-quality supervised multilingual translation data that covers a wide range of language directions remains a significant challenge. Although combining multi-parallel corpora seems like a straightforward solution to expand directional coverage, our empirical results, as shown in Table 11, indicate that such data substantially impairs model performance during instruction fine-tuning. We attribute this phenomenon to the inherent overfitting risk posed by multi-parallel data's simplified learning patterns. Based on these findings, we deliberately avoid indiscriminately increasing direction coverage through multi-parallel data, instead leveraging the LLM's generalization capability to handle uncovered directions.

6 Related Work

Recent work has explored beating ultra-large models in specific abilities by pushing small-scale models with maximum leverage. In this paper, we extend the limitations of the 7B model as a professional interpreter recognized by human annotators. This section reviews this trend of tailoring LLMs for specific tasks, especially in the field of machine translation.

Specialization of LLMs. The specialization aims at higher accuracy and more expert-level responses in the respective fields. Specialized LLMs are trained on targeted datasets that are relevant to a particular domain, such as law [48, 49] and medicine [50], or specific tasks like mathematical reasoning [51, 52] and code [53–55]. These models can be developed in two ways: (1) pre-training from scratch with a combination of general and domain-related text, and (2) starting from an open source model and adapting to the target field through continuous training and task-specific supervised fine-tuning [56]. Although the second method is significantly cheaper, the overall performance may be somewhat limited by the initial capabilities of the base model. Thus, we train from scratch to push the limits of multilingualism in pre-training to lay a solid foundation for the translation task.

Translation-oriented LLMs. Much effort has been made to match the cutting-edge multilingual translation capabilities of GPT-4. The TowerInstruct 13B model [8] outperforms open alternatives but lags behind GPT-4

in automatic evaluation. The ALMA series models [45, 57] perform better than the NLLB-54B model [4] and GPT-3.5, but still fall short of GPT-4. Importantly, existing research only reports automatic evaluation on public MT datasets, despite ultra-large models often outperforming others when evaluated by humans on more varied and authentic translations, especially in challenging scenarios [58]. This paper makes remarkable progress, surpassing ultra-large models on both automatic metrics and human evaluation. We also release a challenging MT benchmark to set a higher standard for conciseness, accuracy, and elegance.

Balancing multilingual capabilities has been crucial in machine translation [59], especially for low-resource languages [60]. This paper proposes an efficient training approach to maximize multilingual translation ability and explores a new perspective on cross-lingual knowledge transfer from primary to secondary languages.

7 Conclusion

In this paper, we present SEED-X, a family of open-source LLMs designed specifically for machine translation, supporting bidirectional translation across 28 languages. We provide a detailed introduction of the complete training pipeline, including pre-training, instruction tuning, and reinforcement learning, while sharing key insights and best practices derived from our hill-climbing optimization process. Remarkably, with only 7B parameters, SEED-X achieves translation quality comparable to, or even surpassing, the most advanced LLMs and leading commercial translation systems, as evidenced by both automatic and human evaluations. By releasing SEED-X's publicly available weights, we aim to make SEED-X an accessible, off-the-shelf tool to support the community in advancing translation research and applications.

References

[1] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics, 1993. URL https://aclanthology.org/J93-2003/.

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

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

[4] Team NLLB, 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.

[5] Xavier Garcia, Yamini Bansal, Colin Cherry, George Foster, Maxim Krikun, Fangxiaoyu Feng, Melvin Johnson, and Orhan Firat. The unreasonable effectiveness of few-shot learning for machine translation, 2023. URL https://arxiv.org/abs/2302.01398.

[6] Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. Multilingual machine translation with large language models: Empirical results and analysis. In Findings of the Association for Computational Linguistics: NAACL 2024, 2024. URL https://aclanthology.org/2024.findings-naacl.176/.

[7] Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. A paradigm shift in machine translation: Boosting translation performance of large language models. In The Twelfth International Conference on Learning Representations, 2024. URL https://openreview.net/pdf?id=farT6XXntP.

[8] Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. Tower: An open multilingual large language model for translation-related tasks, 2024.

[9] Anthropic. Claude 3.5. https://www.anthropic.com/claude, 2025. Accessed: 2025-04-18.

