Language on Demand, Knowledge at Core: Composing LLMs with Encoder-Decoder Translation Models for Extensible Multilinguality

Mengyu Bu[1,2,3], **Yang Feng**[1,2,3†]
[1]Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences (ICT/CAS) [2] State Key Laboratory of AI Safety, Institute of Computing Technology, Chinese Academy of Sciences
[3] University of Chinese Academy of Sciences, Beijing, China
bumengyu23z@ict.ac.cn, fengyang@ict.ac.cn

Abstract

Large language models (LLMs) exhibit strong general intelligence, yet their multilingual performance remains highly imbalanced. Although LLMs encode substantial cross-lingual knowledge in a unified semantic space, they often struggle to reliably interface this knowledge with low-resource or unseen languages. Fortunately, pretrained encoder-decoder translation models already possess balanced multilingual capability, suggesting a natural complement to LLMs. In this work, we propose XBridge, a compositional encoder-LLM-decoder architecture that offloads multilingual understanding and generation to external pretrained translation models, while preserving the LLM as an English-centric core for general knowledge processing. To address the resulting representation misalignment across models, we introduce lightweight cross-model mapping layers and an optimal transport-based alignment objective, enabling fine-grained semantic consistency for multilingual generation. Experiments on four LLMs across multilingual understanding, reasoning, summarization, and generation indicate that XBridge outperforms strong baselines, especially on low-resource and previously unseen languages, without retraining the LLM.[1]

1 Introduction

Large language models (LLMs) have demonstrated remarkable general intelligence and reasoning abilities (Touvron et al., 2023; Üstün et al., 2024; Qwen et al., 2025), which are largely grounded in a unified semantic knowledge space. However, despite possessing substantial cross-lingual knowledge, LLMs exhibit imbalanced multilingual performance: while performing reliably in English and a few high-resource languages, they often fail to robustly understand or generate text in low-resource or unseen languages (Zhu et al., 2023; Chang et al., 2024). This suggests that the core limitation of LLMs lies not in the absence of knowledge, but in the difficulty of interfacing this knowledge with diverse linguistic representation spaces.

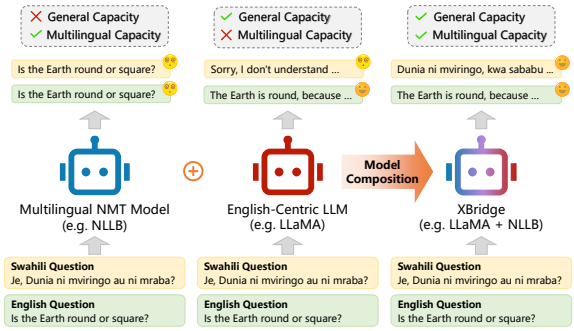


Figure 1: Overview of XBridge. Pretrained multilingual NMT models provide broad language coverage but limited general reasoning capability, while English-centric LLMs excel at general reasoning yet struggle with low-resource or unseen languages. XBridge harmonizes these strengths through model composition, offloading multilingual processing to the pretrained multilingual model while leveraging the LLM as a knowledge core.

Fortunately, a wealth of encoder-decoder based neural machine translation (NMT) models (Xue et al., 2021; Team et al., 2022) specialize in multilingual understanding and generation, and thus provide complementary capabilities to LLMs. These models support semantic transfer across hundreds of languages, including many low-resource ones, by learning a shared semantic representation space across languages. In such models, the encoder maps input text from different languages into the shared semantic space, while the decoder subsequently projects these shared representations into target-language outputs. This closed semantic loop between understanding and generation, along with the modular design of encoder and decoder, naturally complements LLMs. Realizing such a composition would provide LLMs with extensible multilingual capability, particularly for low-resource or

†Corresponding author: Yang Feng.
[1]https://github.com/ictnlp/XBridge

unseen languages that are well modeled by NMT systems but remain challenging for LLMs.

However, existing approaches only partially address this goal, which integrate multilingual encoders to improve multilingual understanding by injecting encoder representations into LLM inputs (Yoon et al., 2024; Huang et al., 2024; Ruan et al., 2025). While effective for input understanding, these approaches leave generation largely English-centric. A natural extension is to further incorporate the multilingual decoder, but doing so introduces a fundamental structural challenge. In NMT, the encoder and decoder are jointly trained within a unified representation space, whereas inserting a frozen LLM in between introduces a transformation from the LLM input space to a different output space shaped by its internal knowledge processing. Consequently, the LLM outputs no longer match the decoder's expected cross-attention representations, resulting in semantic misalignment that cannot be resolved by simple projection.

To address this challenge, we propose XBridge, which composes LLMs with pretrained multilingual NMT models for extensible multilinguality. XBridge adopts an encoder-LLM-decoder architecture, where a multilingual encoder provides robust semantic representations for multilingual inputs, a frozen LLM serves as an English-centric core for knowledge processing, and a multilingual decoder generates outputs in the target language. From a representation perspective, XBridge constructs a semantic bridge that transforms representations from the multilingual semantic space to the LLM input space, through the LLM output space after knowledge transformation, and finally into the decoder's generation space. By explicitly aligning heterogeneous representation spaces across these modules, XBridge resolves the semantic mismatch introduced by inserting a frozen LLM, achieving extensible and generalizable multilingual understanding and generation.

We evaluate XBridge on four LLMs across multilingual understanding, reasoning, summarization, and generation tasks. XBridge outperforms strong baselines, with significant gains on low-resource and unseen languages while preserving LLM's core capability. With minimal additional parameters, limited training data, and parameter-efficient training, XBridge brings low-resource and unseen language performance close to that of external NMT models, substantially narrowing the gap across languages without retraining the LLM.

2 Related Work

2.1 Data-Level Multilingual Enhancement for LLMs

A line of work augments the multilingual capabilities of LLMs at the data level by constructing multilingual training corpora using pretrained multilingual or machine translation models (Li et al., 2023; Zhang et al., 2023, 2024a,b). Typical approaches translate English instruction into multiple languages (Chen et al., 2024), pre-translate non-English inputs into English before task execution (Qin et al., 2023; Chai et al., 2025), or leverage Mix-of-Experts (MoE) for language expansion (Zhang et al., 2025b). Such approaches generally require continual multilingual training of LLMs, which may introduce translation noise and interfere with existing language capabilities. In practice, balancing performance across high- and low-resource languages remains challenging, as gains on low-resource languages often come at degradation on high-resource ones (Gao et al., 2024). In contrast, XBridge achieves multilingual generalization through model composition without multilingual retraining of the LLM.

2.2 Encoder-Augmented Multilingual LLMs

Another line of work augments LLMs with pretrained multilingual encoders, injecting encoder representations into the LLM to improve multilingual understanding. Yoon et al. (2024) leverage multilingual encoders to support cross-lingual understanding, while Huang et al. (2024) reintroduce multilingual inputs to better exploit the complementary strengths of language understanding and reasoning in LLMs. Ruan et al. (2025) further explore layer-wise fusion strategies to enhance the utilization of encoder semantics. These approaches primarily focus on improving multilingual understanding at the input side, while generation remains governed by the LLM's native language distribution, typically English. Moreover, due to differences in training objectives and tokenization schemes, representation gaps persist between multilingual encoders and LLMs, which limit the effective exploitation of encoder semantics. XBridge differs from prior encoder-augmented methods by additionally incorporating a multilingual decoder to support multilingual generation and by explicitly aligning representations across models, enabling more effective end-to-end multilingual behavior.

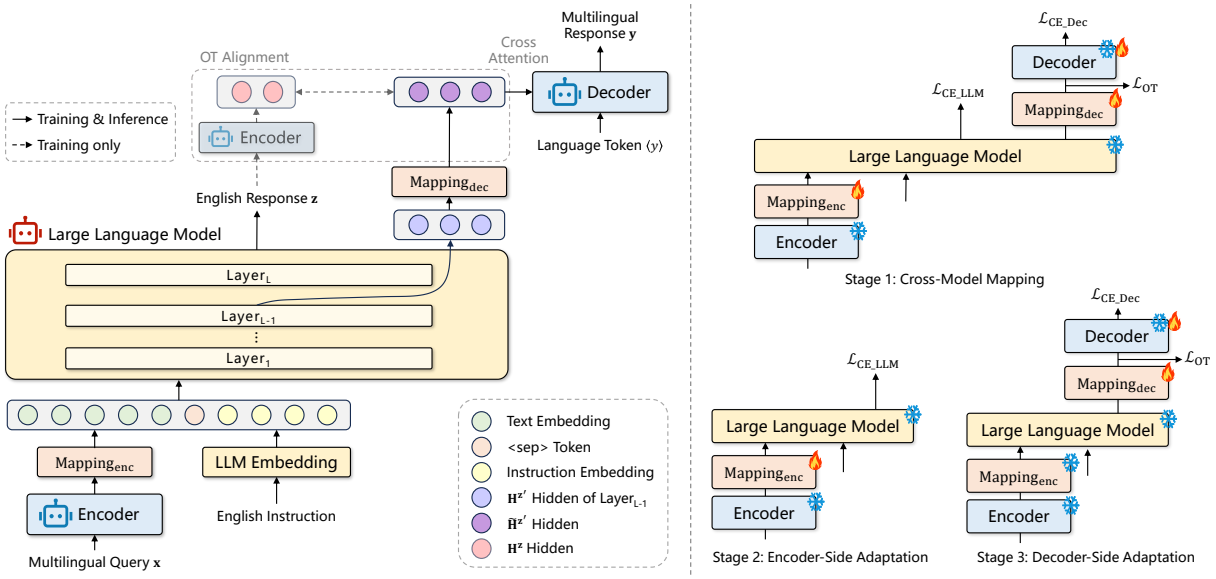


Figure 2: **Left:** XBridge composes a pretrained multilingual encoder-decoder with an LLM via lightweight mapping layers for multilingual understanding and generation, keeping the LLM frozen as a knowledge core. **Right:** A three-stage training strategy progressively aligns heterogeneous representations and adapts the encoder and decoder.

3 Method

Figure 2 presents the framework of our XBridge, a compositional multilingual framework that integrates a pretrained encoder-decoder NMT model with an LLM. XBridge efficiently offloads multilingual burden to the external NMT model while preserving the LLM as an English-centric core for general knowledge processing. XBridge adopts an encoder-LLM-decoder architecture, connected by lightweight cross-model mapping layers (Section 3.1). To facilitate fine-grained semantic transfer for multilingual generation, we introduce an optimal transport-based token alignment objective at the LLM-decoder interface (Section 3.2). For stable optimization, XBridge employs a three-stage training strategy that decouples coarse-grained cross-model alignment from task-specific adaptation (Section 3.3).

3.1 Architecture

XBridge adopts an encoder-LLM-decoder architecture to compose a pretrained encoder-decoder NMT model with an LLM for extensible multilingual understanding and generation.

Formally, given an input sequence $\mathbf{x} = (x_1, \ldots, x_n)$ in language L_x, we first encode it with the pretrained multilingual encoder $\texttt{Enc}(\cdot)$, producing contextual representations $\mathbf{H}^{\mathbf{x}} \in \mathbb{R}^{n \times d_e}$. To bridge the representation gap between the multilingual encoder and LLM, we apply a lightweight mapping $\texttt{Mapping}_{\texttt{enc}}(\cdot)$ that projects $\mathbf{H}^{\mathbf{x}}$ into the LLM representation space, yielding $\tilde{\mathbf{H}}^{\mathbf{x}} \in \mathbb{R}^{n \times d_l}$. The mapped encoder representations are then injected into the LLM together with a high-resource (English) instruction prompt, enabling the LLM to perform general knowledge processing conditioned on encoder semantics. Let $\mathbf{z} = (z_1, \ldots, z_m)$ denote the sequence of English tokens generated by the LLM. Rather than using the final-layer hidden states, we extract the penultimate-layer hidden states, denoted as $\mathbf{H}^{\mathbf{z}'} \in \mathbb{R}^{m \times d_l}$, as Zhang et al. (2025a) show that the last layer is often tightly aligned with the output vocabulary space, while non-final layers retain richer semantic information.

