One Pair Suffices: Unlocking Universal Zero-Shot Translation via Cross-Architecture Alignment

Hao Zong[1,2] , Conghu Yuan[3] ,Chao Bei[3] , Conghu Yuan[3] , Wentao Chen[3], Huan Liu[2] , Kaiyu Huang[4*] and Degen Huang[1,5*]

[1]School of Computer Science and Technology, Dalian University of Technology,
[2]DeepFutureAI,
[3]Global Tone Communication Technology Co., Ltd,
[4]School of Computer Science and Technology of Beijing Jiaotong University,
[5]School of Computing and Artificial Intelligence, Fuyao University of Science and Technology

peterz.nlp@gmail.com

Abstract

Current paradigms for empowering Large Language Models (LLMs) with multilingual capabilities rely heavily on massive instruction tuning. We challenge this view, proposing that the barrier is topological alignment, not data quantity. We introduce **Hybrid Cross-Alignment (HCA)**, fusing a frozen NLLB encoder with a Qwen decoder via a closed-loop dual-adapter architecture. HCA utilizes a Source-Side Adapter to precondition encoder features and a Query-Residual Adapter to preserve generative stability, bridged by an adaptive gated cross-modal interface. Our core finding is the phenomenon of "Source-Side Alignment Generalization." We demonstrate that training HCA on a single language pair (German-English) unlocks state-of-the-art zero-shot transfer to dozens of unseen languages for **X-to-English translation**. Crucially, our "Oracle" experiments reveal that this single-pair training recovers over 96.7% of the performance achievable by training on all available pairs. This suggests that a highly generalizable, source-side projection protocol exists. Evaluated rigorously across **COMET and chrF++**, our ∼5.25B-parameter model significantly outperforms larger baselines, surpassing TowerPlus-9B (+9.0 COMET on low-resource languages) and Aya-101 (13B). Furthermore, performance scales linearly with encoder size; upgrading from 600M to 1.3B yields immediate gains (+3.4 points on Gujarati) with minimal retraining cost.

1 Introduction

The landscape of Machine Translation (MT) is currently bifurcated. On one side, Large Language Models (LLMs) (Yang et al., 2025) demonstrate exceptional generative fluency but suffer from the "curse of multilinguality"—performance degrades sharply on low-resource languages absent from their pre-training data. On the other side, specialized dense encoders like NLLB (Costa-jussà et al., 2022) achieve massive multilingual coverage through highly aligned, isometric embedding spaces, yet they lack the versatile generative power of decoder-only transformers.

Current efforts to bridge this gap typically rely on *Continuous Pre-training* (CPT) or massive *Supervised Fine-Tuning* (SFT) (Xu et al., 2025; Alves et al., 2024). While effective, these strategies are computationally exorbitant and, we argue, **conceptually redundant**: they force the LLM to statistically re-learn multilingual alignments that specialized encoders have already mastered. This raises a fundamental scientific question: *Is the semantic topology of a multilingual encoder universal enough that an LLM can learn to "read" all languages by simply learning to read one?*

In this paper, **focusing on the scope of X-to-English generation**, we answer with a definitive "yes." We propose **Hybrid Cross-Alignment (HCA)**, a parameter-efficient framework that fuses a frozen NLLB encoder with a frozen Qwen decoder. Unlike standard adapters, HCA implements a **closed-loop alignment system**: (1) A **Source-Side Alignment Adapter** topologically preconditions encoder features; (2) A **Cross-Modal Alignment Adapter** projects semantics via gated attention, enabling adaptive information fusion; (3) A **Query-Residual Adapter** preserves the LLM's generative stability.

Our experiments reveal a phenomenon we term **"Source-Side Alignment Generalization,"** supported by three compelling findings:

- **One Pair Suffices for X-to-English:** Training HCA *solely* on German-English (De-En) unlocks state-of-the-art zero-shot translation for dozens of linguistically distinct languages into English. Our ∼5.25B-parameter model significantly outperforms larger supervised baselines on low-resource languages, beating

* Corresponding Author

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 41227–41237
July 2-7, 2026

TowerPlus-9B (+9.0 COMET) and surpassing Aya-101 (13B) (82.6 vs. 80.7). Importantly, training on Arabic-English also yields robust generalization, confirming the broad applicability of the projection protocol across distinct language families.

- **The Oracle Gap is Minimal:** Comparing our Zero-Shot model against an "Oracle" model trained on *all* available language pairs, we find that single-pair training recovers over 96.7% of the upper-bound performance (82.9 vs. 85.8 COMET). This implies that the cross-architecture projection protocol is largely language-agnostic for source-side representations, providing strong empirical backing for the isomorphic alignment hypothesis.
- **Synergistic Superiority:** Crucially, our Oracle model achieves an overall average score of 85.8, surpassing the fully supervised NLLB-1.3B teacher (85.6). This proves that our framework effectively synergizes the superior representation of the dense encoder with the generative fluency of the LLM, yielding a system that exceeds the capabilities of its constituent parts.

By simply upgrading the frozen encoder from 600M to 1.3B, we observe immediate gains (+3.4 COMET on Gujarati) without extra training. It is important to note that our method does not create multilingual knowledge *ex nihilo*; rather, it effectively **unlocks** the massive multilingual supervision already baked into the frozen NLLB encoder. The novelty lies in the topological bridge that transfers this capability to the generative LLM without catastrophic forgetting.