[10] Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In Proceedings of the Ninth Conference on Machine Translation, 2024. URL https://aclanthology.org/2024.wmt-1.1/.

[11] Jianhui Pang, Fanghua Ye, Derek Fai Wong, Dian Yu, Shuming Shi, Zhaopeng Tu, and Longyue Wang. Salute the classic: Revisiting challenges of machine translation in the age of large language models. Transactions of the Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.tacl-1.4/.

[12] Gemma Team. Gemma 3 technical report. https://arxiv.org/abs/2503.19786, 2025.

[13] AI Meta. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. https://ai.meta.com/blog/llama-4-multimodal-intelligence/, checked on, 2025.

[14] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

[15] OpenAI. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/, 2023. Accessed: 2025-04-18.

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

[17] DeepMind. Gemini 2.5. https://deepmind.google/technologies/gemini/, 2025. Accessed: 2025-04-18.

[18] Keming Lu, Hongyi Yuan, Zheng Yuan, Runji Lin, Junyang Lin, Chuanqi Tan, Chang Zhou, and Jingren Zhou. # instag: Instruction tagging for analyzing supervised fine-tuning of large language models. In The Twelfth International Conference on Learning Representations, 2023.

[19] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. arXiv preprint arXiv:2310.06825, 2023.

[20] Meta AI. The llama 3 herd of models. ArXiv, abs/2407.21783, 2024. URL https://api.semanticscholar.org/CorpusID:271571434.

[21] Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yi-Chao Zhang, Yunyang Wan, Yuqi Liu, Zeyu Cui, Zhenru Zhang, Zihan Qiu, Shanghaoran Quan, and Zekun Wang. Qwen2.5 technical report. ArXiv, abs/2412.15115, 2024. URL https://api.semanticscholar.org/CorpusID:274859421.

[22] Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Mohamed A Haggag, Alfonso Amayuelas, et al. Include: Evaluating multilingual language understanding with regional knowledge. arXiv preprint arXiv:2411.19799, 2024.

[23] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. arXiv preprint arXiv:2009.03300, 2020.

[24] Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. Xcopa: A multilingual dataset for causal commonsense reasoning. arXiv preprint arXiv:2005.00333, 2020.

[25] Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, TL Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, et al. Crosslingual generalization through multitask finetuning (2022). arXiv preprint arXiv:2211.01786, 1786.

[26] Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. Few-shot learning with multilingual generative language models. In Proceedings of the 2022 conference on empirical methods in natural language processing, pages 9019–9052, 2022.

[27] Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. Paws-x: A cross-lingual adversarial dataset for paraphrase identification. arXiv preprint arXiv:1908.11828, 2019.

[28] Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. Transactions of the Association for Computational Linguistics, 10:522–538, 2022.

[29] Xingyuan Pan, Luyang Huang, Liyan Kang, Zhicheng Liu, Yu Lu, and Shanbo Cheng. G-dig: Towards gradient-based diverse and high-quality instruction data selection for machine translation. arXiv preprint arXiv:2405.12915, 2024.

[30] Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and Jingren Zhou. Scaling relationship on learning mathematical reasoning with large language models. arXiv preprint arXiv:2308.01825, 2023.

[31] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pages 311–318, 2002.

[32] Thibault Sellam, Dipanjan Das, and Ankur Parikh. Bleurt: Learning robust metrics for text generation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7881–7892, 2020.

[33] Nuno M Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André FT Martins. xcomet: Transparent machine translation evaluation through fine-grained error detection. arXiv preprint arXiv:2310.10482, 2023.

[34] Nuno M Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André FT Martins. xcomet: Transparent machine translation evaluation through fine-grained error detection. arXiv preprint arXiv:2310.10482, 2023.

[35] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. Comet: A neural framework for mt evaluation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2685–2702, 2020.

[36] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347, 2017.

[37] Yinquan Lu, Wenhao Zhu, Lei Li, Yu Qiao, and Fei Yuan. Llamax: Scaling linguistic horizons of llm by enhancing translation capabilities beyond 100 languages. arXiv preprint arXiv:2407.05975, 2024.

[38] Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, and Dahua Lin. Internlm2 technical report, 2024.

[39] John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. Aya expanse: Combining research breakthroughs for a new multilingual frontier, 2024. URL https://arxiv.org/abs/2412.04261.