To support multilingual generation, XBridge further integrates a pretrained multilingual decoder $\texttt{Dec}(\cdot)$ at the output side. Specifically, we apply a decoder-side mapping $\texttt{Mapping}_{\texttt{dec}}(\cdot)$ to project the LLM hidden states into the decoder representation space, obtaining $\tilde{\mathbf{H}}^{\mathbf{z}'} \in \mathbb{R}^{m \times d_d}$, which are used as key-value representations for cross-attention in the decoder. Given target-language tokens $\langle y \rangle$ in language L_y as decoder inputs, the decoder generates the output sequence $\mathbf{y}$ by attending to $\tilde{\mathbf{H}}^{\mathbf{z}'}$, producing text that follows the target-language distribution while remaining semantically grounded in the LLM's knowledge processing results.

3.2 Optimal Transport-Based Alignment

Although the mapped LLM representations $\tilde{\mathbf{H}}^{\mathbf{z}'}$ can be directly used as cross-attention inputs for

multilingual decoding, token-level semantic misalignment may arise due to heterogeneous tokenizations and representation spaces across models. To encourage fine-grained semantic consistency at the LLM-decoder interface, we introduce an optimal transport (OT)-based alignment objective.

Specifically, given the English token sequence $\mathbf{z} = (z_1, \ldots, z_m)$ generated by the LLM, we re-encode it using the same multilingual encoder $\texttt{Enc}(\cdot)$, obtaining encoder representations $\mathbf{H}^{\mathbf{z}} \in \mathbb{R}^{k \times d_e}$, where the sequence length k may differ from m due to heterogeneous tokenizers. Since $\mathbf{H}^{\mathbf{z}}$ and the decoder-side LLM representations $\tilde{\mathbf{H}}^{\mathbf{z}'}$ are both derived from the same LLM output, they are semantically equivalent in expectation, despite residing in different representation spaces. We therefore align $\mathbf{H}^{\mathbf{z}}$ with $\tilde{\mathbf{H}}^{\mathbf{z}'}$ to enforce token-level semantic alignment.

Due to sequence length mismatch by heterogeneous tokenizers, we formulate the alignment as an optimal transport problem (Peyré et al., 2019), which computes a soft, many-to-many matching between the two sequences. Concretely, we define the OT distance between $\mathbf{H}^{\mathbf{z}}$ and $\tilde{\mathbf{H}}^{\mathbf{z}'}$ as:

$$\begin{aligned} \mathcal{D}^*(\mathbf{H}^{\mathbf{z}}, \tilde{\mathbf{H}}^{\mathbf{z}'}) &= \min_{\mathbf{T} \geq 0} \sum_{i,j} \mathbf{T}_{ij}\, c(H_i^{\mathbf{z}}, \tilde{H}_j^{\mathbf{z}'}), \\ \text{s.t.} \sum_{j=1}^{m} \mathbf{T}_{ij} &= m_i^{\mathbf{z}}, \quad \forall i \in \{1, \ldots, k\}. \end{aligned} \tag{1}$$

where $\mathbf{T}_{ij}$ denotes the transport mass from $H_i^{\mathbf{z}}$ to $\tilde{H}_j^{\mathbf{z}'}$, and $c(\cdot, \cdot)$ is the transport cost computed using cosine distance. The mass distribution $\{m_i^{\mathbf{z}}\}$ is obtained by normalizing $\mathbf{H}^{\mathbf{z}}$. Appendix A presents details of the OT formulation and optimization.

The OT loss provides flexible, token-level supervision that is robust to length mismatch. By regularizing the decoder-side mapping with encoder-derived representations of the LLM's own outputs, the OT objective encourages $\tilde{\mathbf{H}}^{\mathbf{z}'}$ to preserve semantic structures compatible with the multilingual encoder-decoder space. This alignment not only improves multilingual generation quality, but also indirectly facilitates more effective utilization of multilingual encoder signals by the LLM.

3.3 Three-Stage Training Strategy

To ensure stable optimization across models and objectives, XBridge employs a three-stage training strategy that progressively aligns heterogeneous representations and adapts the model to downstream tasks, keeping the LLM frozen throughout.

Stage 1: Cross-Model Mapping Due to the substantial representation gaps between the multilingual encoder and the LLM, as well as between the LLM and the multilingual decoder, directly bridging heterogeneous components is non-trivial. We therefore first establish coarse-grained semantic alignment among the multilingual encoder, the LLM, and the multilingual decoder using trilingual translation data $(\mathbf{x}, \mathbf{z}, \mathbf{y})$, where $\mathbf{z}$ is an English sequence generated by the LLM. In this stage, only the encoder-side mapping $\texttt{Mapping}_{\texttt{enc}}$, the decoder-side mapping $\texttt{Mapping}_{\texttt{dec}}$, and the decoder cross-attention layers are trained, optimizing the LLM English generation loss, the multilingual decoder generation loss, and the optimal transport alignment loss. This stage enables the LLM to interpret multilingual encoder representations and allows the decoder to attend to LLM hidden states for multilingual generation.

Stage 2: Encoder-Side Adaptation After cross-model semantic alignment is established, the second stage adapts multilingual input representations to downstream instruction-following tasks. We fine-tune only the encoder-side mapping layer $\texttt{Mapping}_{\texttt{enc}}$ on task-specific instruction data by optimizing the LLM English generation loss, while keeping all decoder-related components frozen. This stage teaches the LLM how to use multilingual representations to perform tasks, building upon the aligned representation space learned in stage 1.

Stage 3: Decoder-Side Adaptation The third stage focuses on improving multilingual generation quality by adapting the LLM-decoder interface. We update only $\texttt{Mapping}_{\texttt{dec}}$ and the decoder cross-attention layers, optimizing the multilingual decoder generation loss together with the optimal transport alignment loss. Separating this stage from stage 2 avoids conflicts between LLM and decoder objectives: stage 2 first stabilizes the conditional distribution of the LLM outputs, which stage 3 then exploits to enhance decoder performance without degrading task understanding.

Training Objectives Given encoder input sequence $\mathbf{x}$ with encoder representations $\mathbf{H}^{\mathbf{x}}$, the LLM-generated English sequence $\mathbf{z}$ with penultimate-layer hidden states $\mathbf{H}^{\mathbf{z}'}$, decoder-mapped representations $\tilde{\mathbf{H}}^{\mathbf{z}'}$, and multilingual decoder output sequence $\mathbf{y}$, the cross-entropy losses of LLM and decoder are defined as:

$$\mathcal{L}_{\text{CE\_LLM}} = -\log p_{\text{LLM}}(\mathbf{z} \mid \mathbf{x}, \text{inst}). \tag{2}$$

System	Low-Resource Languages				High-Resource Languages				Average	
	Bn-En	En-Bn	Sw-En	En-Sw	Ja-En	En-Ja	De-En	En-De	X-En	En-X
NLLB-200-1.3B	37.78	32.83	42.66	36.28	29.60	19.07	46.23	39.91	37.51	31.00
MetaMath-7B										
MetaMath-7B	1.46	0.67	3.33	1.75	**27.62**	16.76	34.36	19.42	18.62	11.92
MindMerger	30.76	-	39.43	-	22.50	-	40.05	-	31.57	-
LayAlign	30.91	-	39.02	-	22.36	-	39.43	-	31.98	-
XBridge (Ours)	**35.47**	**29.23**	**42.02**	**34.28**	24.52	**19.60**	**41.42**	**35.39**	**33.37**	**29.80**
LLaMA3-8B										
LLaMA3-8B	29.83	13.18	35.87	19.31	**27.71**	**25.40**	45.28	**36.24**	35.19	27.36
MindMerger	33.86	-	41.81	-	25.48	-	42.52	-	33.88	-
LayAlign	32.95	-	41.35	-	24.62	-	41.29	-	33.18	-
XBridge (Ours)	**37.09**	**28.42**	**44.73**	**34.68**	27.63	20.12	**45.75**	35.45	**36.21**	**29.82**
Aya-23-8B										
Aya-23-8B	8.59	2.43	7.89	1.16	**29.11**	**29.34**	**45.46**	**38.03**	28.13	23.71
MindMerger	33.41	-	41.56	-	24.96	-	41.78	-	33.44	-
LayAlign	32.42	-	40.22	-	24.16	-	41.44	-	32.92	-
XBridge (Ours)	**34.67**	**28.00**	**42.88**	**34.25**	26.35	19.14	44.40	33.78	**33.70**	**28.85**
Qwen2.5-7B-Instruct										
Qwen2.5-7B-Instruct	22.15	8.30	15.05	4.35	25.92	**25.76**	42.32	32.10	30.21	24.75
MindMerger	34.20	-	42.75	-	25.43	-	43.46	-	**34.71**	-
LayAlign	33.39	-	41.26	-	**26.11**	-	42.12	-	34.02	-
XBridge (Ours)	**35.89**	**27.59**	**43.24**	**34.55**	25.50	18.66	**44.55**	**33.02**	34.69	**28.64**

Table 1: FLORES-101 translation results for stage 1. For clarity, we report results on two low-resource languages (Bengali, Swahili) and two high-resource languages (Japanese, German), with complete results and COMET scores in Appendix C. "X" denotes all languages except for English. We bold the best scores for each LLM group.

$$\mathcal{L}_{\text{CE\_Dec}} = -\log p_{\text{Dec}}(\mathbf{y} \mid \tilde{\mathbf{H}}^{\mathbf{z}'}, \langle y \rangle). \quad (3)$$

Across stages, the overall training objective is:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{CE\_LLM}} + \lambda_2 \mathcal{L}_{\text{CE\_Dec}} + \lambda_3 \mathcal{L}_{\text{OT}}. \quad (4)$$

where different loss terms are activated depending on the training stage, as illustrated in Figure 2.

4 Experiment

4.1 Experiment Setup

Base Models We evaluate XBridge on four representative base LLMs: MetaMath-7B-V1.0 (Yu et al., 2024), LLaMA3-8B (Grattafiori et al., 2024), Aya-23-8B (Üstün et al., 2024), and Qwen2.5-7B-Instruct (Qwen et al., 2025). As the pretrained encoder-decoder NMT model, we adopt NLLB-200-1.3B (Team et al., 2022), which covers 200 languages with strong multilingual capacity.

Baselines We compare XBridge with these strong baselines: (1) **SFT** performs multilingual instruction fine-tuning directly on each base LLM. (2) **Translate-Test** (Artetxe et al., 2023) translates inputs to English, queries the English-SFT LLM, and translates the output back to the target language. (3) **MindMerger** (Huang et al., 2024) augments the LLM input with a pretrained multilingual encoder to enhance multilingual understanding, forming a strong multilingual-to-English system. (4) **LayAlign** (Ruan et al., 2025) further extends MindMerger with layer-wise fusion strategies to better integrate encoder representations into the LLM.

Language Setup Following Chen et al. (2024), we experiment on ten languages: Bengali (Bn), German (De), English (En), Spanish (Es), French (Fr), Japanese (Ja), Russian (Ru), Swahili (Sw), Thai (Th), and Chinese (Zh). These languages span diverse language families and resource levels. We treat Bn, Sw, and Th as low-resource languages, and the remaining as high-resource ones.