2 Related Work

LLMs for Machine Translation. While foundational models like GPT-4 (Josh Achiam and others., 2024) and LLaMA-3 (Dubey et al., 2024) exhibit fluency, they struggle with low-resource languages due to sparse coverage and vocabulary bottlenecks (Zheng et al., 2021). To address this, specialized efforts—including X-ALMA (Xu et al., 2025), HunyuanMT (Zheng et al., 2025), MarcoMT (Wang et al., 2025), and TowerLLM (Alves et al., 2024)—enhance capability via massive Continuous Pre-training (CPT). However, these methods rely on a **data-scaling paradigm**, requiring full-weight updates on billions of tokens. This incurs prohibitive costs and hits a ceiling for long-tail languages where data is scarce. Our work differs by shifting from data scaling to **topological alignment**, requiring zero target data.

Multilingual Encoders and Transfer. Discriminative encoders like XLM-R (Liu et al., 2019), LaBSE (Feng et al., 2022), and NLLB (Costa-jussà et al., 2022) established shared semantic spaces. Recent works like SeamlessM4T (Barrault et al., 2023) and SONAR (Duquenne et al., 2023) extended this to multimodal embeddings. Pioneering the zero-shot transfer direction, SixT (Chen et al., 2021) demonstrated that fine-tuning an encoder with a randomly initialized decoder on a single pair could enable multilingual translation. However, SixT relies on training a lightweight decoder from scratch, missing out on the reasoning and fluency of modern LLMs. Our HCA framework evolves this paradigm by replacing the weak decoder with a powerful, frozen LLM (Qwen3), significantly boosting generation quality.

Modular Composition and Cross-Modal Fusion. Drawing from foundational work on parameter-efficient transfer learning using adapters (Houlsby et al., 2019) and modular learning (Zhang et al., 2023), recent works explore bridging external encoders with LLMs. In this context, "cross-modal fusion" broadly refers to integrating information across distinct latent spaces or modalities (e.g., injecting vision encoders as seen in BLIP-2 (Li et al., 2023)). Within the textual domain, recent approaches like MT-LLM (Schmidt et al., 2024) align multilingual encoders with LLMs using standard LoRA and simple input-level projection, primarily focusing on cross-lingual Natural Language Understanding (NLU) tasks via self-distillation. Other works explore model merging (Wan et al., 2025) or composition (Bansal et al., 2024). Most notably, LaMaTE (Luo et al., 2025) explores the inverse approach: using LLMs as encoders for NMT decoders.

In contrast to input-level projection or shallow adaptation methods, HCA implements a **deep fusion architecture** tailored for the strict semantic fidelity required in translation. Instead of merely projecting tokens into the input layer, we inject preconditioned NLLB representations into *every layer* of the Qwen decoder via gated cross-attention. This closed-loop, layer-wise integration ensures precise semantic control and generative stability, distinguishing HCA from prior shallow fusion or model merging approaches.

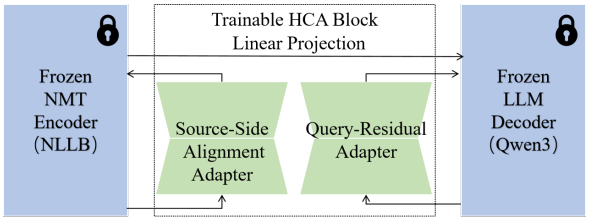

Figure 1: The schematic overview of our proposed framework. The trainable HCA Block acts as a bridge between the frozen NLLB Encoder and Qwen Decoder. It strictly decouples representation learning (Source-Side) from generation (Query-Residual) via a cross-modal interface.

3 Methodology

3.1 Overall Framework

We propose a parameter-efficient framework to bridge the gap between discriminative multilingual representations and generative Large Language Models (LLMs). As illustrated in Figure 1, our architecture freezes two pre-trained giants: the NLLB encoder ($\mathcal{E}$) on the left and the Qwen LLM decoder ($\mathcal{D}$) on the right.

To facilitate communication between these distinct architectures without disrupting their original capabilities, we introduce the Hybrid Cross-Alignment (HCA) Block. As detailed in the microscopic view in Figure 2, the data flow within this block forms a closed-loop system organized into three distinct functional zones:

I. **Source-Side Alignment (Green Dashed Box):** Pre-conditioning the NLLB features via an intra-encoder adapter.

II. **Cross-Modal Alignment (Blue Dashed Box):** Projecting and fusing the aligned features into the LLM via gated cross-attention.

III. **Query-Residual Adaptation (Red Dashed Box):** Refining the LLM's query representations to maintain generative stability.

In the following subsections, we detail the implementation of these components corresponding to the dashed regions in Figure 2.

3.2 Source-Side Alignment

Standard cross-attention mechanisms often fail when the source feature distribution diverges significantly from the target decoder's expected input manifold. As encapsulated by the **green dashed box** in the left panel of Figure 2, we intervene directly within the NLLB encoder layers.

Let $\mathbf{H}^l_{attn}$ denote the output of the self-attention layer in the encoder. Instead of feeding this directly to the projection layer, we employ a **Source-Side Alignment Adapter**. This adapter operates in parallel to the main feed-forward path to refine the features (as shown within the green dashed box, where the adapter's output is added via $\bigoplus$ before 'Final Norm'):

$$\mathbf{H}_{align} = \text{ReLU}(\mathbf{H}^l_{attn}\mathbf{W}^{src}_{down})\mathbf{W}^{src}_{up} \quad (1)$$

The refined source representation $\mathbf{H}'_{enc}$ (denoted as $\mathbf{H}$ in the figure) is obtained by summing the original self-attention output, FFN output, and the adapter's output before the final layer normalization:

$$\mathbf{H}'_{enc} = \text{FinalNorm}(\text{FFN}(\mathbf{H}^l_{attn}) + \mathbf{H}_{align}) \quad (2)$$

This step effectively "pre-conditions" the dense multilingual vectors, transforming them into a topological space compatible with the subsequent projection.