[40] Mistral AI. Mistral large. https://mistral.ai/news/mistral-large-2407, 2024.

[41] Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. Deepseek llm: Scaling open-source language models with longtermism. arXiv preprint arXiv:2401.02954, 2024.

[42] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. arXiv:1803.05457v1, 2018.

[43] Liwei Wu, Shanbo Cheng, Mingxuan Wang, and Lei Li. Language tags matter for zero-shot neural machine translation. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 3001–3007, 2021.

[44] Yang Liu, Peng Sun, and Hang Li. Large language models as agents in two-player games. arXiv preprint arXiv:2402.08078, 2024.

[45] Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. A paradigm shift in machine translation: Boosting translation performance of large language models, 2023.

[46] Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. How do large language models handle multilingualism? arXiv preprint arXiv:2402.18815, 2024.

[47] OpenCompass Contributors. Opencompass: A universal evaluation platform for foundation models. https://github.com/open-compass/opencompass, 2023.

[48] Haitao Li, Qingyao Ai, Jia Chen, Qian Dong, Yueyue Wu, Yiqun Liu, Chong Chen, and Qi Tian. Sailer: structure-aware pre-trained language model for legal case retrieval. In Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 1035–1044, 2023.

[49] Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. Chatlaw: Open-source legal large language model with integrated external knowledge bases. arXiv preprint arXiv:2306.16092, 2023.

[50] Hongjian Zhou, Boyang Gu, Xinyu Zou, Yiru Li, Sam S Chen, Peilin Zhou, Junling Liu, Yining Hua, Chengfeng Mao, Xian Wu, et al. A survey of large language models in medicine: Progress, application, and challenge. arXiv preprint arXiv:2311.05112, 2023.

[51] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, YK Li, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300, 2024.

[52] Ke Wang, Houxing Ren, Aojun Zhou, Zimu Lu, Sichun Luo, Weikang Shi, Renrui Zhang, Linqi Song, Mingjie Zhan, and Hongsheng Li. Mathcoder: Seamless code integration in llms for enhanced mathematical reasoning. arXiv preprint arXiv:2310.03731, 2023.

[53] Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, et al. Code llama: Open foundation models for code. arXiv preprint arXiv:2308.12950, 2023.

[54] Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, et al. Starcoder: may the source be with you! arXiv preprint arXiv:2305.06161, 2023.

[55] Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. Wizardcoder: Empowering code large language models with evol-instruct. In The Twelfth International Conference on Learning Representations, 2023.

[56] Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mammoth: Building math generalist models through hybrid instruction tuning. arXiv e-prints, pages arXiv–2309, 2023.

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

[58] Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. Is chatgpt a good translator? yes with gpt-4 as the engine. arXiv preprint arXiv:2301.08745, 2023.

[59] Jiahuan Li, Hao Zhou, Shujian Huang, Shanbo Cheng, and Jiajun Chen. Eliciting the translation ability of large language models via multilingual finetuning with translation instructions. Transactions of the Association for Computational Linguistics, 12:576–592, 2024.

[60] Wei Zou, Sen Yang, Yu Bao, Shujian Huang, Jiajun Chen, and Shanbo Cheng. Trans-zero: Self-play incentivizes large language models for multilingual translation without parallel data. arXiv preprint arXiv:2504.14669, 2025.

[61] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. Neurocomputing, 568:127063, 2024.

[62] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1715–1725, 2016.

[63] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437, 2024.

Appendix

A Supported Languages

Languages	Abbr.	Languages	Abbr.
Arabic	ar	Malay	ms
Czech	cs	Norwegian Bokmal	nb
Danish	da	Dutch	nl
German	de	Norwegian	no
English	en	Polish	pl
Spanish	es	Portuguese	pt
Finnish	fi	Romanian	ro
French	fr	Russian	ru
Croatian	hr	Swedish	sv
Hungarian	hu	Thai	th
Indonesian	id	Turkish	tr
Italian	it	Ukrainian	uk
Japanese	ja	Vietnamese	vi
Korean	ko	Chinese	zh

**Table 12** Supported languages.

B Implementation Details

Architecture. To enhance accessibility and adoption within the open-source community, we construct our 7B model following the architecture of Mistral-7B [19]. This architecture is based on the Transformer [3], and it consists of 32 Transformer decoder layers, each containing 32 self-attention heads, an embedding size of 4,096 dimensions, and a feed-forward neural network size of 14,336 dimensions. Layer normalization is used to stabilize learning. Additionally, to improve the model's ability to handle positional information, the model equipped with Rotary Position Embedding [ROPE, 61]. ROPE modifies the standard positional encoding by using a rotation matrix, allowing the model to capture relative positions more effectively. Additionally, we set the maximum sequence length to 2,048, precisely calibrated to match our real-world translation requirements.

