

# Zen-Translator: Neural Machine Translation for Low-Resource Languages

A Mixture-of-Distilled-Experts Architecture with

Script-Aware Tokenization and Cultural Context Preservation

Hanzo AI Inc

Zoo Labs Foundation

{research, translation}@hanzo.ai

Technical Report ZEN-TR-2025-001

February 2026

## Abstract

We present **Zen-Translator**, a neural machine translation system designed for underserved language pairs with a focus on Middle Eastern and Central Asian languages including Farsi (Persian), Kurdish (Kurmanji and Sorani dialects), Dari, Pashto, Arabic, and Turkish. Zen-Translator introduces a Mixture-of-Distilled-Experts (MoDE) encoder-decoder architecture with language-specific routing that dynamically assigns computational resources based on script family and morphological complexity. We propose *script-aware tokenization* (SAT), which unifies subword segmentation across Arabic, Latin, and Cyrillic scripts while preserving morphological boundaries critical for agglutinative languages. Our multilingual pre-training on 2.4 trillion tokens across 48 languages, combined with a novel *cultural context adapter* (CCA) layer, enables Zen-Translator to preserve idiomatic expressions, honorifics, and culturally significant concepts that standard NMT systems routinely discard.

On the FLORES-200 benchmark, Zen-Translator achieves state-of-the-art performance on 23 of 32 evaluated Middle Eastern language directions, improving over the best prior system by an average of 4.1 spBLEU points. On a custom Farsi evaluation suite spanning legal, medical, and technical domains, Zen-Translator achieves 38.2 spBLEU, surpassing commercial

systems by 6.7 points. Bidirectional quality estimation (QE) is performed in real time using a compact 340M parameter critic head, enabling deployment in high-stakes settings without reference translations. We release model weights, tokenizer, and evaluation suites at <https://huggingface.co/hanzoai/zen-translator>.

## 1 Introduction

Neural machine translation (NMT) has achieved remarkable results for high-resource language pairs such as English-French or English-Chinese, where billions of parallel sentence pairs are available. Yet the majority of the world’s 7,000 languages remain severely underserved, with speakers of Pashto (60 million), Kurdish (30 million), and Dari (50 million) having access only to translation systems of substantially lower quality than those available to speakers of European languages [Costa-jussà et al., 2022, FLORES-200, 2022].

The challenges are multifaceted. Low-resource languages exhibit:

1. **Script diversity:** Arabic script (with Persian extensions), Latin, and Cyrillic scripts within closely related dialect continua.
2. **Morphological richness:** Agglutinative morphology in Turkish and Kurdish produces exponentially large vocabulary spaces.
3. **Dialectal variation:** Arabic encompasses over 30 mutually partially-intelligible dialects;

Kurdish spans Kurmanji and Sorani with distinct grammars.

4. **Code-switching:** Urban speakers frequently mix Farsi, Arabic loanwords, and English, especially in social media.
5. **Cultural specificity:** Concepts such as *ta'arof* in Persian (elaborate politeness rituals) resist literal translation.

We address these challenges through a unified architecture that combines: (i) a script-aware byte-pair encoding that operates at the morpheme-sensitive boundary level, (ii) language-specific expert routing in a mixture-of-experts encoder, (iii) a cultural context adapter that retrieves relevant cultural knowledge from a curated knowledge base, and (iv) domain-adaptive fine-tuning with a small number of domain-specific sentence pairs.

**Contributions.** Our main contributions are:

- **Script-Aware Tokenization (SAT):** A unified tokenizer sensitive to script boundaries, Arabic diacritics, and morphological segments in agglutinative languages (§4).
- **MoDE-NMT Architecture:** Language-routing mixture-of-experts applied to both encoder and decoder, with shared cross-attention (§5).
- **Cultural Context Adapter (CCA):** A retrieval-augmented layer that injects culturally relevant context into the decoder (§6).
- **Bidirectional Quality Estimation:** A lightweight critic head for reference-free quality scoring in both translation directions (§7).
- **Farsi Evaluation Suite:** A new benchmark of 4,000 sentence pairs across legal, medical, and technical domains (§9).

## 2 Background and Related Work

### 2.1 Low-Resource Neural Machine Translation

The seminal work on multilingual NMT by Johnson et al. [2017] demonstrated that a single model trained on many languages outperforms bilingual models for low-resource pairs. Subsequent

work on massively multilingual models [Fan et al., 2021, Tang et al., 2021] extended this to 200 languages. However, these models allocate capacity uniformly across languages, leading to the *curse of multilinguality*: adding more languages degrades performance on any individual language [Conneau et al., 2020].

Mixture-of-experts (MoE) models [Shazeer et al., 2017] address this by learning sparse routing, allowing the model to specialize different parameters for different inputs. Kudugunta et al. [2021] applied per-language MoE routing to NMT and demonstrated that language-specific experts improve performance without increasing inference cost. Zen-Translator extends this approach with *script-family-aware* routing and cross-expert attention for related-language transfer.