Training Datasets For stage 1 training, we extract English-centric translation pairs from OPUS-100 (Zhang et al., 2020). For XBridge, we further translate the English sentences into other languages L_y using NLLB-200-3.3B, constructing trilingual *x-en-y* data. For stage 2 and stage 3, we adopt multilingual mathematical reasoning data from Ruan et al. (2025) and multilingual abstractive summarization data from XL-Sum (Hasan et al., 2021). For XBridge, we construct bilingual responses using NLLB-200-3.3B. Appendix B presents details

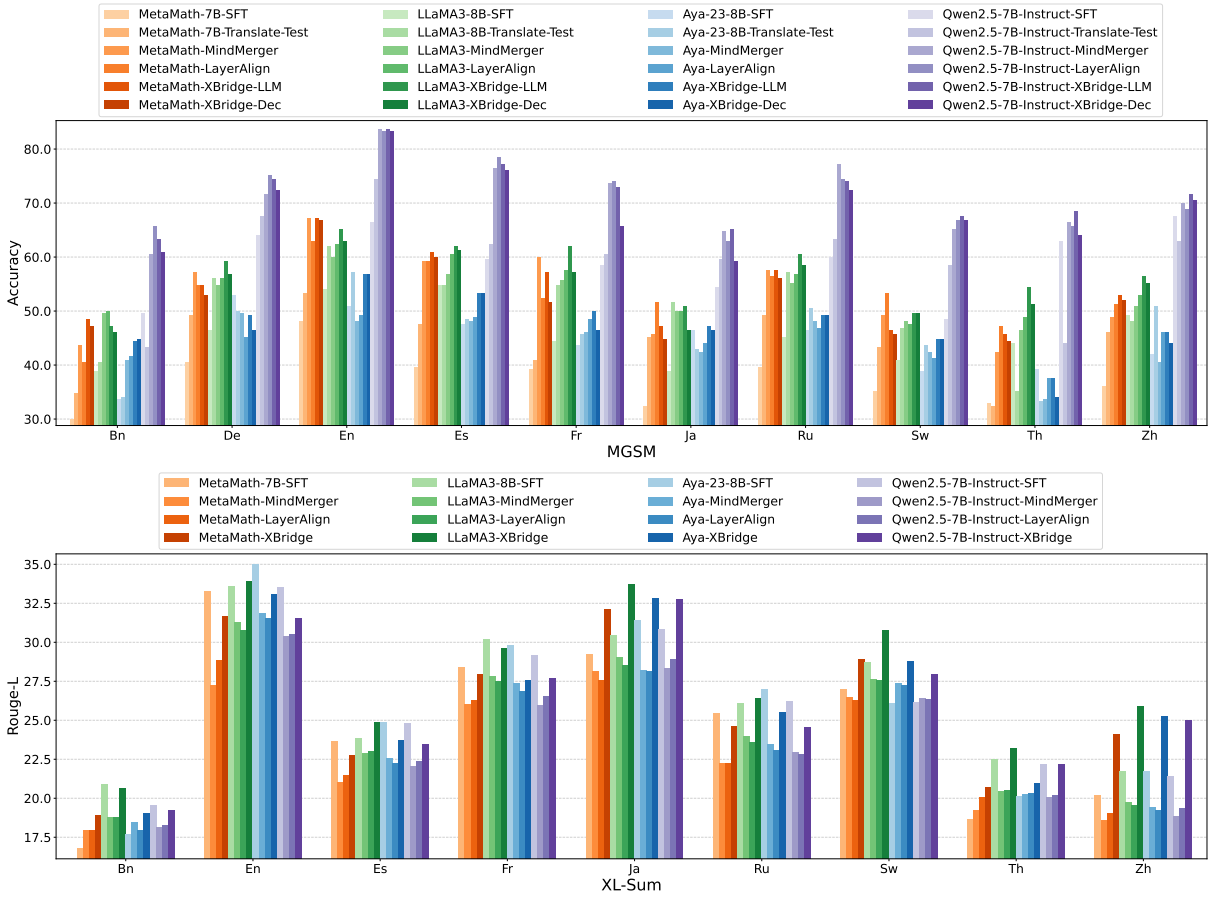


Figure 3: Multilingual reasoning accuracy on MGSM and multilingual summarization Rouge-L on XL-Sum, with complete results in Appendix C. Models with the same base LLM share the same color scheme, where lighter shades denote baselines and darker shades denote XBridge. "XBridge-LLM" refers to English reasoning by the LLM, while "XBridge-Dec" refers to multilingual reasoning by the composed decoder. For XL-Sum, since the baselines produce English-only summaries, we translate them into target languages using NLLB-200-1.3B for evaluation.

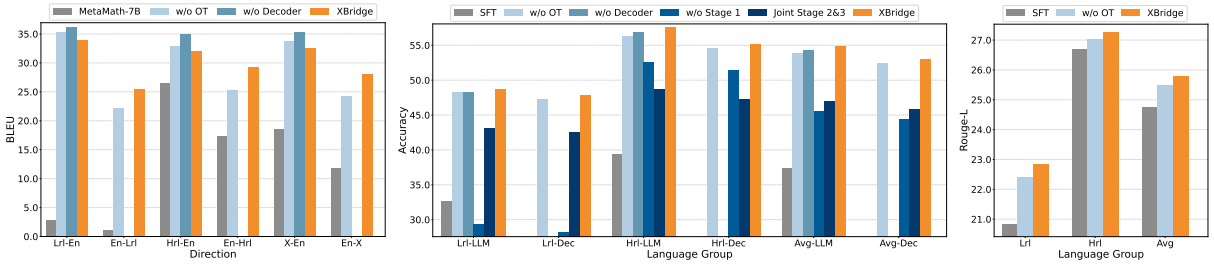
about data processing and statistics.

Evaluation Benchmarks For stage 1, we evaluate cross-model mapping quality on FLORES-101 (Goyal et al., 2022). Given the strong English ability of LLMs, we use *x-en* and *en-x* translation performance to measure multilingual understanding and generation, respectively, and report BLEU (Papineni et al., 2002) and COMET (Rei et al., 2020) scores. For base LLMs, we leverage *MMT-LLM* (Zhu et al., 2024) framework to evaluate translation capability in a 1-shot setting. For stage 2 and stage 3, we evaluate multilingual reasoning on MGSM (Shi et al., 2023) with Accuracy, and multilingual abstractive summarization on XL-Sum with multilingual Rouge-L (Lin, 2004).

Model Configuration and Training Details The encoder-side mapping is implemented as a two-layer multi-layer perceptron (MLP), while the decoder-side mapping is a four-layer MLP composed of two stacked two-layer MLP blocks. All intermediate dimensions are aligned with the LLM hidden size. We use the AdamW optimizer with a learning rate of 2×10^{-5}, train each stage for 3 epochs with a batch size of 128, and conduct experiments on 8 NVIDIA H800 GPUs. We empirically set $\lambda_1 = 1.0$, $\lambda_2 = 1.0$, and $\lambda_2 = 6.0$ when the corresponding losses are active, with detailed activation schedules described in Section 3.3.

4.2 Experimental Results

XBridge effectively offloads multilingual capability to the external multilingual model, while preserving the LLM as a knowledge and reasoning core. Table 1 evaluates the cross-model mapping learned in stage 1 on FLORES-101. Across all base LLMs, XBridge substantially improves both multilingual understanding and generation, with especially large gains on low-resource languages where base LLMs have limited capability. The



(a) FLORES-101 (b) MGSM (c) XL-Sum

Figure 4: Ablation analysis of XBridge. We compare different ablated variants of XBridge: encoder-only augmentation "w/o Decoder", loss ablation "w/o OT", removal of stage 1 "w/o Stage 1", and joint optimization of stage 2&3 "Joint Stage 2&3". “Lrl”, “Hrl”, and “Avg” denote low-, high-resource, and average performance, respectively.

performance of XBridge approaches that of the external NLLB-200-1.3B and outperforms encoder-augmented baselines, showing that XBridge can effectively offload multilingual ability to external NMT models while keeping the LLM frozen as a knowledge and reasoning core. Importantly, performance on high-resource languages remains comparable to base LLMs, indicating that offloading does not degrade the original strengths of LLMs.

Encoder adaptation improves multilingual understanding without degrading English performance. Figure 3 presents multilingual reasoning accuracy on MGSM after encoder adaptation. XBridge outperforms the base LLM, encoder-only baselines, and the Translate-Test pipeline. Since MGSM accuracy is language-agnostic, these gains directly reflect better semantic transfer between multilingual encoder representations and the LLM reasoning space. These results indicate that encoder-side adaptation facilitates more effective utilization of multilingual representations by the LLM, improving multilingual reasoning without sacrificing its English-centric reasoning capability.

Decoder adaptation achieves faithful multilingual generation. We further evaluate decoder adaptation on MGSM and XL-Sum in Figure 3. On MGSM, decoder-generated multilingual reasoning (XBridge_Dec) achieves accuracy comparable to English LLM outputs, suggesting that the decoder can faithfully express reasoning content across languages. On XL-Sum, XBridge consistently outperforms encoder-augmented baselines and achieves better average performance than the SFT baseline, with particularly clear gains on languages where multilingual generation is more challenging. While translation-cascaded systems are limited by the NMT model, XBridge directly leverages the LLM’s knowledge through decoder adaptation, resulting in more stable multilingual generation across languages. These results demonstrate the importance of decoder adaptation for robust multilingual generation.

5 Analysis

5.1 Ablation Analysis

We conduct the ablation study on MetaMath-7B-V1.0 to analyze the contribution of each component and training strategies in XBridge, and evaluate ablated variants on FLORES-101, MGSM, and XL-Sum. Figure 4 presents the results, and Appendix C provides detailed results.

Encoder-Decoder Collaboration Removing the decoder (*w/o Decoder*) achieves competitive multilingual-to-English understanding but fails to support multilingual generation, and underperforms XBridge on MGSM. This confirms that encoder-only augmentation is insufficient for multilingual reasoning and generation.

OT Alignment Objectives Similarly, removing the OT alignment (*w/o OT*) leads to performance degradation on all benchmarks, particularly for multilingual generation, indicating that token-level soft alignment plays a crucial role in bridging heterogeneous representation spaces between the LLM and the multilingual decoder.

Stage-Wise Optimization Skipping stage 1 (*w/o Stage 1*) results in a substantial performance drop across all metrics, suggesting that direct task-level adaptation is insufficient when the representation gap between the LLM and the multilingual model remains large. Moreover, jointly training stage 2 and stage 3 (*Joint Stage 2&3*) underperforms the stage-wise optimization, reflecting a trade-off be-

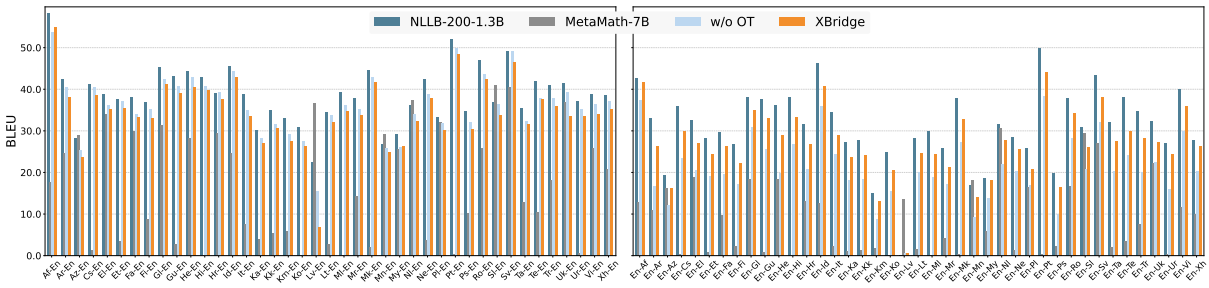
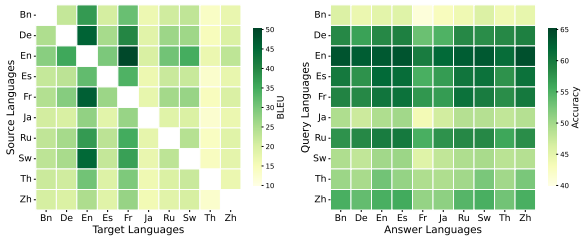
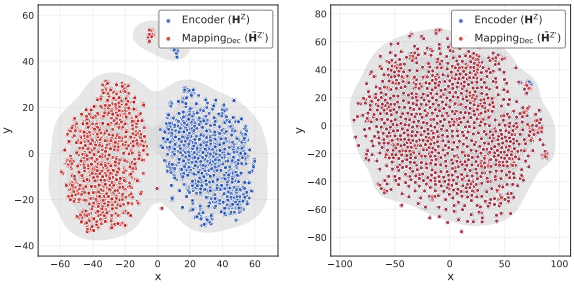


Figure 5: Cross-lingual generalization to 41 untuned languages in FLORES-101. Left: *X*→*En* direction. Right: *En*→*X* direction. We directly evaluate the ablation variants described in Section 5.1. Appendix C lists the included untuned languages and provides detailed results.