Vocabulary. Building upon established techniques, we design our tokenization approach utilizing the Byte Pair Encoding [BPE, 62] method, similar to what Mistral uses, with a significant enhancement by expanding the original vocabulary from 32,000 to 65,269 tokens. Compared to the original vocabulary, our enhanced tokenizer boosts compression rates from 3.17 to 3.74 characters per token across a diverse set of multilingual data. This strategic extension enables our model to better handle the multilingual data and nuances of translation tasks, ensuring more comprehensive linguistic coverage.

Training. Through rigorous experimentation, we methodically test and optimize various batch sizes and learning rates. We find that utilizing a batch size of 2M tokens alongside a learning rate of 3e-4 constitutes the optimal setup for pretraining. The models undergo a warm-up phase involving 2,000 linear scaling steps to reach the peak learning rate, followed by cosine decay to gradually reduce it to a minimum of 10% of the peak rate. During the supervised fine-tuning stage, we utilize a batch size of 64 sentences and a learning rate of 3e-6, striking a precise balance that refines the model's performance, ultimately leading to superior translation quality.

Inference. During inference, we utilize beam search (with a beam size of 4) to generate translation results. We observed that beam search slightly outperforms greedy decoding, though at the expense of inference speed.
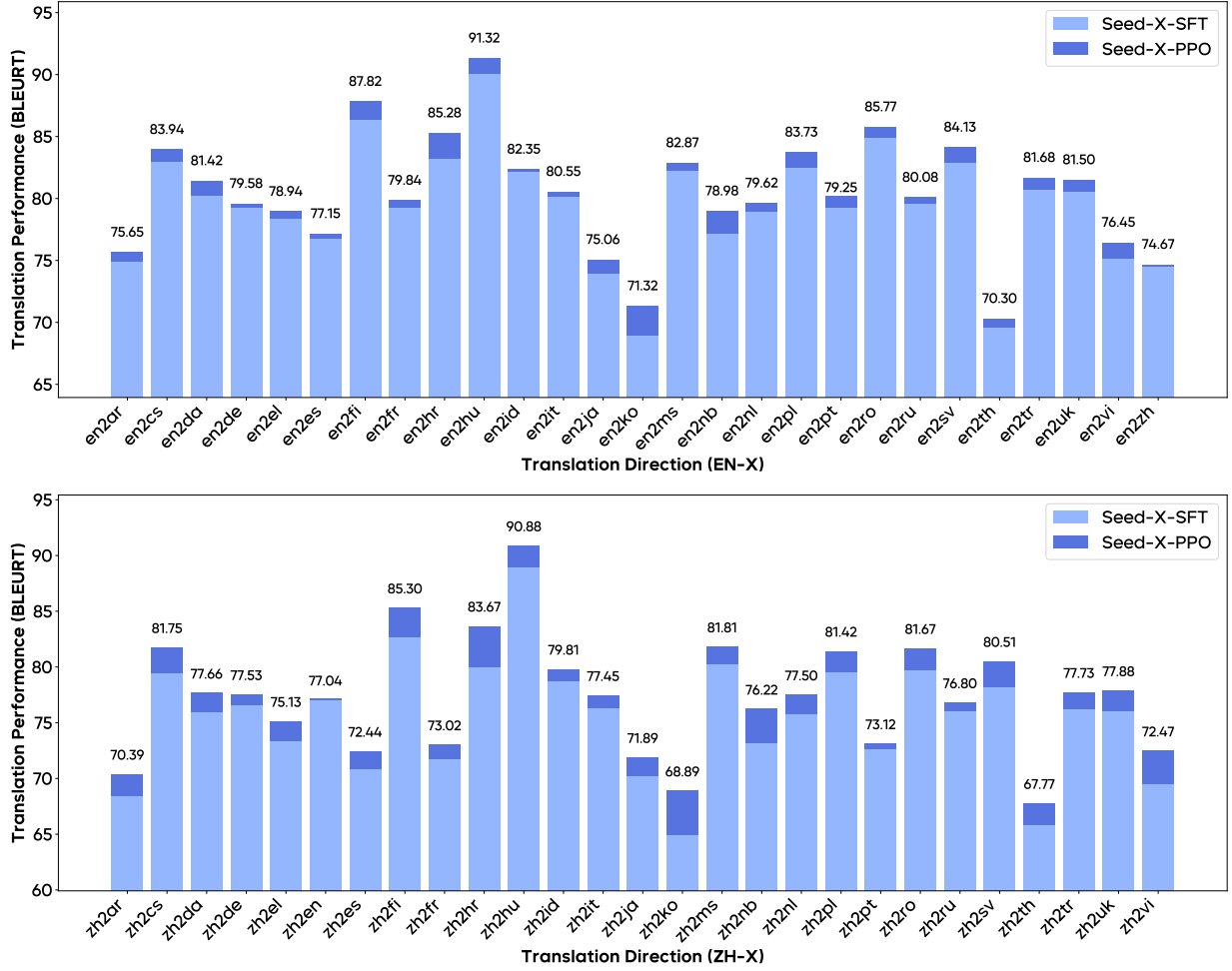
Figure 6 Performance comparison between Seed-X-SFT and Seed-X-PPO across evaluated translation directions on Flores-200.