3.3 Cross-Modal Alignment Adapter

As highlighted by the blue dashed box in Figure 2, we design a specific Cross-Modal Alignment Adapter to bridge the topological gap between the NLLB encoder and the Qwen decoder. This module aligns the frozen encoder's hidden states with the LLM's generative space through linear projection and attention operations.

I. Linear Projection: First, the refined source feature $\mathbf{H}'_{enc}$ (the output labeled $\mathbf{H}$) serves as the input. To match the dimension of the LLM, it passes through a Projection layer:

$$\mathbf{K}, \mathbf{V} = \mathbf{H}'_{enc}\mathbf{W}_{proj} \quad (3)$$

where $\mathbf{W}_{proj} \in \mathbb{R}^{d_{enc} \times d_{llm}}$. The output acts as both Keys ($\mathbf{K}$) and Values ($\mathbf{V}$) for the subsequent attention mechanism.

II. Gated Cross-Attention: Inside the adapter, we employ a standard dot-product attention mechanism to synthesize source information.

- **Correlation Calculation (Bottom Matmul):** The LLM's hidden state serves as the Query ($\mathbf{Q}$). It interacts with the projected Keys $\mathbf{K}$ to compute attention scores via the first Matrix Multiplication (Matmul) and Softmax:

$$\mathbf{A} = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right) \quad (4)$$

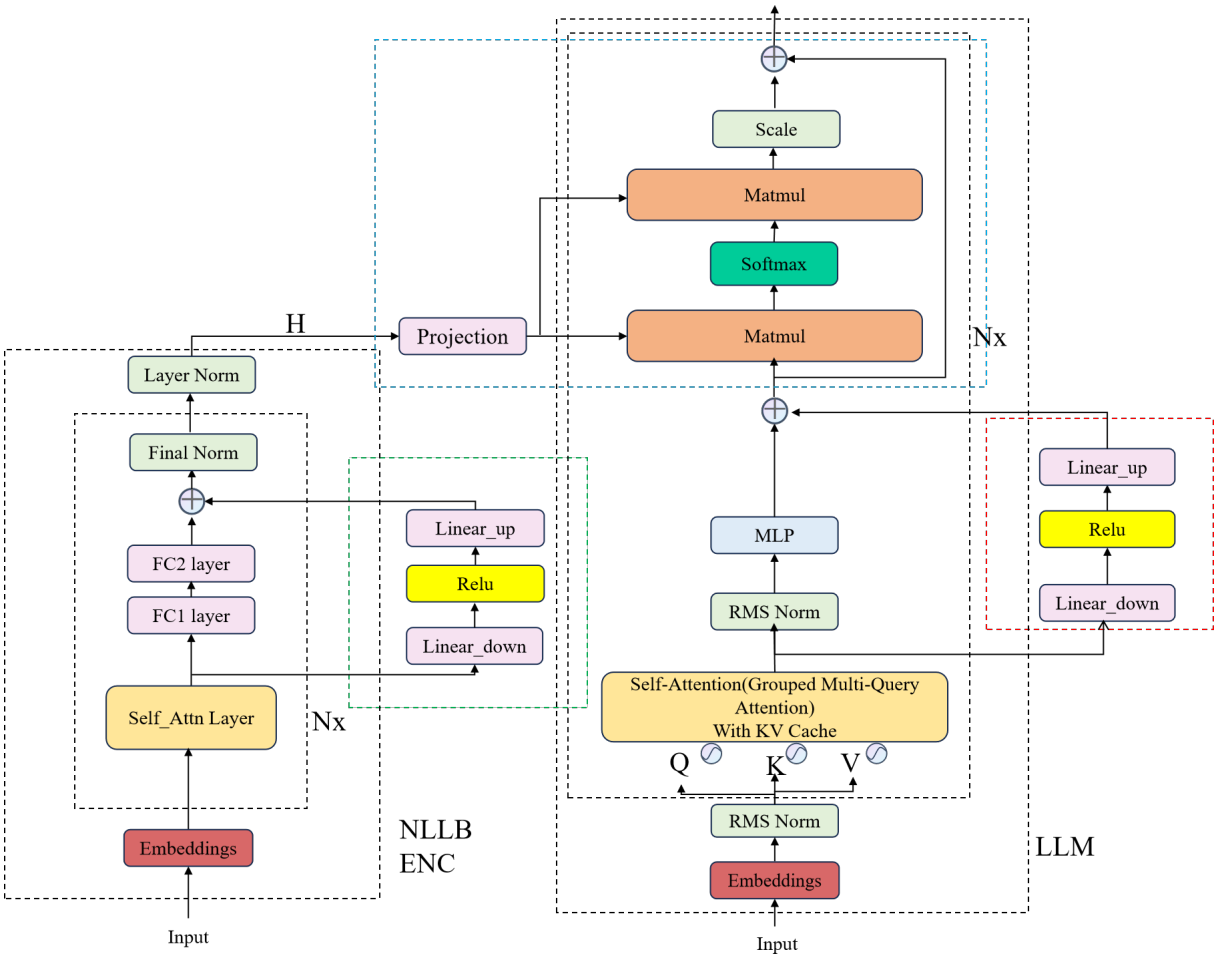

Figure 2: Detailed architecture of the Hybrid Cross-Alignment (HCA) Block. **Green Dashed Box (Left):** The Source-Side Alignment Adapter refines NLLB features before they leave the encoder. **Blue Dashed Box (Center):** The Cross-Modal Alignment Adapter projects these features and injects them into the LLM via Gated Cross-Attention. **Red Dashed Box (Right):** A Query-Residual Adapter operates in parallel to the cross-attention to preserve the LLM's internal linguistic knowledge. The "Scale" factor employs a learnable initialization strategy. The input to the NLLB is the source language, and the input to the LLM is a prompt: translate this sentence to English. The loss is the cross-entropy between the target language and the LLM output.