(a) FLORES-101 (b) MGSM

Figure 6: Evaluation for language-on-demand generation. Appendix C presents detailed results.



(a) w/o OT (b) w/ OT

Figure 7: Visualization of sentence-level representation alignment for Chinese (Zh). We compare models trained without OT (left) and with OT (right) using t-SNE.

tween LLM- and decoder-side generation objectives. These results support the design of stage-wise adaptation, where coarse-grained cross-model alignment is first established, followed by fine-grained encoder and decoder specialization, enabling XBridge to achieve stable and effective multilingual reasoning and generation.

5.2 Generalization to Untuned Languages

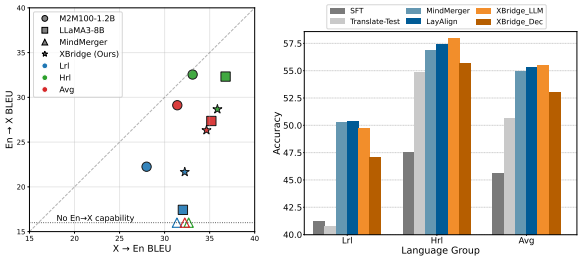
To examine whether the cross-model mappings learned by XBridge are language-agnostic rather than simply tied to specific training languages, we evaluate cross-model cross-lingual transfer on 41 untuned languages (listed in Table 3) in Figure 5, based on variants of Section 5.1.

XBridge yields substantial gains on untuned languages over the base LLM, with performance approaching the external NLLB model. This indicates that stage 1 cross-model mapping learns language-agnostic semantic transfer that generalizes beyond tuned languages, rather than language-specific mapping. Meanwhile, performance in the *En*→*X* directions highlights the importance of optimal transport. Removing the OT objective leads to a substantial drop in generation quality, particularly where tokenization length differs across different tokenizers. These results suggest that OT enables robust alignment between heterogeneous tokenizations, which is crucial for generalizable multilingual generation. Overall, the results demonstrate that cross-model semantic alignment generalizes across languages, while OT is crucial for achieving reliable generation-level generalization.

5.3 Language-on-Demand Generation

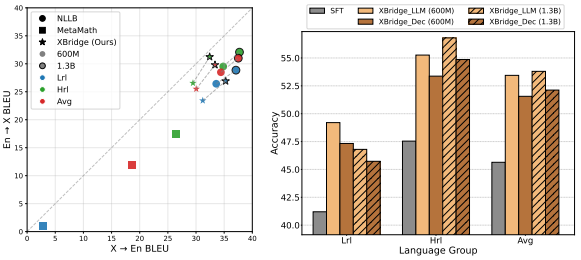
We verify the language-on-demand property of XBridge by switching the target language token $\langle y \rangle$ to generate outputs in arbitrary languages without retraining in Figure 6.

On FLORES-101, we evaluate translation between all language pairs. With the target language fixed, changing the source language causes only minor performance differences, while variations primarily depend on the target language. On MGSM, we force the decoder to generate responses in languages different from the input query language. For each input language, performance remains largely stable across different output languages. These results indicate that XBridge enables stable language-on-demand generation, supporting flexible multilingual outputs while preserving a language-agnostic reasoning core in the LLM.



(a) FLORES-101 (b) MGSM

Figure 8: LLaMA3-8B composed with M2M100-1.2B. "Lrl", "Hrl", and "Avg" denote low-resource, high-resource, and average performance, respectively. Hollow markers placed on the bottom boundary indicate models that lack *En*→*X* translation capability. Appendix C presents detailed results.



(a) FLORES-101 (b) MGSM

Figure 9: MetaMath-7B composed with NLLB in different sizes (600M vs. 1.3B). "Lrl", "Hrl", and "Avg" denote low-, high-resource, and average performance, respectively. Appendix C presents detailed results.

5.4 Representation Visualization

To analyze the effect of optimal transport (OT) on aligning heterogeneous representations, we visualize sentence-level hidden states for each language. Specifically, we compare encoder representations of LLM English outputs $\mathbf{H}^{\mathbf{z}}$ with decoder-side representations after mapping $\tilde{\mathbf{H}}^{\mathbf{z}'}$. We obtain sentence-level vectors via average pooling and project them to 2-dim for visualization using t-SNE (Van der Maaten and Hinton, 2008).

As shown in Figure 7, without OT, the two sets of representations form largely separate clusters, reflecting a substantial distribution gap at the LLM-decoder interface. In contrast, when OT is applied, the two sets of representations overlap substantially, with density contours largely merged, indicating that OT promotes fine-grained semantic consistency and reduces token-level misalignment across model components.

5.5 Composing with Different NMT Models

To further examine the generality of XBridge beyond a specific NMT backbone, we replace the multilingual NMT model with M2M100-1.2B (Fan et al., 2021) in Figure 8 while keeping the same training and evaluation settings as in Section 4.

XBridge remains effective when composed with M2M100-1.2B. On FLORES-101, XBridge achieves strong cross-model transfer across low- and high-resource language directions, demonstrating that the lightweight mapping layers can reliably bridge NMT models and LLMs. On MGSM, XBridge outperforms the translation-cascaded baseline, indicating that the benefits of XBridge extend beyond translation quality to multilingual reasoning. Overall, these results demonstrate that XBridge is an architecture-agnostic framework that generalizes across both different LLM backbones and multilingual NMT backbones.

5.6 Impact of NMT Model Size

We further investigate the impact of NMT model capacity on XBridge by comparing NLLB-200-600M and NLLB-200-1.3B, both integrated with MetaMath-7B. Figure 9 presents the results.

On FLORES-101, larger NLLB models consistently improve multilingual capability across languages, indicating that stronger multilingual capacity in the composed NMT model leads to better multilingual understanding and generation. On MGSM, increasing the NLLB size brings only marginal changes in reasoning accuracy, suggesting that reasoning performance is primarily determined by the LLM core. These results align with our design, indicating that the quality of the composed NMT model directly influences cross-model mapping, while reasoning remains governed by the LLM.

5.7 Supplementary Analysis

We conduct supplementary analysis, including efficiency analysis (Appendix D.1), the case study on MGSM (Appendix D.2), and evaluation on multilingual commonsense reasoning (Appendix D.3).

6 Conclusion

In this paper, we propose XBridge, a compositional framework that offloads multilingual capability to an external encoder-decoder NMT model, while preserving the LLM as an English-centric core for general knowledge processing. Extensive experiments demonstrate that XBridge enables efficient multilingual extension, raising low-resource and unseen language performance to near external NMT, without compromising LLM's core abilities.

Limitations

While XBridge substantially mitigates multilingual imbalance, notably improving performance on low-resource and previously unseen languages for LLMs, the overall model still exhibits some imbalance in multilingual capabilities. This is primarily due to the combined influence of the external encoder-decoder NMT model and the base LLM, which limits complete uniformity across languages. Future work could further explore strategies to harmonize these components.

Acknowledgements

We thank all the anonymous reviewers for their insightful and valuable comments on this paper. This work was supported by the grant from the Beijing Natural Science Foundation (No. L257006).

References

Mikel Artetxe, Vedanuj Goswami, Shruti Bhosale, Angela Fan, and Luke Zettlemoyer. 2023. Revisiting machine translation for cross-lingual classification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6489–6499, Singapore. Association for Computational Linguistics.

Linzheng Chai, Jian Yang, Tao Sun, Hongcheng Guo, Jiaheng Liu, Bing Wang, Xinnian Liang, Jiaqi Bai, Tongliang Li, Qiyao Peng, and 1 others. 2025. xcot: Cross-lingual instruction tuning for cross-lingual chain-of-thought reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23550–23558.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Ben Bergen. 2024. When is multilinguality a curse? language modeling for 250 high- and low-resource languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4074–4096, Miami, Florida, USA. Association for Computational Linguistics.

Nuo Chen, Zinan Zheng, Ning Wu, Ming Gong, Dongmei Zhang, and Jia Li. 2024. Breaking language barriers in multilingual mathematical reasoning: Insights and observations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7001–7016.

Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, and 1 others. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Changjiang Gao, Hongda Hu, Peng Hu, Jiajun Chen, Jixing Li, and Shujian Huang. 2024. Multilingual pretraining and instruction tuning improve cross-lingual knowledge alignment, but only shallowly. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6101–6117, Mexico City, Mexico. Association for Computational Linguistics.

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Zixian Huang, Wenhao Zhu, Gong Cheng, Lei Li, and Fei Yuan. 2024. Mindmerger: Efficiently boosting llm reasoning in non-english languages. *Advances in Neural Information Processing Systems*, 37:34161–34187.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023. Bactrian-x: Multilingual replicable instruction-following models with low-rank adaptation. *arXiv preprint arXiv:2305.15011*.

Bill Yuchen Lin, Seyeon Lee, Xiaoyang Qiao, and Xiang Ren. 2021. Common sense beyond English: Evaluating and improving multilingual language models for commonsense reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1274–1287, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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.

Gabriel Peyré, Marco Cuturi, and 1 others. 2019. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2695–2709, Singapore. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Zhiwen Ruan, Yixia Li, He Zhu, Longyue Wang, Weihua Luo, Kaifu Zhang, Yun Chen, and Guanhua Chen. 2025. LayAlign: Enhancing multilingual reasoning in large language models via layer-wise adaptive fusion and alignment strategy. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1481–1495, Albuquerque, New Mexico. Association for Computational Linguistics.

Adriaan MJ Schakel and Benjamin J Wilson. 2015. Measuring word significance using distributed representations of words. *arXiv preprint arXiv:1508.02297*.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, and 1 others. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, and 1 others. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Yujia Xie, Xiangfeng Wang, Ruijia Wang, and Hongyuan Zha. 2020. A fast proximal point method for computing exact wasserstein distance. In *Uncertainty in artificial intelligence*, pages 433–453. PMLR.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Sho Yokoi, Ryo Takahashi, Reina Akama, Jun Suzuki, and Kentaro Inui. 2020. Word rotator's distance. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2944–2960.

Dongkeun Yoon, Joel Jang, Sungdong Kim, Seungone Kim, Sheikh Shafayat, and Minjoon Seo. 2024. LangBridge: Multilingual reasoning without multilingual supervision. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7502–7522, Bangkok, Thailand. Association for Computational Linguistics.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng YU, Zhengying Liu, Yu Zhang, James Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2024. Metamath: Bootstrap your own mathematical questions for large language models. In *The Twelfth International Conference on Learning Representations*.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Shaolei Zhang, Qingkai Fang, Zhuocheng Zhang, Zhengrui Ma, Yan Zhou, Langlin Huang, Mengyu Bu, Shangtong Gui, Yunji Chen, Xilin Chen, and 1 others. 2023. Bayling: Bridging cross-lingual alignment and instruction following through interactive translation for large language models. *arXiv preprint arXiv:2306.10968*.

Shaolei Zhang, Kehao Zhang, Qingkai Fang, Shoutao Guo, Yan Zhou, Xiaodong Liu, and Yang Feng. 2024a. Bayling 2: A multilingual large language model with efficient language alignment. *arXiv preprint arXiv:2411.16300*.

Shimao Zhang, Zhejian Lai, Xiang Liu, Shuaijie She, Xiao Liu, Yeyun Gong, Shujian Huang, and Jiajun Chen. 2025a. How does alignment enhance llms' multilingual capabilities? a language neurons perspective. *Preprint*, arXiv:2505.21505.