C Translation Performance Across Individual post-training stages

In Figure 6, we present the the detailed performance of SEED-X across multiple translation directions after instruction-tuning (SFT) and reinforcement learning (PPO). The results show that, after the SFT stage, the model already achieves strong translation quality, as measured by BLEURT scores. PPO further improves upon the SFT model, pushing the performance ceiling for machine translation.

D Human Evaluation

Figure 7 shows the human evaluation results for the Chinese-English translation task. For this high-resource language pair, we use a strict deduction system to evaluate translation quality, accumulating deductions for major and minor errors in the translated text. It demonstrates that SEED-X outperforms most models with a distinct advantage on EN⇒ZH, trailing only behind Deepseek-R1. Table 13 shows the human evaluation scores for each translation direction corresponding to Figure 3. In 14 language directions, SEED-X-PPO achieved 6 first places and one second place, with the highest overall average score, showing a clear advantage in EN⇒XX language directions.
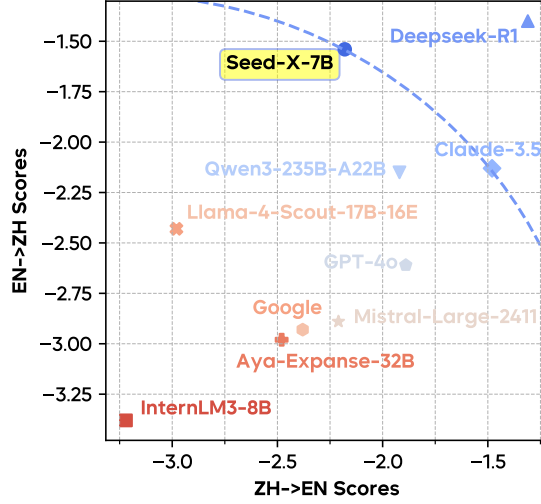
Figure 7 Human evaluation of models on the **Seed-X-Challenge** for the Chinese-English translation task. We use a strict deduction system to evaluate the quality of translations, accumulating deductions for major and minor errors in the translated text, so the final scores presented are all negative numbers. The average Chinese-English translation ability of the models within the arc in the figure is worse than that of SEED-X.