- **Content Aggregation (Top Matmul):** The attention scores $\mathbf{A}$ are then used to weight the Values $\mathbf{V}$ via the second Matmul, retrieving the relevant semantic context from the NLLB encoder.

III. Gated Scaling. To adaptively control the fusion ratio, the aggregated context passes through a **Scale** layer. This corresponds to a learnable gating factor g:

$$\mathbf{C}_{cross} = g \cdot (\mathbf{AV}) \tag{5}$$

We initialize $\boldsymbol{g = 1}$. This sets the cross-attention mechanism to be fully active at the start of training, encouraging the LLM to immediately attend to the aligned encoder representations.

3.4 Query-Residual Adaptation

Directly injecting cross-modal information can sometimes disrupt the internal reasoning flow of the LLM. To address this, we introduce the Query-Residual Adapter, demarcated by the red dashed box on the right side of Figure 2.

This adapter processes the LLM's query features $\mathbf{H}_{llm}$ independently of the source information. It creates a dedicated pathway to preserve and refine the target-side linguistic features:

$$\mathbf{H}_{res} = \text{ReLU}(\mathbf{H}_{llm}\mathbf{W}^{res}_{down})\mathbf{W}^{res}_{up} \tag{6}$$

Finally, as shown in the fusion summation $\bigoplus$ in Figure 2, the output of the HCA block combines three streams: the original LLM residual stream, the cross-attention context, and the query-residual

adaptation:

$$\mathbf{H}_{out} = \mathbf{H}_{llm} + \mathbf{C}_{cross} + \mathbf{H}_{res} \quad (7)$$

This dual-loop design ensures that the model benefits from the *Source-Side Adapter*'s alignment capability while the *Query-Residual Adapter* maintains the generative fluency of the LLM.

4 Experiments

4.1 Datasets and Evaluation Protocols

Table 1 summarizes the data configurations. We adhere to a strict "Train-on-One, Test-on-Many" protocol.

Training Configurations: We construct three distinct training sets, all randomly sampled from the original NLLB corpus, to verify different hypotheses:

1. **Main Zero-Shot (De-En):** We use a large-scale NLLB subset (580k pairs) for the primary German-English model to ensure robust alignment learning.
2. **Robustness Ablation (Ar-En):** We use a distinct Arabic-English subset (270k pairs) to verify the universality of the alignment protocol across language families.
3. **Oracle (All Pairs):** To establish a theoretical upper bound, we construct a balanced multi-lingual dataset covering all 10 language pairs. This includes High-Resource pairs (De, Zh, Ru, Ar) and Low-Resource pairs (Zu, Xh, My, Lo, Gu, Prs). Crucially, we *down-sample* the high-resource pairs to match the scale of low-resource pairs (∼50k), preventing data imbalance.

Data Efficiency and Convergence: It is worth noting that while the complete NLLB training corpora were available for these pairs, our HCA module exhibited extremely rapid convergence. We employed an early stopping strategy based on validation loss plateauing. Consequently, the model achieved optimal alignment using only a fraction of the available data—approximately **580k** pairs for De-En and **270k** pairs for Ar-En. This phenomenon further corroborates our hypothesis that *topological alignment* requires significantly less data density than traditional language modeling or generative fine-tuning.

Evaluation: We employ the **FLORES-200** benchmark (Costa-jussà et al., 2022) (devtest split), utilizing 1,012 sentences per language pair for rigorous zero-shot evaluation. The test set covers 10 languages: De, Zh, Ru, Ar, Zu, Xh, My, Lo, Gu and Prs(refer to Appendix Table 6 for details).

Setting	Lang. Pairs	Source	Size
1. Single-Pair Training (Zero-Shot)			
Main	De → En (Full)	NLLB	580k
Robustness	Ar → En (Full)	NLLB	270k
2. Oracle Training (All-Pairs Balanced)			
Pure NLLB Data			
	De (Sub.) → En	NLLB	40k
	Zh → En	NLLB	59k
Mixed Data (NLLB + NTREX)			
	Ar (Sub.), Ru → En	Mixed	∼45k/ea
	Zu, Xh, My → En Lo, Gu, Prs → En	Mixed	∼52k/ea
3. Evaluation			
Test Set	All 10 Langs → En	FLORES	1,012/ea

Table 1: Dataset Statistics: For Single-Pair experiments, we use the full NLLB capacity. For the Oracle experiment, we create a balanced distribution across all 10 languages. High-resource pairs (De/Zh/Ar/Ru) are down-sampled (marked as "Sub."). The 6 low-resource languages (Zu, Xh, My, Lo, Gu, Prs) supplement NLLB seed data with NTREX. All directions are X → English; "/ea" denotes per language pair.

4.2 Base Models

Our framework consists of two frozen components and one trainable module:

- **Generative Backbone:** We employ Qwen3-4B (Decoder-only) as the target-side generator.
- **Dense Encoders:** We investigate two variants of the NLLB encoder to verify scaling laws:
  1. **NLLB-200-600M**: A distilled variant with 600M parameters.
  2. **NLLB-200-1.3B**: A larger variant with 1.3B parameters.