Xue Zhang, Yunlong Liang, Fandong Meng, Songming Zhang, Yufeng Chen, Jinan Xu, and Jie Zhou. 2025b. Less, but better: Efficient multilingual expansion for LLMs via layer-wise mixture-of-experts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17948–17963, Vienna, Austria. Association for Computational Linguistics.

Yuanchi Zhang, Yile Wang, Zijun Liu, Shuo Wang, Xiaolong Wang, Peng Li, Maosong Sun, and Yang Liu. 2024b. Enhancing multilingual capabilities of large language models through self-distillation from resource-rich languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11189–11204, Bangkok, Thailand. Association for Computational Linguistics.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

Wenhao Zhu, Yunzhe Lv, Qingxiu Dong, Fei Yuan, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Extrapolating large language models to non-english by aligning languages. *arXiv preprint arXiv:2308.04948*.

A Optimal Transport Algorithm for Heterogeneous Representations

In this section, we briefly review the standard OT formulation and describe how it is adapted to align heterogeneous and unequal-length representation sequences in XBridge.

A.1 Optimal Transport Between Discrete Distributions

Optimal Transport (OT) provides a principled framework for measuring the discrepancy between two probability distributions by minimizing the cost of transporting probability mass from one distribution to the other (Peyré et al., 2019). Consider the following discrete transport problem: given two probability distributions P and Q,

$$\begin{aligned} P &= \{(w_i, m_i)\}_{i=1}^{n}, \quad \text{s.t.} \sum_{i=1}^{n} m_i = 1, \\ Q &= \{(w'_j, m'_j)\}_{j=1}^{n'}, \quad \text{s.t.} \sum_{j=1}^{n'} m_j = 1. \end{aligned} \tag{5}$$

where each support point $w_i, w'_j \in \mathbb{R}^d$ is associated with a non-negative probability mass m_i, m'_j. Given a cost function $c(w_i, w'_j)$ that measures the unit cost of transporting mass from w_i to w'_j, the transport cost between P and Q is defined as:

$$\begin{aligned} \mathcal{D}(P, Q) = \min_{\mathbf{T} \geq 0} \sum_{i,j} \mathbf{T}_{ij}\, c(w_i, w'_j), \\ \text{s.t.} \sum_{j=1}^{n'} \mathbf{T}_{ij} = m_i, \forall i \in \{1, \ldots, n\}, \\ \sum_{i=1}^{n} \mathbf{T}_{ij} = m'_j, \forall j \in \{1, \ldots, n'\}. \end{aligned} \tag{6}$$

where $\mathbf{T}_{ij}$ denotes the mass transported from w_i to w'_j.

A.2 OT for Aligning Unequal-Length Representation Sequences

In XBridge, we apply OT to align two heterogeneous token representation sequences:

$$\begin{aligned} \mathbf{H}^{\mathbf{z}} &= (H_1^{\mathbf{z}}, \ldots, H_k^{\mathbf{z}}), \\ \tilde{\mathbf{H}}^{\mathbf{z}'} &= (\tilde{H}_1^{\mathbf{z}'}, \ldots, \tilde{H}_m^{\mathbf{z}'}). \end{aligned} \tag{7}$$

where $k \neq m$ in general due to different tokenization schemes. Both sequences originate from the same underlying LLM output but are obtained

through different encoding pathways, making explicit token-wise correspondence unavailable.

We formulate their alignment as the following OT problem:

$$\begin{aligned}\mathcal{D}(\mathbf{H}^{\mathbf{z}}, \tilde{\mathbf{H}}^{\mathbf{z}'}) &= \min_{\mathbf{T}\geq 0} \sum_{i,j} \mathbf{T}_{ij}\, c(H_i^{\mathbf{z}}, \tilde{H}_j^{\mathbf{z}'}), \\ \text{s.t.} \sum_{j=1}^{m} \mathbf{T}_{ij} &= m_i^{\mathbf{z}}, \forall i \in \{1,\dots,k\}, \\ \sum_{i=1}^{k} \mathbf{T}_{ij} &= m_j^{\mathbf{z}'}, \forall j \in \{1,\dots,m\}.\end{aligned} \tag{8}$$

where the cost function $c(\cdot,\cdot)$ is defined as cosine distance.

The probability masses $m_i^{\mathbf{z}}$ and $m_j^{\mathbf{z}'}$ are obtained by normalizing the ℓ_1 norms of the corresponding representations. This choice is motivated by prior work (Schakel and Wilson, 2015; Yokoi et al., 2020) showing that embedding norms correlate with token importance, with semantically salient words exhibiting larger magnitudes.

A.3 Approximate OT via Relaxed Marginal Constraints

Solving the exact OT problem requires $O(n^3)$ linear programming, which is computationally prohibitive for long sequences. While entropic regularization methods such as Sinkhorn (Cuturi, 2013) or IPOT (Xie et al., 2020) provide approximate solutions, they still introduce significant overhead during training.

Following Kusner et al. (2015), we adopt a relaxed OT formulation by removing the second marginal constraint:

$$\begin{aligned}\mathcal{D}^*(\mathbf{H}^{\mathbf{z}}, \tilde{\mathbf{H}}^{\mathbf{z}'}) &= \min_{\mathbf{T}\geq 0} \sum_{i,j} \mathbf{T}_{ij}\, c(H_i^{\mathbf{z}}, \tilde{H}_j^{\mathbf{z}'}), \\ \text{s.t.} \sum_{j=1}^{m} \mathbf{T}_{ij} &= m_i^{\mathbf{z}}, \quad \forall i \in \{1,\dots,k\}.\end{aligned} \tag{9}$$

This relaxation yields a lower bound of the exact OT distance and admits a closed-form solution: each representation $H_i^{\mathbf{z}}$ transports all its probability mass to the most similar $\tilde{H}_j^{\mathbf{z}'}$ under the cosine distance. The resulting transport plan naturally supports unequal-length alignments, making it well-suited for sequences with heterogeneous tokenizations.

A.4 Role in XBridge

The proposed OT alignment provides a principled mechanism for aligning heterogeneous representations without assuming positional correspondence. Moreover, since the multilingual encoder is frozen during training, the relaxed OT objective anchors the alignment to encoder-defined semantic geometry, encouraging decoder-side representations to remain compatible with the multilingual encoder-decoder space. Despite its simplicity, this approximation is sufficient for our setting, as the goal is semantic compatibility regularization rather than exact distribution matching.

B Details for Training Data

Translation Data in Stage 1 We sample English-centric translation pairs from OPUS-100 (Zhang et al., 2020) and filter the off-target pairs, with 50k samples per translation direction. For XBridge, we further translate English sentences into other languages L_y using NLLB-200-3.3B to construct trilingual *x-en-y* data. To mitigate translation noise in generation, we train XBridge using *y-en-x*, where the encoder processes translated sentences and the decoder processes natural sentences.

Multilingual Reasoning Data and Multilingual Abstractive Summarization Data in Stage 2 and Stage 3 We extract multilingual reasoning data from Ruan et al. (2025), which contains 30K samples per language across ten languages (the same as in Section 4.1). We extract multilingual abstractive summarization data from XL-Sum (Hasan et al., 2021). XL-Sum contains imbalanced multilingual data, and we have set the data upper limit to 30K. For XBridge, we additionally construct bilingual responses using NLLB-200-3.3B.

Data Statistics Figure 10 presents detailed data statistics for the training data.

C Detailed Results

Table 4, Table 6 and Table 7 present detailed BLEU scores on FLORES-101 (COMET scores in Table 5), accuracy on MGSM, and multilingual Rouge-L on XL-Sum for the main experiments in Section 4.

Table 8, Table 9 and Table 10 present results for the ablation study in Section 5.1.

Table 3 presents the included untuned languages and corresponding language codes. Table 11 presents detailed results for untuned language generalization in Section 5.2.

Table 12 presents BLEU scores for cross-lingual generation on FLORES-101, and Table 13 presents

Task	Languages	Data Composition	Total Size
Translation	Bn, De, En, Es, Fr, Ja, Ru, Sw, Th, Zh	50K * 72 translation directions	3.6M
Multilingual Reasoning	Bn, De, En, Es, Fr, Ja, Ru, Sw, Th, Zh	30K * 10 language	300K
Multilingual Abstractive Summarization	Bn, En, Es, Fr, Ja, Ru, Sw, Th, Zh	Bn-8K, En-30K, Es-30K, Fr-9K, Ja-7K, Ru-30K, Sw-8K, Th-7K, Zh-30K	158K

Figure 10: Statistics of training datasets used in different stages.

System	Training	Inference
SFT	1.0x	1.0x
Translate-Test	-	0.55x
MindMerger	1.42x	0.85x
XBridge	0.91x	0.66x

Table 2: Relative speed comparison.

Accuracy for language-on-demand generation on MGSM in Section 5.3.

Table 14 and Table 15 present results for LLaMA3-8B composed with M2M100-1.2B in Section 5.5.

Table 17 and Table 16 present results for MetaMath-7B composed with NLLB-200 in different sizes (600M vs. 1.3B) in Section 5.6.

D Supplementary Analysis

D.1 Efficiency Analysis

We compare the training and inference efficiency of XBridge with SFT (LLM-only), MindMerger (Encoder-LLM), and the cascaded Translate-Test pipeline in Table 2. XBridge introduces only a limited training overhead despite the additional encoder and decoder, due to its parameter-efficient design. For inference, XBridge is slower than the LLM-only method due to the additional decoding for multilingual generation, but it remains faster than the cascaded Translate-Test pipeline. Overall, XBridge trades moderate computational cost for improved multilingual generation quality and robustness, while avoiding the inefficiency and error accumulation of cascaded systems.

D.2 Case Study on MGSM

In the case study, we compare the outputs of MindMerger, LayAlign, and XBridge in Figure 11. MindMerger and LayAlign adopt encoder-augmented architectures, which only enable multilingual-to-English processing. As a result, the generated responses are always in English, which is less friendly for multilingual users who expect outputs in their native languages. In contrast, XBridge supports controllable target-language generation by explicitly specifying the decoder language token, allowing the model to produce responses in different languages as required. This demonstrates the advantage of the compositional encoder-LLM-decoder design in providing flexible multilingual generation.

D.3 Multilingual Commonsense Reasoning

We further extend XBridge to X-CSQA (Lin et al., 2021) to evaluate both cross-lingual understanding and logical reasoning. We report results on MetaMath-7B and LLaMA3-8B across eight languages in Figure 12. XBridge outperforms MindMerger and LayAlign on MetaMath-7B across most languages, yielding the highest average accuracy. Notably, the gains are more pronounced on non-English languages, while English performance is preserved. This suggests that XBridge effectively enhances multilingual understanding without degrading the base model's reasoning ability, and achieves more balanced multilingual reasoning performance across languages.