Models	EN⇒XX							XX⇒EN							Avg.
	es	de	fr	ru	pt	it	ar	es	de	fr	ru	pt	it	ar	
InternLM3-8B[†] [38]	2.56	2.77	3.11	2.70	3.74	2.60	1.69	3.35	1.85	3.06	3.16	3.07	3.05	2.50	2.80
LLaMAX3-8B[†] [37]	2.85	3.04	3.30	3.16	3.76	3.11	2.66	3.23	2.10	3.06	3.55	2.72	3.10	3.26	3.06
Google-Translator	3.66	3.55	3.44	3.77	3.84	3.42	3.53	3.29	3.32	2.99	3.50	3.11	3.19	3.57	3.44
Mistral-Large-2411-123B[†] [40]	3.55	3.61	3.38	3.64	3.79	3.31	3.55	3.30	3.20	3.28	3.68	3.07	3.53	3.60	3.46
Llama-4-Scout-17B-16E[†] [13]	3.62	3.60	3.62	3.66	3.86	3.50	3.38	3.56	3.18	3.70	3.47	3.31	3.65	3.60	3.55
Gemma3-27B[†] [12]	3.72	**3.66**	3.43	3.76	3.81	3.39	3.57	3.75	3.47	3.50	3.75	3.29	3.61	3.73	3.60
Qwen3-235B-A22B[†] [14]	3.62	3.56	3.75	3.69	3.87	3.61	3.25	3.56	3.32	3.77	3.76	3.41	3.74	3.64	3.61
Aya-32B[†] [39]	3.59	3.58	3.79	3.75	3.88	**3.64**	3.60	3.73	3.25	3.66	3.25	3.34	3.72	3.89	3.62
DeepSeek-V3[†] [63]	3.81	3.53	3.60	3.73	3.81	3.31	3.53	3.80	3.52	3.69	3.79	3.37	3.55	3.91	3.64
DeepSeek-R1[†] [16]	3.82	3.57	3.51	3.69	3.81	3.31	3.59	3.75	3.66	3.79	3.81	3.21	3.64	3.84	3.64
GPT-4o [15]	3.74	3.65	3.63	3.74	3.82	3.60	3.54	3.58	3.27	3.62	3.63	**3.67**	3.84	**3.91**	3.66
Gemini-2.5-Pro [17]	3.76	3.59	3.53	3.61	3.88	3.58	3.67	3.69	3.42	3.85	3.52	3.63	3.86	3.68	3.66
Claude-3.5-Sonnet [9]	3.84	3.50	3.55	3.75	3.82	3.48	3.59	**3.87**	**3.76**	3.67	3.79	3.35	3.69	3.87	3.68
Seed-X-PPO	**3.87**	3.61	**3.88**	**3.88**	**3.88**	3.55	**3.78**	3.85	3.49	3.74	3.55	3.55	**3.87**	3.75	**3.73**

Table 13 Human evaluation scores (0-4) on the **Seed-X-Challenge** for 14 language directions corresponding to Figure 3. Open-source models are marked with †. Models are arranged in ascending order by average score, with **the highest score** in bold and the second-highest score underlined.

E Case Study

Table 14 presents representative Chinese-English translation examples, highlighting key testing points and corresponding translations by SEED-X. These cases demonstrate SEED-X's contextual awareness and competence in handling complex translations.

Scenario 1: Internet buzzwords

Input #1	在某宝上看到一件裙子 绝绝子 红色YYDS，加油瘦 然后去拥有它！
Testing Points	"某宝": Taobao, "绝绝子": amazing/incredible, "YYDS": forever the best, "加油瘦": the subject is the speaker.
Google	I saw a skirt on Taobao. It's **absolutely red. YYDS**, go ahead and lose weight, then **go get it**.
Seed-X	I saw a skirt on Taobao that's absolutely amazing! Red is **the ultimate choice**! Keep working hard to **lose weight and then go get it'll be yours**!
Input #2	光腿神器
Testing Points	"光腿神器": a pair of leggings that look just like bare legs when worn.
Google	**Bare Legs Artifact**
Seed-X	**Pantyhose that make legs look bare**
Input #3	INTJ总是装E
Testing Points	"E": in psychology, it usually refers to people with an extraverted personality.
Google	INTJ always pretends to be **E**?
Seed-X	INTJs always pretend to be **extraverts**?
Input #4	集美们在集美区做什么？
Testing Points	"集美们": it has the similar pronunciation with "姐妹" (girls).
Google	What do **Jimei** people do in Jimei District?
Seed-X	What are you **girls** doing in Jimei District?
Input #4	u1s1，这次城市峰会真是泰裤辣！
Testing Points	"u1s1": it is an abbreviation, which means "to be honest". "泰裤辣": it has the same pronunciation with "太酷啦", which means "awesome".
Google	**u1s1**, this city summit is really **hot!**
Seed-X	**To be honest**, this city summit is really **awesome!**