4.3 Comparative Systems

To benchmark our approach against state-of-the-art multilingual translation paradigms, we compare with three categories of systems:

I. **Base LLMs (Zero-Shot/LoRA):** Standard Qwen3-4B with prompting or standard LoRA fine-tuning on De-En.

II. **Translation-Specialized LLMs:** Recent open-source models optimized via Continuous Pre-training (CPT) and Supervised Fine-Tuning (SFT). We include Bayling2-13B

(Zhang et al., 2024), and Aya-101 (Üstün et al., 2024).

III. **Advanced Scaled Models:** Top-performing large-scale systems, including X-ALMA-29B (Xu et al., 2025) (massively multilingual CPT) , TowerPlus-9B (Rei et al., 2025) and LaMaTE-s2-8B (Luo et al., 2025) (LLM-as-Encoder architecture).

4.4 Implementation Details

Model Configuration: We implement the framework using PyTorch. We insert HCA blocks into all decoder layers with a bottleneck dimension $r = \mathbf{2048}$. The total trainable parameter count for the HCA block is ∼485M. The total inference footprint is approximately **5.25B** ($\mathbf{0.766}$B Encoder + $\mathbf{4.0}$B Decoder + $\mathbf{0.485}$B HCA Adapter). While this creates a dual-model system, the encoder runs only once per source sentence, resulting in an inference speed of **25.2 tokens/s** on a single NVIDIA H200 GPU. For comparison, the standalone Qwen3-4B decoder achieves **40.1 tokens/s**; thus, our framework maintains $\sim\mathbf{63\%}$ of the base generation speed while unlocking universal zero-shot translation capabilities.

Training Protocol: We train the HCA adapter on the NLLB dataset for a single epoch, demonstrating the rapid convergence of our method. We utilize the Adam optimizer with a learning rate of $\mathbf{1}e^{-5}$ and a global batch size of 128. No extensive hyperparameter search was performed, suggesting the robustness of the proposed architecture.

Inference and Evaluation: All experiments were conducted on NVIDIA H200 GPUs. We report performance using COMET scores (Rei et al., 2020) (model: `wmt22-comet-da`) as our primary metric. Following recent WMT recommendations (Kocmi et al., 2021; Freitag et al., 2022), we prioritize neural metrics over traditional lexical overlap metrics like BLEU (Papineni et al., 2002), as neural metrics exhibit significantly higher correlation with human judgment, especially in data-scarce domains. Furthermore, to provide a comprehensive evaluation and robust lexical assessment, we supplement our main COMET findings with **chrF++** scores (Popović, 2015).

5 Results and Analysis

5.1 Main Results

Table 2 presents the comparative performance. We analyze the results through three critical lenses: generalization efficiency, coverage, and synergy.

I. Zero-Shot vs. Massive SFT (Beating the Strongest Baselines). Comparison with Category III models highlights the extreme data efficiency of our approach. While TowerPlus-9B struggles with low-resource languages (Avg: 73.6) due to data scarcity, Aya-101-13B establishes a strong baseline (Avg: 80.7) by leveraging massive instruction tuning across 101 languages. Remarkably, our HCA model (NLLB-1.3B), trained *exclusively* on German-English, achieves a Low-Resource average of **82.6**. This represents a **+1.9 point margin** over the strongest SFT competitor (Aya-101) and a massive **+9.0 point margin** over TowerPlus. Crucially, HCA achieves this using fewer parameters (∼5.25B vs. 13B) and zero target language supervision. This confirms that transferring the dense, pre-aligned topological structure of NLLB is more effective for low-resource translation than brute-force generative fine-tuning.

II. High-Resource Peak vs. Universal Coverage. X-ALMA (29B) sets the upper bound for high-resource translation (Avg High-Res: 88.2), benefiting from its massive scale (6× larger than HCA). However, this performance comes at the cost of coverage; it fails to support half of our test languages. Our method offers a strategic trade-off: while we trail X-ALMA by ∼4.6 points on high-resource languages (where data scaling rules), we provide robust, high-quality coverage for long-tail languages where massive models often lack support. This validates HCA as a superior solution for *universal* and *inclusive* translation systems.

III. Isomorphic Transfer and Synergy. The comparison between our *Zero-Shot (De-En)* and *Oracle (All Pairs)* models reveals a remarkably narrow gap (82.6 vs. 84.5 on Low-Res). This small delta (∼2.1 points) indicates that learning the alignment on a single language pair captures the vast majority of the universal projection protocol. Furthermore, our Oracle model (**85.8**) surpasses the fully supervised NLLB-1.3B Teacher (85.6). This confirms a positive synergy: the HCA block successfully combines the encoder's semantic precision with the decoder's generative fluency, creating a system that exceeds the capabilities of its constituent parts.

IV. Comprehensive Lexical Validation (chrF++). To ensure the robustness of our findings beyond neural metrics, we additionally evaluate the models using chrF++ (Table 3). The results strictly corrob-

orate our primary COMET evaluations. Notably, our Zero-Shot HCA model (trained exclusively on De-En) achieves a Low-Resource average of 56.8. This vastly outperforms massive SOTA baselines, including Aya-101 (50.4) and TowerPlus-9B (44.1), and demonstrates an overwhelming superiority over other specialized models like Bayling2 (27.5) and LaMaTE-s2 (18.7).

On the overall average, the Zero-Shot model attains 58.2, easily surpassing all evaluated general-purpose and translation-specialized LLMs. Furthermore, our Oracle model achieves the highest overall average (59.7), slightly edging out the fully supervised NLLB-1.3B teacher (59.5). The remarkably narrow gap between the Zero-Shot HCA (58.2) and the Oracle upper bound (59.7) confirms that our single-pair topological alignment is highly effective. It proves that the model preserves not only deep semantic equivalence (as measured by COMET) but also guarantees high surface-level structural and lexical precision across unseen, low-resource languages.