Query: 制作一件袍子需要 2 匹蓝色纤维布料和这个数量一半的白色纤维布料。它一共需要用掉多少匹布料?	
System	**Responses**
MindMerger	If 2 blue fibers are needed, then half that amount of white fibers is 2/2 = 1 white fiber.\nSo, in total, 2 blue fibers + 1 white fiber = 3 fibers are needed.\n#### 3\nThe answer is: 3
LayAlign	If it takes 2 yards of blue fabric to make a robe, then it takes 2/2 = 1 yard of white fabric.\nIn total, it takes 2 yards of blue fabric + 1 yard of white fabric = 3 yards of fabric.\n#### 3\nThe answer is: 3
XBridge	**<ben_Beng>** নীল ফাইবারের সংখ্যা ২। অর্ধেক সংখ্যা ২/২ = ১। সুতরাং, সাদা ফাইবারের সংখ্যা ১। সুতরাং, ফাইবারের মোট সংখ্যা ২ টি নীল ফাইবার + ১ টি সাদা ফাইবার = ৩ টি ফাইবার। #### 3 উত্তর হল: ৩
	<deu_Latn> Die Anzahl der gebrauchten blauen Stoffe beträgt 2. Die Hälfte dieser Anzahl ist 2/2 = 1. Also ist die Anzahl der benötigten weißen Stoffe 1 . Daher beträgt die Gesamtzahl der benötigten Stoffe 2 blaue Stoffe + 1 weiße Stoffe = 3 Stoffe.
	<eng_Latn> The number of blue fibers needed is 2. Half of that number is 2/2 = 1. So, the number of white fibers needed is 1. Therefore, the total number of fibers needed is 2 blue fibers + 1 white fiber = 3 fibers. #### 3 The answer is: 3
	<spa_Latn> El número de fibras azules necesarias es 2. La mitad de ese número es 2/2 = 1. Así, el número de fibras blancas necesarias es 1. Por lo tanto, el número total de fibras necesarias es 2 fibras azules + 1 fibra blanca = 3 fibras. #### 3 La respuesta es: 3
	<fra_Latn> Le nombre de fibres bleues nécessaires est de 2. La moitié de ce nombre est de 2/2 = 1. Donc, le nombre de fibres blanches nécessaires est de 1. Par conséquent, le nombre total de fibres nécessaires est de 2 fibres bleues + 1 fibres blanches = 3 fibres.
	<jpn_Jpan> 青い繊維の数は2です. その数の半分は2/2 = 1 です. したがって,必要な白い繊維の数は1です. したがって,必要な繊維の総数は2つの青い繊維+1つの白い繊維=3つの繊維です. #### 3 答えは: 3
	<rus_Cyrl> Количество синих тканей, необходимых, равняется 2. Половина этого числа равняется 2/2 = 1. Таким образом, количество белых тканей, необходимых, равняется 1. Таким образом, общее количество тканей, необходимых, равняется 2 синим тканям + 1 белому тканю = 3 тканям. #### 3 Ответ: 3
	<swh_Latn> Idadi ya nyuzi za bluu zinahitajika ni 2. Nusu ya idadi hiyo ni 2/2 = 1. Kwa hiyo, idadi ya nyuzi nyeupe zinahitajika ni 1. Kwa hiyo, jumla ya idadi ya nyuzi zinahitajika ni nyuzi 2 za bluu + nyuzi 1 nyeupe = nyuzi 3. #### 3 Jibu ni: 3
	<tha_Thai> จำนวนเซื้อสีฟ้าที่ต้องการคือ 2 ครึ่งหนึ่งของจำนวนนั้นคือ 2/2 = 1 ดังนั้น จำนวนเซื้อสีขาวที่ต้องการคือ 1 ดังนั้น จำนวนเซื้อที่ต้องการทั้งหมดคือ 2 เซื้อสีฟ้า + 1 เซื้อสีขาว = 3 เซื้อ
	<zho_Hans> 蓝色的数量是2. 需要的蓝色数量是2/2=1. 所以,需要的白色数量是1. 因此,所需的总数是2个蓝色的数量+1个白色的数量=3个. #### 3 答案是: 3

Figure 11: Case study on multilingual reasoning. Red text indicates the language token fed to the decoder, which controls the target generation language.

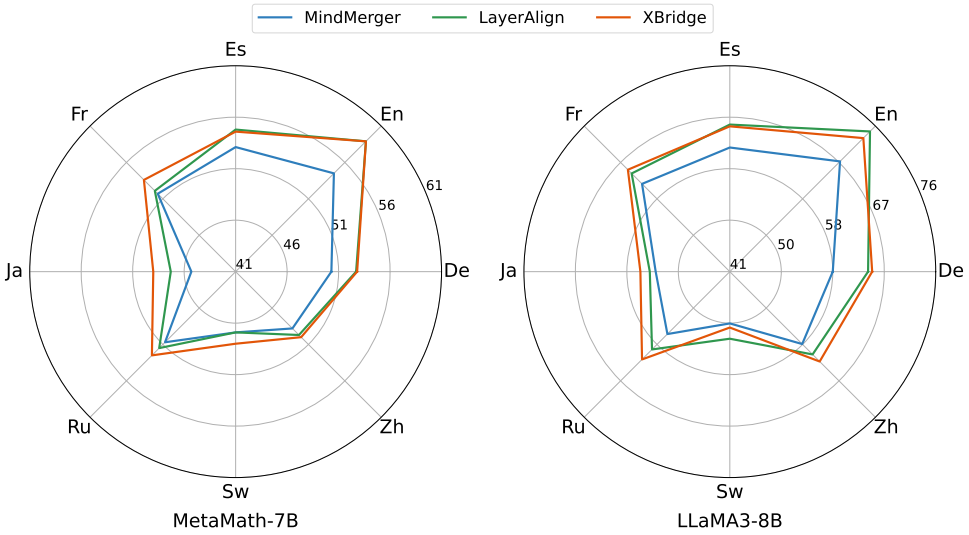


Figure 12: Radar plot comparison on X-CSQA.

ISO 639-1	Language
Af	Akrikaans
Ar	Arabic
Az	Azerbaijani
Cs	Czech
El	Modern Greek
Et	Estonian
Fa	Persian
Fi	Finnish
Gl	Galician
Gu	Gujarati
He	Hebrew
Hi	Hindi
Hr	Croatian
Id	Indonesian

ISO 639-1	Language
It	Italian
Ka	Georgian
Kk	Kazakh
Km	Khmer
Ko	Korean
Lt	Lithuanian
Lv	Latvian
Mk	Macedonian
Ml	Malayalam
Mn	Mongolian
Mr	Marathi
My	Burmese
Ne	Nepali
Nl	Dutch

ISO 639-1	Language
Pl	Polish
Ps	Pashto
Pt	Portuguese
Ro	Romanian
Sl	Slovenian
Sv	Swedish
Ta	Tamil
Te	Telugu
Tr	Turkish
Uk	Ukrainian
Ur	Urdu
Vi	Vietnamese
Xh	Xhosa

Table 3: The included 41 untuned languages and corresponding language codes.

System	Bn-En	En-Bn	De-En	En-De	Es-En	En-Es	Fr-En	En-Fr	Ja-En	En-Ja
NLLB-200-1.3B	37.78	32.83	46.23	39.91	35.37	31.35	47.14	51.33	29.60	19.07
MetaMath-7B										
MetaMath-7B	1.46	0.67	34.36	19.42	**36.33**	27.42	15.54	11.64	**27.62**	16.76
MindMerger	30.76	-	40.05	-	28.77	-	38.65	-	22.50	-
LayerAlign	30.91	-	39.43	-	30.16	-	40.48	-	22.36	-
XBridge (Ours)	**35.47**	**29.23**	**41.42**	**35.39**	30.51	**28.88**	**42.14**	**49.49**	24.52	**19.60**
LLaMA3-8B										
LLaMA3-8B	29.83	13.18	45.28	**36.24**	**33.65**	28.86	**46.18**	47.04	**27.71**	**25.40**
MindMerger	33.86	-	42.52	-	30.70	-	41.47	-	25.48	-
LayerAlign	32.95	-	41.29	-	30.36	-	39.40	-	24.62	-
XBridge (Ours)	**37.09**	**28.42**	**45.75**	35.45	32.00	**29.59**	46.10	**49.38**	27.63	20.12
Aya-23-8B										
Aya-23-8B	8.59	2.43	**45.46**	**38.03**	**33.22**	**30.50**	**46.17**	**49.45**	**29.11**	**29.34**
MindMerger	33.41	-	41.78	-	30.75	-	40.26	-	24.96	-
LayerAlign	32.42	-	41.44	-	30.75	-	39.86	-	24.16	-
XBridge (Ours)	**34.67**	**28.00**	44.40	33.78	31.36	28.71	42.05	48.47	26.35	19.14
Qwen2.5-7B-Instruct										
Qwen2.5-7B-Instruct	22.15	8.30	42.32	32.10	**32.15**	**29.31**	**43.86**	45.09	25.92	**25.76**
MindMerger	34.20	-	43.46	-	31.99	-	42.00	-	25.43	-
LayerAlign	33.39	-	42.12	-	31.17	-	40.58	-	**26.11**	-
XBridge (Ours)	**35.89**	**27.59**	**44.55**	**33.02**	31.52	28.22	42.64	**47.80**	25.50	18.66

System	Ru-En	En-Ru	Sw-En	En-Sw	Th-En	En-Th	Zh-En	En-Zh	X-En	En-X
NLLB-200-1.3B	38.06	33.72	42.66	36.28	30.82	17.46	29.90	17.07	37.51	31.00
MetaMath-7B										
MetaMath-7B	26.21	19.34	3.33	1.75	3.82	0.72	18.93	9.58	18.62	11.92
MindMerger	32.82	-	39.43	-	26.63	-	24.49	-	31.57	-
LayerAlign	34.28	-	39.02	-	26.56	-	**24.59**	-	31.98	-
XBridge (Ours)	**34.97**	**31.03**	**42.02**	**34.28**	**28.30**	**17.23**	20.97	**23.04**	**33.37**	**29.80**
LLaMA3-8B										
LLaMA3-8B	**37.31**	30.49	35.87	19.31	30.39	**19.80**	**30.46**	**25.90**	35.19	27.36
MindMerger	34.16	-	41.81	-	29.14	-	25.76	-	33.88	-
LayerAlign	34.32	-	41.35	-	28.66	-	25.63	-	33.18	-
XBridge (Ours)	37.08	**30.57**	**44.73**	**34.68**	**30.61**	17.09	24.89	23.11	**36.21**	**29.82**
Aya-23-8B										
Aya-23-8B	**37.10**	**33.01**	7.89	1.16	14.72	2.11	**30.89**	**27.32**	28.13	23.71
MindMerger	34.50	-	41.56	-	28.34	-	25.40	-	33.44	-
LayerAlign	34.64	-	40.22	-	27.62	-	25.15	-	32.92	-
XBridge (Ours)	33.74	28.90	**42.88**	**34.25**	**28.95**	**16.52**	18.94	21.87	**33.70**	**28.85**
Qwen2.5-7B-Instruct										
Qwen2.5-7B-Instruct	32.81	27.98	15.05	4.35	26.27	**19.89**	**31.34**	**29.99**	30.21	24.75
MindMerger	**35.47**	-	42.75	-	**29.52**	-	27.59	-	**34.71**	-
LayerAlign	34.97	-	41.26	-	28.91	-	27.66	-	34.02	-
XBridge (Ours)	35.06	**29.77**	**43.24**	**34.55**	29.01	16.26	24.78	21.89	34.69	**28.64**

Table 4: Detailed BLEU scores on FLORES-101 for stage 1. "X" denotes all languages except for English. We bold the best scores for each LLM group.