Scenario 2: Chinese Slang

Input #6	不撞南墙是会不知道有多痛的。。。疼了痛了才会明白！
Testing Points	"撞南墙": to learn something the hard way.
Google	You won't know how painful it is until you **hit your head against the wall**... You won't understand until you feel the pain!
Seed-X	You won't know how painful it is until you **hit a wall head-on**... Only when it hurts will you realize it!
Input #5	有点厉害，不知道水分大不大
Testing Points	"水分": data or descriptions that contain exaggerated or untrue elements.
Google	It's a bit powerful, I don't know if it has a lot of **water**.
Seed-X	That's pretty impressive. I wonder if there's **a lot of hype** around it.
Input #6	我说一句你说一车啊
Testing Points	"说一车": an exaggerated expression, which means that there's a lot to say.
Google	I said one thing, you said **one car**
Seed-X	I say one thing, and you come up with **a whole bunch of responses**

Scenario 3: Classical Chinese

Input #8	若夫淫雨霏霏，连月不开，阴风怒号，浊浪排空
Testing Points	释义：像那阴雨连绵，接连几个月不放晴，寒风怒吼，浑浊的浪冲向天空。
Google	If the rain is **heavy and the moon is not bright**, the wind is howling, and the turbid waves are rolling
Seed-X	As for the **persistent** rain, it lasted for several months without stopping. The **gloomy wind howled furiously**, and the turbulent waves filled the sky.

Input #9	桃李不言，下自成蹊。
Testing Points	释义：桃树、李树不会说话，但因其果实可口，人们纷纷去摘取，于是便在树下踩出一条路来。比喻只要人品高尚，就能得到别人的尊敬和景仰。
Google	The peaches and plums do not speak, but **they create a trail of their own**.
Seed-X	**A good example always speaks for itself, and others will naturally follow suit.**
Input #10	天高地迥，觉宇宙之无穷；兴尽悲来，识盈虚之有数。
Testing Points	释义：苍天高远，大地寥廓，令人感到宇宙的无穷无尽。欢乐逝去，悲哀袭来，意识到万事万物皆有定数。
Google	The sky is high and the earth is vast, and one feels the infinity of the universe; when the excitement ends and sadness comes, one realizes **the finiteness of fullness and emptiness**.
Seed-X	The sky is high and the earth is vast, making me feel the boundlessness of the universe. When my joy reaches its peak, sorrow inevitably follows, and I realize that **all things in the world have their own limits**.
Scenario 4: English Slang	
Input #12	Their relationship is a total situationship.
Testing Points	It's the "grey area" where two people are involved romantically but haven't defined what they are to each other.
Google	他们的关系完全是一种关系。
Seed-X	他们的关系完全是逢场作戏。
Input #13	I'm totally hangry right now.
Testing Points	"hangry": it combines the words "hungry" and "angry" to describe the bad mood.
Google	我现在非常饿。
Seed-X	我现在真是又饿又生气。
Input #14	Out of the frying pan and into the fire.
Testing Points	It means you've escaped from one bad situation to end up in an even worse one.
Google	刚从油锅里跳进火里。
Seed-X	真是屋漏偏逢连夜雨。

Table 14 Comparative case studies across different scenarios. Red text highlights translation errors in Google Translate outputs, demonstrating Seed-X's enhanced accuracy.

F   Contributions

Project Lead

Shanbo Cheng

Core Contributors

Yu Bao, Qian Cao[1], Luyang Huang, Liyan Kang, Zhicheng Liu, Yu Lu, Wenhao Zhu

Contributors

Jingwen Chen[1], Zhichao Huang, Tao Li, Yifu Li, Huiying Lin[1], Sitong Liu, Ningxin Peng, Shuaijie She[2], Lu Xu, Nuo Xu, Sen Yang[2], Runsheng Yu, Yiming Yu, Liehao Zou

Supervision

Hang Li, Lu Lu, Yuxuan Wang, Yonghui Wu

(Last-Name in Alphabetical Order)

---

1 indicates the contributor has already left ByteDance. We appreciate their contribution to this work.
2 indicates work done during internship, Nanjing University.