5.2 Ablation Study

To rigorously validate our architectural choices and the "Universal Alignment" hypothesis, we conduct comprehensive ablation studies using the **NLLB-1.3B** encoder. Table 4 details the impact on both Low-Resource and Overall averages.

5.2.1 Impact of HCA Components

We dissect the HCA block by selectively removing its key components. The results highlight the distinct roles of each module:

- **Query-Residual Adapter (The Firewall):** Removing this adapter causes performance to plummet to **56.9** (Low Avg), falling significantly below the Qwen Zero-shot baseline (71.3). This is a critical finding: it confirms that without a residual pathway to preserve internal reasoning, raw cross-attention signals act as **destructive noise**, effectively shattering the decoder's language model.
- **Source-Side Pre-conditioning:** Removing the intra-encoder adapter leads to a **2.0-point drop** on low-resource languages ($\mathbf{82.6} \rightarrow \mathbf{80.6}$). This suggests that while NLLB's embedding space is globally isometric, the local topology for long-tail languages requires fine-grained adjustment (rotation or scaling) to align perfectly with the decoder's input manifold.

5.2.2 Robustness to Pivot Language Selection

To verify that our success isn't due to the specific linguistic proximity between German and English, we trained HCA on **Arabic-English (Ar-En)**. As shown in Table 2, this model achieves **77.7** on low-resource languages. While lower than De-En (82.6), it still generalizes effectively to unrelated languages like Zulu and Lao, outperforming TowerPlus (73.6). This confirms the HCA block learns a **language-agnostic projection protocol**—a universal mathematical mapping—rather than overfitting to source-specific linguistic features.

5.2.3 The "Zero-Shot" vs. "Oracle" Gap

Finally, we investigate the theoretical upper bound by comparing the *Zero-Shot (De-En)* model against the *Oracle* model (trained on all pairs).

- **High-Res Avg:** $\mathbf{83.6}$ (Zero-Shot) vs $\mathbf{87.7}$ (Oracle) $\rightarrow$ Gap = **4.1 points**.
- **Low-Res Avg:** $\mathbf{82.6}$ (Zero-Shot) vs $\mathbf{84.5}$ (Oracle) $\rightarrow$ Gap = **1.9 points**.

Single-pair training recovers over **96.7%** of the model's potential capacity. A profound insight here is that the gap is significantly smaller for low-resource languages ($\mathbf{\Delta 1.9}$) than for high-resource ones ($\mathbf{\Delta 4.1}$). This suggests that universal topological alignment is particularly effective for long-tail languages, where translation relies on shared, deep semantic structures captured by NLLB. In contrast, high-resource translation involves more surface-level lexical patterns or idiomatic expressions, which benefit more from direct supervision seen in the Oracle setting. Thus, our Zero-Shot approach is uniquely optimized for solving the long-tail coverage problem.

5.3 Qualitative Analysis: Mitigating Hallucinations

Large Language Models often suffer from "hallucinations" when translating low-resource languages—generating fluent but factually incorrect content. We illustrate this with a Gujarati-to-English biomedical example (Table 5).

Insight: The "Grounding" Effect. This example powerfully demonstrates the *Source-Side Adapter's* role. The Qwen decoder, lacking sufficient probability mass for low-resource Gujarati tokens, guesses a likely animal subject. However, the HCA block injects the precise semantic vector for "Mouse" from NLLB, effectively "grounding" the generation in reality.

Model	Param Size	High-Res De	Zh	Ru	Ar	Low-Res Zu	Xh	My	Lo	Gu	Prs	Average Metrics High	Low	All
Category I: Supervised Dense Encoders (Teacher Models)														
NLLB-600M	0.6B	87.9	84.1	85.0	85.5	78.9	78.0	84.3	84.5	89.1	84.5	85.6	83.2	84.2
NLLB-1.3B	1.3B	88.7	85.4	85.9	87.0	**80.9**	**80.1**	**86.1**	85.7	89.9	86.0	86.8	**84.8**	85.6
Category II: General-Purpose LLMs														
Qwen3 (0-shot)	4B	89.1	87.2	86.3	86.2	51.0	53.4	75.0	76.8	87.4	84.4	87.2	71.3	77.7
Qwen3+LoRA	4B	88.6	86.8	85.9	86.0	64.0	61.9	74.4	79.9	86.8	83.4	86.8	75.1	79.8
Category III: Translation-Specialized SOTA (CPT / SFT)														
Bayling2	13B	88.5	86.3	85.4	78.6	47.6	48.5	56.2	57.0	74.2	78.1	84.7	60.3	70.0
Aya-101	13B	88.1	84.3	85.4	85.6	74.9	71.5	80.3	85.2	86.6	85.6	85.9	80.7	82.8
TowerPlus	9B	**90.0**	**87.8**	87.5	87.1	63.7	63.1	71.6	66.7	89.2	**87.1**	**88.1**	73.6	79.4
X-ALMA	29B	89.7	87.5	**87.7**	**87.9**	-	-	-	-	**90.2**	-	**88.2**	-	-
LaMaTE-s2	8B	84.7	83.5	84.1	83.9	76.5	73.4	80.1	78.2	81.5	82.1	84.1	78.6	80.8
Ours (HCA Framework - Trained on De-En ONLY)														
Ours (600M)	4B*	81.9	82.3	83.9	83.8	77.5	76.2	82.1	82.4	84.1	83.0	83.0	80.9	81.7
Ours (1.3B)	4B*	82.1	82.5	84.6	85.0	78.8	78.1	83.8	82.8	87.5	84.2	83.6	82.6	82.9
Ours (Ablation: Trained on Ar-En ONLY)														
Ours (600M)	4B*	70.3	75.8	78.2	79.7	72.0	71.0	76.8	78.0	81.1	77.8	76.0	76.1	76.1
Ours (1.3B)	4B*	82.3	75.5	80.1	81.7	74.1	72.3	77.4	79.3	82.8	80.3	79.9	77.7	78.6
Ours (Oracle: Trained on All Pairs)														
Ours (600M)	4B*	88.9	85.6	85.1	86.8	78.9	78.5	83.1	83.2	87.6	84.7	86.6	82.7	84.2
Ours (1.3B)	4B*	89.6	87.2	86.7	87.2	80.8	79.6	**86.1**	**85.9**	88.9	85.8	87.7	84.5	**85.8**