System	Bn-En	En-Bn	De-En	En-De	Es-En	En-Es	Fr-En	En-Fr	Ja-En	En-Ja
NLLB-200-1.3B	88.30	85.67	88.80	86.12	86.72	85.68	88.83	87.05	87.05	86.45
MetaMath-7B										
MetaMath-7B	50.85	35.10	86.25	76.59	84.39	78.49	86.84	78.45	82.62	76.35
MindMerger	86.87	-	87.81	-	85.52	-	87.64	-	84.94	-
LayerAlign	87.10	-	87.77	-	85.75	-	87.77	-	84.78	-
XBridge (Ours)	**88.03**	**84.98**	**88.17**	**85.31**	**86.16**	**85.26**	**88.07**	**86.53**	**86.04**	**85.99**
LLaMA3-8B										
LLaMA3-8B	86.21	69.38	88.83	**85.62**	**86.89**	85.22	**88.96**	86.45	86.90	**87.47**
MindMerger	87.52	-	88.29	-	85.89	-	87.86	-	85.68	-
LayerAlign	87.35	-	87.88	-	85.45	-	87.10	-	85.53	-
XBridge (Ours)	**88.53**	**84.77**	**88.97**	85.57	86.66	**85.23**	88.92	**86.53**	**87.01**	85.33
Aya-23-8B										
Aya-23-8B	68.76	40.01	88.40	**85.27**	86.42	**85.26**	**88.37**	**86.85**	**86.97**	**89.75**
MindMerger	87.89	-	88.32	-	86.28	-	88.08	-	86.11	-
LayerAlign	87.66	-	88.28	-	86.05	-	87.84	-	85.70	-
XBridge (Ours)	**88.12**	**84.40**	**88.80**	84.45	**86.48**	84.86	88.33	86.22	86.78	85.45
Qwen2.5-7B-Instruct										
Qwen2.5-7B-Instruct	83.02	61.32	88.13	84.31	86.28	**85.35**	88.21	**86.42**	86.06	**88.76**
MindMerger	88.07	-	88.75	-	86.52	-	88.31	-	85.94	-
LayerAlign	88.07	-	88.52	-	86.32	-	88.01	-	86.65	-
XBridge (Ours)	**88.25**	**84.65**	**88.80**	**84.36**	**86.58**	84.70	**88.52**	85.87	**86.48**	84.87

System	Ru-En	En-Ru	Sw-En	En-Sw	Th-En	En-Th	Zh-En	En-Zh	X-En	En-X
NLLB-200-1.3B	86.20	88.02	85.11	83.07	86.39	80.17	85.66	76.98	87.01	84.36
MetaMath-7B										
MetaMath-7B	83.41	72.45	49.89	43.48	58.68	39.53	82.32	45.69	73.92	60.68
MindMerger	85.06	-	84.88	-	85.14	-	83.96	-	85.76	-
LayerAlign	85.51	-	**84.95**	-	85.30	-	**84.42**	-	85.93	-
XBridge (Ours)	**85.38**	**86.65**	84.87	**80.23**	**86.07**	**81.38**	82.00	**83.52**	**86.09**	**84.43**
LLaMA3-8B										
LLaMA3-8B	86.19	**87.17**	82.74	74.06	**87.04**	**83.54**	**86.33**	**85.93**	86.68	82.76
MindMerger	85.33	-	85.30	-	85.83	-	84.48	-	86.24	-
LayerAlign	85.24	-	85.25	-	85.74	-	84.13	-	85.96	-
XBridge (Ours)	**86.23**	86.27	**85.63**	**80.19**	86.79	81.12	84.64	83.46	**87.04**	**84.27**
Aya-23-8B										
Aya-23-8B	85.87	**88.02**	54.36	31.49	76.40	48.21	**86.14**	**86.88**	80.19	71.31
MindMerger	85.80	-	**85.43**	-	86.07	-	84.53	-	86.50	-
LayerAlign	85.74	-	85.32	-	85.78	-	84.55	-	86.32	-
XBridge (Ours)	**85.88**	85.87	85.36	**80.03**	**86.39**	**81.03**	83.23	82.80	**86.60**	**83.90**
Qwen2.5-7B-Instruct										
Qwen2.5-7B-Instruct	84.22	84.68	64.52	46.92	84.94	**84.00**	**86.86**	**88.30**	83.58	78.90
MindMerger	85.95	-	**85.82**	-	86.49	-	85.62	-	**86.83**	-
LayerAlign	85.82	-	85.66	-	86.42	-	85.61	-	86.79	-
XBridge (Ours)	**85.91**	**85.80**	85.43	**80.23**	**86.45**	80.51	84.99	82.79	86.82	**83.75**

Table 5: Detailed COMET scores on FLORES-101 for stage 1. "X" denotes all languages except for English. We bold the best scores for each LLM group.

System	Bn	De	En	Es	Fr	Ja	Ru	Sw	Th	Zh	Avg
					MetaMath-7B						
SFT	30.00	40.40	48.00	39.60	39.20	32.40	39.60	35.20	32.80	36.00	37.32
Translate-Test	34.80	49.20	53.20	47.60	40.80	45.20	49.20	43.20	32.40	46.00	44.16
MindMerger	43.60	**57.20**	**67.20**	59.20	**60.00**	45.60	**57.60**	49.20	42.40	48.80	53.08
LayerAlign	40.40	54.80	62.80	59.20	52.40	**51.60**	56.40	**53.20**	**47.20**	51.20	52.92
XBridge_LLM (Ours)	**48.40**	54.80	**67.20**	**60.80**	57.20	47.20	**57.60**	46.40	45.60	**52.80**	**53.80**
XBridge_Dec (Ours)	47.20	52.80	66.80	60.00	51.60	44.80	56.00	45.60	44.40	52.00	52.12
					LLaMA3-8B						
SFT	38.80	46.40	54.00	54.80	44.40	38.80	45.20	40.80	44.00	49.20	45.64
Translate-Test	40.40	56.00	62.00	54.80	54.80	**51.60**	57.20	46.80	35.20	48.00	50.68
MindMerger	**49.60**	54.80	60.00	56.80	55.60	50.00	55.20	48.00	46.40	50.80	52.72
LayerAlign	50.00	56.00	62.40	60.40	57.60	50.00	56.80	47.60	48.80	52.80	54.24
XBridge_LLM (Ours)	47.20	**59.20**	**65.20**	**62.00**	**62.00**	50.80	**60.40**	**49.60**	**54.40**	**56.40**	**56.72**
XBridge_Dec (Ours)	46.00	56.80	62.80	61.20	57.20	46.40	58.40	**49.60**	51.20	55.20	54.48
					Aya-23-8B						
SFT	33.60	**52.80**	50.80	47.60	43.60	46.40	46.40	38.80	**39.20**	42.00	44.12
Translate-Test	34.00	50.00	**57.20**	48.40	45.60	42.80	**50.40**	43.60	33.20	**50.80**	45.60
MindMerger	40.80	49.60	48.00	48.00	46.00	42.40	48.00	42.40	33.60	40.40	43.92
LayerAlign	41.60	45.20	49.20	48.80	48.40	44.00	46.80	41.20	37.60	46.00	44.88
XBridge_LLM (Ours)	44.40	49.20	56.80	**53.20**	**50.00**	**47.20**	49.20	**44.80**	37.60	46.00	**47.84**
XBridge_Dec (Ours)	**44.80**	46.40	56.80	**53.20**	46.40	46.40	49.20	**44.80**	34.00	44.00	46.60
					Qwen2.5-7B-Instruct						
SFT	49.60	64.00	66.40	59.60	58.40	54.40	60.00	48.40	62.80	67.60	59.12
Translate-Test	43.20	67.60	74.40	62.40	60.40	59.60	63.20	58.40	44.00	62.80	59.60
MindMerger	60.40	71.60	83.60	76.40	73.60	64.80	**77.20**	65.20	66.40	70.00	70.92
LayerAlign	**65.60**	**75.20**	83.20	**78.40**	**74.00**	62.80	74.40	66.80	65.60	68.80	71.48
XBridge_LLM (Ours)	63.20	74.40	**83.60**	77.20	72.80	**65.20**	74.00	**67.60**	**68.40**	**71.60**	**71.80**
XBridge_Dec (Ours)	60.80	72.40	83.20	76.00	65.60	59.20	72.40	66.80	64.00	70.40	69.08

Table 6: Detailed accuracy results on MGSM. "XBridge-LLM" denotes English reasoning outputs from the LLM, while "XBridge-Dec" denotes multilingual reasoning outputs via the external decoder. We bold the best scores for each LLM group.

System	Bn	En	Es	Fr	Ja	Ru	Sw	Th	Zh	Avg
				MetaMath-7B						
SFT	16.79	**33.27**	**23.68**	**28.42**	29.20	**25.42**	27.01	18.68	20.16	24.74
MindMerger	17.93	27.27	21.03	26.05	28.11	22.25	26.49	19.20	18.56	22.99
LayerAlign	17.94	28.84	21.49	26.31	27.55	22.22	26.29	20.09	19.02	23.30
XBridge (Ours)	**18.93**	31.66	22.75	27.96	**32.12**	24.64	**28.90**	**20.70**	**24.12**	**25.75**
				LLaMA3-8B						
SFT	**20.87**	33.59	23.84	**30.18**	30.42	26.12	28.73	22.51	21.72	26.44
MindMerger	18.76	31.29	22.88	27.84	29.02	23.97	27.63	20.47	19.75	24.62
LayerAlign	18.78	30.78	22.98	27.47	28.55	23.59	27.58	20.48	19.56	24.42
XBridge (Ours)	20.65	**33.94**	**24.87**	29.62	**33.74**	**26.42**	**30.75**	**23.19**	**25.91**	**27.68**
				Aya-23-8B						
SFT	17.68	**34.97**	**24.88**	**29.78**	31.43	**26.96**	26.11	20.14	21.70	25.96
MindMerger	18.49	31.87	22.54	27.40	28.21	23.49	27.36	20.23	19.45	24.34
LayerAlign	17.93	31.56	22.23	26.83	28.17	23.07	27.25	20.35	19.23	24.07
XBridge (Ours)	**19.03**	33.10	23.70	27.59	**32.82**	25.48	**28.78**	**20.96**	**25.23**	**26.30**
				Qwen2.5-7B-Instruct						
SFT	**19.52**	**33.52**	**24.79**	**29.14**	30.81	**26.20**	26.13	22.16	21.41	25.97
MindMerger	18.17	30.38	22.06	25.95	28.33	22.92	26.40	20.04	18.85	23.68
LayerAlign	18.27	30.51	22.35	26.56	28.90	22.82	26.32	20.18	19.34	23.92
XBridge (Ours)	19.22	31.52	23.48	27.72	**32.77**	24.53	**27.95**	**22.19**	**25.02**	**26.04**

Table 7: Detailed multilingual Rouge-L results on XL-Sum. For XL-Sum, since the baselines produce English summaries only, we translate them into target languages using NLLB-200-1.3B for comparison. We bold the best scores for each LLM group.

Variants	Lrl-En	En-Lrl	Hrl-En	En-Hrl	X-En	En-X
MetaMath-7B	2.87	1.05	26.50	17.36	18.62	11.92
w/o OT	35.33	22.20	32.98	25.30	33.76	24.27
w/o Decoder	**36.22**	-	**35.01**	-	**35.41**	-
XBridge	33.93	**25.56**	31.98	**29.37**	32.63	**28.10**

Table 8: Ablation results on FLORES-101 for stage 1. "Lrl" and "Hrl" denote low- and high-resource languages, respectively. Following Shi et al. (2023), we treat Bn, Sw, and Th as low-resource languages, and the remaining as high-resource ones. "X" denotes all languages except for English. We bold the best scores for each LLM group.

Variants	Lrl-LLM	Lrl-Dec	Hrl-LLM	Hrl-Dec	Avg-LLM	Avg-Dec
MetaMath-7B	5.60	-	50.06	-	36.72	-
SFT	32.67	-	39.31	-	37.32	-
w/o OT	48.27	47.20	56.23	54.57	53.84	52.36
w/o Decoder	48.27	-	56.91	-	54.32	-
w/o Stage 1	29.33	28.13	52.51	51.37	45.56	44.40
Joint Stage 2&3	43.07	42.53	48.63	47.26	46.96	45.84
XBridge	**48.67**	**47.87**	**57.49**	**55.09**	**54.84**	**52.92**

Table 9: Ablation accuracy results on MGSM. "Lrl" and "Hrl" denote low- and high-resource languages, respectively. "-LLM" denotes English reasoning outputs from the LLM, while "-Dec" denotes multilingual reasoning outputs via the external decoder. We bold the best scores.

Variants	Bn	En	Es	Fr	Ja	Ru	Sw	Th	Zh	Lrl	Hrl	Avg
SFT	16.79	**33.27**	**23.68**	**28.42**	29.20	**25.42**	27.01	18.68	20.16	20.83	26.69	24.74
w/o OT	18.55	31.49	22.68	27.86	32.05	24.27	28.35	20.27	23.81	22.39	27.03	25.48
XBridge	**18.85**	31.50	22.84	28.18	**32.26**	24.36	**28.57**	**21.08**	**24.40**	**22.83**	**27.26**	**25.78**

Table 10: Ablation Rouge-L results of XL-Sum on MetaMath-7B. "Lrl" and "Hrl" denote low- and high-resource languages, respectively. We bold the best scores.

Languages	X→En				En→X			
	NLLB	MetaMath	w/o OT	XBridge	NLLB	MetaMath	w/o OT	XBridge
Af	58.33	17.64	53.69	**54.90**	42.69	12.79	37.30	**41.75**
Ar	42.42	24.59	**40.61**	38.20	33.13	10.94	16.81	**26.37**
Az	28.25	**28.97**	25.35	23.57	19.39	**16.23**	12.25	16.18
Cs	41.17	1.34	**40.47**	38.54	35.85	0.32	23.39	**30.03**
El	37.68	3.59	**37.19**	35.43	28.16	0.80	19.07	**24.34**
Et	38.22	29.84	**33.99**	33.30	29.77	9.62	19.61	**26.28**
Fa	36.91	8.67	**35.10**	33.06	26.86	2.27	17.28	**22.33**
Fi	45.36	31.28	**42.34**	41.34	38.09	18.29	30.98	**34.92**
Gl	43.18	2.67	**40.69**	39.12	37.72	0.78	25.71	**33.11**
Gu	44.32	28.17	**43.02**	40.53	36.21	18.51	19.81	**28.92**
He	42.93	0.21	**40.75**	39.82	38.07	0.12	26.84	**33.31**
Hi	39.18	29.48	**39.40**	37.57	31.65	13.01	20.71	**26.83**
Hr	45.60	24.70	**44.42**	42.83	46.27	12.54	35.95	**40.81**
Id	38.74	7.49	**34.99**	33.54	34.57	2.30	24.36	**28.92**
It	30.15	3.87	**28.25**	26.93	27.28	1.12	18.13	**23.72**
Ka	34.89	5.43	**31.69**	30.60	27.81	1.23	18.33	**24.24**
Kk	33.11	5.87	**29.29**	27.58	15.01	1.92	8.74	**13.13**
Km	30.88	0.08	**27.63**	26.25	24.84	0.16	15.51	**20.58**
Ko	22.49	**36.64**	15.39	6.83	0.18	**13.52**	0.70	0.51
Lt	34.38	2.85	**33.89**	32.04	28.25	1.59	19.97	**24.62**
Lv	39.26	0.23	**36.20**	34.78	29.87	0.09	18.89	**24.50**
Mk	37.98	14.19	**35.15**	33.72	25.91	4.29	17.13	**21.20**
Ml	44.49	2.05	**42.90**	41.80	37.80	0.45	27.26	**32.83**
Mn	26.73	**29.23**	25.88	24.91	16.95	**18.22**	9.37	14.18
Mr	29.24	25.51	26.04	**26.38**	18.53	5.84	13.92	**18.26**
My	36.07	**37.48**	34.04	32.39	31.54	**30.64**	22.08	27.72
Ne	42.52	3.61	**38.93**	37.98	28.37	1.22	20.28	**25.64**
Nl	33.18	**32.03**	31.83	30.20	25.95	16.41	16.94	**20.80**
Pl	52.13	0.26	**49.96**	48.55	49.84	0.26	38.40	**44.17**
Ps	34.68	10.20	**32.04**	30.48	19.93	2.18	9.93	**16.49**
Pt	47.08	25.82	**43.58**	42.52	37.89	16.67	28.34	**34.26**
Ro	36.87	**41.00**	36.54	33.82	30.79	**29.45**	20.87	26.02
Sl	35.35	27.14	**30.83**	29.90	31.36	20.21	24.99	**27.54**
Sv	49.10	40.61	**49.21**	46.40	43.34	26.96	32.16	**38.06**
Ta	35.56	12.89	**32.42**	31.67	31.99	2.05	20.41	**27.55**
Te	41.87	10.57	**37.98**	37.60	38.21	3.45	24.07	**29.98**
Tr	40.87	18.17	**37.78**	35.94	34.67	7.62	20.11	**28.19**
Uk	41.39	36.98	**39.21**	33.53	32.30	22.16	22.45	**27.34**
Ur	37.23	0.56	**35.27**	33.65	27.06	0.27	15.98	**24.49**
Vi	38.84	25.91	**36.37**	34.11	39.95	11.68	29.97	**35.84**
Xh	38.49	20.87	**37.20**	35.22	27.76	9.99	20.24	**26.42**
Avg	38.71	17.29	**36.28**	34.57	30.78	8.98	21.10	**26.64**

Table 11: Detailed results on 41 untuned languages, composing MetaMath-7B with NLLB-200-1.3B. We evaluate cross-model mapping quality on FLORES-101, following the main experiments. "X" denotes all languages except for English. We bold the best scores for the LLM group.

	Bn	De	En	Es	Fr	Ja	Ru	Sw	Th	Zh	Avg
Bn	-	22.47	37.09	20.56	31.36	16.69	21.03	22.85	13.62	17.42	22.57
De	24.74	-	45.75	25.57	39.82	18.73	26.51	26.34	14.82	19.81	26.90
En	28.42	35.45	-	29.59	49.38	20.12	30.57	34.68	17.09	23.11	29.82
Es	22.21	23.42	32.00	-	33.74	16.82	21.78	22.01	12.93	17.57	22.50
Fr	24.26	28.62	46.10	26.64	-	17.75	25.90	26.99	14.47	19.96	25.63
Ja	20.76	20.05	27.63	18.59	26.94	-	18.82	19.76	12.60	17.16	20.26
Ru	22.89	25.65	37.08	23.41	34.80	17.85	-	23.96	13.91	18.82	24.26
Sw	23.16	25.36	44.73	22.37	36.35	17.44	23.27	-	14.24	18.18	25.01
Th	21.59	20.98	30.61	19.34	29.39	16.37	19.78	21.86	-	17.12	21.89
Zh	20.36	19.78	24.89	18.90	27.12	15.89	18.74	20.21	12.94	-	19.87
Avg	23.15	24.64	36.21	22.77	34.32	17.52	22.93	24.30	14.07	18.79	23.87

Table 12: BLEU scores for cross-lingual generation of XBridge, composing LLaMA3-8B with NLLB-200-1.3B. Rows denote the source language and columns denote the target language. The source text is encoded by the multilingual encoder, the LLM produces an English translation, and the decoder generates the target-language text conditioned on the target-language token. For *X→En* directions, we directly evaluate the LLM outputs.

MGSM	**Bn**	**De**	**En**	**Es**	**Fr**	**Ja**	**Ru**	**Sw**	**Th**	**Zh**	**Avg**
Bn	46.00	44.40	45.20	45.60	40.80	42.80	44.00	46.00	44.00	44.80	44.36
De	58.40	56.40	58.40	59.20	53.20	54.80	58.40	58.00	58.40	59.20	57.44
En	63.60	62.80	63.20	63.60	59.60	61.60	62.80	63.20	61.20	64.00	62.56
Es	60.00	57.60	61.60	61.20	55.20	56.80	60.40	60.80	57.60	60.80	59.20
Fr	58.80	58.40	60.40	60.80	57.20	56.80	60.80	59.60	58.00	60.00	59.08
Ja	49.20	47.20	49.20	50.00	43.60	46.80	48.80	48.80	48.00	48.80	48.04
Ru	57.60	58.40	59.60	60.00	55.20	57.60	58.40	58.40	56.00	58.00	57.92
Sw	49.20	48.00	50.00	50.40	45.60	48.00	49.20	49.60	48.40	48.80	48.72
Th	50.40	49.60	52.80	50.80	49.20	49.20	50.00	52.40	50.00	52.40	50.68
Zh	55.20	53.60	55.20	55.60	50.80	51.20	54.40	55.20	52.80	55.20	53.92
Avg	54.84	53.64	55.56	55.72	51.04	52.56	54.72	55.20	53.44	55.20	54.19

Table 13: Accuracy for langauge-on-demand generation of XBridge, composing LLaMA3-8B with NLLB-200-1.3B. Rows denote the query language and columns denote the response language. The query is first processed by the encoder, the LLM produces an English response, and the decoder then generates the target language response via the target language token.

System	**Lrl-En**	**En-Lrl**	**Hrl-En**	**En-Hrl**	**X-En**	**En-X**
M2M100-1.2B	28.00	22.25	33.11	32.55	31.41	29.12
LLaMA3-8B	32.03	17.43	36.77	32.32	35.19	27.36
MindMerger	31.37	-	32.67	-	32.24	-
LayerAlign	31.28	-	32.47	-	32.08	-
XBridge (Ours)	**32.24**	**21.66**	**35.86**	**28.67**	**34.65**	**26.33**

Table 14: Detailed FLORES-101 translation results for LLaMA3-8B composed with M2M100-1.2B. "Lrl" and "Hrl" denote low- and high-resource languages, respectively. "X" denotes all languages except for English. We bold the best scores for the LLM group.

System	**Bn**	**De**	**En**	**Es**	**Fr**	**Ja**	**Ru**	**Sw**	**Th**	**Zh**	**Avg**
SFT	38.80	46.40	54.00	54.80	44.40	38.80	45.20	40.80	44.00	49.20	45.64
Translate-Test	40.40	56.00	62.00	54.80	54.80	51.60	57.20	46.80	35.20	48.00	50.68
MindMerger	**52.40**	55.60	62.40	55.60	**59.60**	50.40	58.40	47.60	50.80	56.40	54.92
LayerAlign	51.20	**60.00**	65.20	58.00	55.60	50.00	**59.20**	**49.20**	50.80	54.00	55.32
XBridge_LLM (Ours)	49.20	57.20	**66.00**	**60.80**	52.80	**54.40**	54.80	47.20	**52.80**	**60.00**	**55.52**
XBridge_Dec (Ours)	48.40	53.20	62.80	59.20	48.40	53.20	54.00	44.40	48.40	58.80	53.08

Table 15: Detailed MGSM results for LLaMA3-8B composed with M2M100-1.2B. "XBridge-LLM" denotes English reasoning outputs from the LLM, while "XBridge-Dec" denotes multilingual reasoning outputs via the external decoder. We bold the best scores for the LLM group.

System	Lrl-En	En-Lrl	Hrl-En	En-Hrl	X-En	En-X
NLLB-200-600M	33.59	26.43	34.83	29.52	34.42	28.49
NLLB-200-1.3B	37.09	28.86	37.72	32.08	37.51	31.00
MetaMath-7B	2.87	1.05	26.50	17.36	18.62	11.92
XBridge (600M)	31.20	23.42	29.46	26.55	30.04	25.50
XBridge (1.3B)	**35.26**	**26.91**	**32.42**	**31.24**	**33.37**	**29.80**

Table 16: Detailed FLORES-101 translation results for MetaMath-7B composed with NLLB-200 in different sizes (600M vs. 1.3B). "Lrl" and "Hrl" denote low- and high-resource languages, respectively. "X" denotes all languages except for English. We bold the best scores for the LLM group.

System	Bn	De	En	Es	Fr	Ja	Ru	Sw	Th	Zh	Avg
SFT	38.80	46.40	54.00	54.80	44.40	38.80	45.20	40.80	44.00	49.20	45.64
XBridge_LLM (600M)	47.60	**56.00**	65.60	56.00	56.00	**47.20**	56.00	**50.80**	**49.20**	50.00	53.44
XBridge_Dec (600M)	45.20	52.80	65.60	55.60	51.60	44.40	54.40	**50.80**	46.00	49.20	51.56
XBridge_LLM (1.3B)	**48.40**	54.80	**67.20**	**60.80**	**57.20**	**47.20**	**57.60**	46.40	45.60	**52.80**	**53.80**
XBridge_Dec (1.3B)	47.20	52.80	66.80	60.00	51.60	44.80	56.00	45.60	44.40	52.00	52.12

Table 17: Detailed FLORES-101 translation results for MetaMath-7B composed with NLLB-200 in different sizes (600M vs. 1.3B). "X" denotes all languages except for English. We bold the best scores for the LLM group.