Table 2: Comprehensive Main Results (COMET). Green numbers indicate pairs included in training. **Bold** is best, underline is second best. Our **Zero-Shot (De-En)** model (Ours 1.3B) significantly outperforms massive baselines like TowerPlus (Low-Res Avg: 82.6 vs 73.6). Crucially, our **Oracle** model (All Pairs) achieves the highest overall average (85.8), confirming the efficacy of the HCA architecture.

Model	Size	High	Low	Avg.
Category I & II: Baselines & Teacher				
NLLB-1.3B (Teacher)	1.3B	60.8	**58.7**	59.5
Qwen3(0-shot)	4B	59.4	39.2	47.3
Qwen3+LoRA	4B	58.9	42.6	49.1
Category III: SOTA-MT Baselines				
Bayling2	13B	56.4	27.5	39.0
Aya-101	13B	56.1	50.4	52.6
LaMaTE-s2	8B	44.5	18.7	29.0
TowerPlus	9B	**63.1**	44.1	51.7
Ours				
Ours (HCA De-En)	4B*	60.3	56.8	58.2
Ours (Oracle All-Pairs)	4B*	62.2	58.1	**59.7**

Table 3: Comprehensive chrF++ Evaluation. Consistent with our COMET findings, the Zero-Shot HCA model significantly outperforms both massive SOTA baselines (Aya-101, TowerPlus) and specialized modular models (Bayling2, LaMaTE-s2) on low-resource languages. This robustly validates the structural and lexical precision of our deep-fusion alignment.

Configuration	Avg (Low)	Avg (All)
Full HCA (De-En)	**82.6**	**82.9**
w/o Source-Side Adapter	80.6 (↓ 2.0)	81.7 (↓ 1.2)
w/o Query-Residual	56.9 (↓ 25.7)	56.6 (↓ 26.3)
Ref: Qwen3 Zero-shot	*71.3*	*77.7*

Table 4: Component Ablation. Removing the Query-Residual Adapter leads to a performance collapse below the Qwen3 zero-shot baseline, indicating that raw cross-attention signals cause destructive interference without a dedicated residual pathway to preserve generative stability.

6 Discussion

The negligible gap between Zero-Shot and Oracle performance invites a re-examination of multilingual scaling laws.

6.1 The Isomorphic Alignment Hypothesis

We propose the **Isomorphic Alignment Hypothesis**: The embedding spaces of dense encoders (NLLB) and LLMs (Qwen) share high topological isomorphism. The geometric transformation mapping *German* to *English* is structurally identical to that for *Zulu* or *Gujarati*. Consequently, the HCA adapter learns a **global rotation matrix** rather than language-specific features. Our robustness experiment with Arabic-English (Ar-En) confirms this: the universal protocol transfers even from a linguistically distant anchor. This suggests future research should pivot from scaling data quantity to optimiz-

Model	Translation & Analysis
Source *(Gu)*	તેમણે ઉમેર્યું, "હવે અમારી પાસે 4 મહિના પહેલા જન્મેલા ઉંદરો છે જેઓ પહેલા ડાયાબિટીસથી પીડાતા હતા પરંતુ હવે તેઓને ડાયાબિટીસ નથી." ***Reference:*** *He added, "Now we have 4-month-old mice who used to suffer from diabetes but now they do not."*
Qwen *(0-shot)*	"He added, "Now we have **monkeys** born four months ago who previously suffered from diabetes, but they no longer have diabetes."" **Analysis:** *Critical Hallucination. Qwen produces a grammatically fluent sentence but catastrophically mistranslates the core entity "**mice**" as "**monkeys**", a factual error rendering the translation dangerous.*
NLLB *(Teacher)*	"We now have **mice** born 4 months ago who were previously suffering from diabetes..." **Analysis:** *Rigid Phrasing. Semantically correct (identifies mice), but the phrasing is verbose and lacks the natural flow of English.*
HCA *(Ours)*	**"We now have 4-month-old rats that used to have diabetes but now they don't have diabetes," he added."** **Analysis:** *Grounded & Fluent. HCA inherits the correct entity (**rats/mice**) from NLLB, correcting the hallucination while utilizing the LLM's fluency.*

Table 5: Qualitative case study on unseen Gujarati.

ing *topological alignment*.

6.2 Efficiency and Modularity

Green AI. Unlike baselines requiring massive CPT, our "Train-on-One" paradigm is highly sustainable. HCA is trained for only 1 epoch on De-En (∼580K pairs), consuming less than 1% of the compute required for full fine-tuning.

Plug-and-Play Upgradability. By decoupling "understanding" (Encoder) from "generation" (LLM), our framework enables independent component upgrades. As shown in our scaling experiments, upgrading the encoder (600M → 1.3B) yields immediate performance gains without retraining the LLM, significantly extending system lifespan.

7 Conclusion

We challenge the dogma that multilingual translation requires massive data scaling. We introduce Hybrid Cross-Alignment (HCA), a parameter-efficient framework synergizing NLLB and Qwen via a closed-loop dual-adapter.

Our experiments confirm that One Pair Suffices: single-pair training recovers 96.7% of the Oracle performance. This "Universal Alignment Generalization" allows our **∼5.25B** model to significantly outperform TowerPlus-9B (by +9.0 points on low-resource languages) and offer robust coverage where X-ALMA-29B lacks support.

We conclude that the barrier to universal multilinguality is *alignment topology*, not data scarcity. HCA paves the way for the next generation of modular, efficient, and universally accessible translation systems.

8 Limitations

While promising, our framework has limitations. First, the system's performance is strictly bounded by the **encoder's upper bound**. If the frozen encoder fails to capture the semantics of a low-resource language, the HCA block cannot recover this information. Second, this work focuses on **X-to-English** translation. Extending this to *English-to-X* or *Many-to-Many* translation remains a challenge, as it requires the LLM to generate in low-resource languages, which is often a capability gap in English-centric LLMs. Finally, the introduction of an encoder adds inference latency compared to decoder-only architectures, although this is often a worthwhile trade-off for the massive gain in low-resource coverage.

9 Acknowledgment

This work is supported by the Fundamental Research Funds for the Central Universities of China under Grant 2024JBGP008 and the National Natural Science Foundation of China (U1936109,62406018).

References

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. 2024. Tower: An open multilingual large language model for translation-related tasks. *Preprint*, arXiv:2402.17733.

Rachit Bansal, Bidisha Samanta, Siddharth Dalmia, Nitish Gupta, Sriram Ganapathy, Abhishek Bapna, Prateek Jain, and Partha Talukdar. 2024. LLM augmented LLMs: Expanding capabilities through composition. In *The Twelfth International Conference on Learning Representations*.

Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, and 1 others. 2023. Seamlessm4t: massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*.

Guanhua Chen, Shuming Ma, Yun Chen, Li Dong, Dongdong Zhang, Jia Pan, Wenping Wang, and Furu Wei. 2021. Zero-shot cross-lingual transfer of neural machine translation with multilingual pretrained encoders. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 15–26, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marta R Costa-jussà, James Tran, Artem Sokolov, Tripti Dewan, Guillaume Wenzek, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. 2023. Sonar: sentence-level multimodal and language-agnostic representations. *arXiv preprint arXiv:2308.11466*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*.

Sandhini Agarwal Lama Ahmad Ilge Akkaya Josh Achiam, Steven Adler and others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 19730–19742.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yingfeng Luo, Tong Zheng, Yongyu Mu, Bei Li, Qinghong Zhang, Yongqi Gao, Ziqiang Xu, Peinan Feng, Xiaoqian Liu, Tong Xiao, and Jingbo Zhu. 2025. Beyond decoder-only: Large language models can be good encoders for machine translation. *Preprint*, arXiv:2503.06594.

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*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, João Alves, Pedro Teixeirinha, Amin Farajian, and André F. T. Martins. 2025. Tower+: Bridging generality and translation specialization in multilingual llms. *Preprint*, arXiv:2506.17080.

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.

Fabian David Schmidt, Philipp Borchert, Ivan Vulić, and Goran Glavaš. 2024. Self-distillation for model stacking unlocks cross-lingual NLU in 200+ languages. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6724–6743, Miami, Florida, USA. Association for Computational Linguistics.

Fanqi Wan, Longguang Zhong, Ziyi Yang, Ruijun Chen, and Xiaojun Quan. 2025. FuseChat: Knowledge fusion of chat models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 21618–21642, Suzhou, China. Association for Computational Linguistics.

Hao Wang, Linlong Xu, Heng Liu, Yangyang Liu, Xiaohu Zhao, Bo Zeng, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2025. Marco large translation model at wmt2025: Transforming translation capability in llms via quality-aware training and decoding.

In *Proceedings of the Tenth Conference on Machine Translation*, pages 587–593.

Haoran Xu, Kenton Murray, Philipp Koehn, Hieu Hoang, Akiko Eriguchi, and Huda Khayrallah. 2025. X-ALMA: Plug & play modules and adaptive rejection for quality translation at scale. In *The Thirteenth International Conference on Learning Representations*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. 2023. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*.

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

Bo Zheng, Li Dong, Shaohan Huang, Saksham Singhal, Wanxiang Che, Ting Liu, Xia Song, and Furu Wei. 2021. Allocating large vocabulary capacity for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3203–3215.

Mao Zheng, Zheng Li, Yang Du, Bingxin Qu, and Mingyang Song. 2025. Shy-hunyuan-mt at wmt25 general machine translation shared task. In *Proceedings of the Tenth Conference on Machine Translation*, pages 607–613.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. *Preprint*, arXiv:2402.07827.

A Language Codes and Details

Table 6 lists the ISO language codes, FLORES-200 identifiers, and full language names used in our experiments. We categorize them into High-Resource and Low-Resource groups consistent with our experimental setup.

Category	Abbr.	FLORES Code	Language
Source	De	deu_Latn	German
	En	eng_Latn	English
High-Res	Zh	zho_Hans	Chinese
	Ru	rus_Cyrl	Russian
	Ar	arb_Arab	Arabic
Low-Res	Zu	zul_Latn	Zulu
	Xh	xho_Latn	Xhosa
	My	mya_Mymr	Burmese
	Lo	lao_Laoo	Lao
	Gu	guj_Gujr	Gujarati
	Prs	prs_Arab	Dari

Table 6: **Language Details.** Mapping of abbreviations used in this paper to standard FLORES-200 codes and full language names. *Prs* (Dari) is a variety of Persian spoken in Afghanistan, distinct from *pes* (Western Persian).