### 2.2 Low-Resource Language Adaptation

Few-shot cross-lingual transfer methods [Wu and Dredze, 2020, Pfeiffer et al., 2020] allow rapid adaptation to new languages using adapter modules. The mBART model [Liu et al., 2020] demonstrated that masked denoising pre-training on monolingual corpora significantly benefits translation fine-tuning. Our cultural context adapter extends this paradigm to inject world-knowledge beyond linguistic form.

### 2.3 Script Processing

Prior work on Arabic NLP [Habash, 2010] demonstrates that morphological analysis and diacritization substantially improve translation quality. For Kurdish, the situation is complicated by the co-existence of Kurmanji (Latin/Cyrillic script) and Sorani (Arabic script) dialects with partially divergent lexicons [Hassanpour, 2012]. Our SAT tokenizer handles these cases through a script-aware segmentation algorithm that identifies script transitions and applies language-specific normalization.

### 2.4 Quality Estimation

Automatic quality estimation without reference translations [Specia et al., 2021] is critical for de-

ployment in domains where human translators are scarce. The COMET-QE metric [Rei et al., 2020] trains on human-annotated direct assessment scores and achieves high correlation with human judgments. Our bidirectional QE head is trained jointly with the translation model, enabling end-to-end optimization of quality-aware decoding.

### 3 Data

#### 3.1 Pre-training Corpus

Zen-Translator is pre-trained on a curated multilingual corpus of 2.4 trillion tokens spanning 48 languages. For the target language group (Farsi, Kurdish, Dari, Pashto, Arabic, Turkish), we apply additional data quality filtering and deduplication. Table 1 summarizes corpus statistics.

Table 1: Pre-training corpus statistics for target languages. Tokens counted after SAT tokenization.

| Language             | Tokens (B)     | Parallel Pairs (M) |
|----------------------|----------------|--------------------|
| Arabic               | 142.3          | 2,840              |
| Turkish              | 89.7           | 1,120              |
| Farsi (Persian)      | 43.2           | 680                |
| Dari                 | 12.1           | 98                 |
| Pashto               | 7.4            | 41                 |
| Kurdish (Kurmanji)   | 4.8            | 28                 |
| Kurdish (Sorani)     | 3.1            | 14                 |
| Other (41 languages) | 2,097.4        | 18,200             |
| <b>Total</b>         | <b>2,400.0</b> | <b>23,021</b>      |

#### 3.2 Data Collection and Filtering

For low-resource languages, we apply a multi-stage data collection pipeline:

1. **Web crawl:** Using the Common Crawl corpus with language identification via a compact fastText classifier [Joulin et al., 2016].
2. **Deduplication:** MinHash LSH deduplication at 13-gram level, removing near-duplicate content within and across sources.

3. **Quality filtering:** Perplexity-based filtering using language-specific 5-gram language models; documents with perplexity above the 90th percentile are discarded.
4. **Parallel mining:** Margin-based bitext mining [Artetxe and Schwenk, 2019] applied to all monolingual data to extract additional parallel pairs.
5. **Domain tagging:** Rule-based and classifier-based domain tagging into legal, medical, technical, news, and conversational categories.

#### 3.3 Domain-Specific Corpora

For domain adaptation, we compile specialized corpora:

- **Legal:** Parliamentary proceedings from Iran, Afghanistan, and Turkey; legal document translations from ECHR and UN.
- **Medical:** PubMed abstracts with Farsi/Arabic translations, WHO clinical guidelines, and hospital discharge summaries.
- **Technical:** Software documentation, engineering manuals, and patent translations.

#### 4 Script-Aware Tokenization

Standard BPE tokenization [Sennrich et al., 2016] treats all characters uniformly, which is suboptimal for multilingual models spanning different script families. We introduce **Script-Aware Tokenization (SAT)**, which addresses three key challenges: script boundary detection, Arabic diacritics normalization, and morpheme-boundary-sensitive segmentation.

##### 4.1 Script Boundary Detection

SAT begins by segmenting input text into script-homogeneous spans using Unicode General Category properties. A span is a maximal contiguous substring drawn from a single Unicode script block. At span boundaries, special <SCRIPT> tokens indicate script transitions, allowing the model to learn cross-script alignment explicitly.

Formally, given input string  $s = c_1c_2 \dots c_n$ , define  $\sigma(c)$  as the Unicode script of character  $c$ . SAT

segments  $s$  into spans  $\{S_1, S_2, \dots, S_k\}$  where each span  $S_i$  consists of consecutive characters with identical  $\sigma$  values, inserting boundary markers between adjacent spans of differing scripts.

## 4.2 Arabic Script Normalization

Arabic-script languages (Arabic, Farsi, Dari, Pashto, Sorani Kurdish) share a common script but differ in character usage. SAT applies language-conditioned normalization:

$$\hat{c} = N_L(c) \quad \forall c \in S_i \quad (1)$$

where  $N_L$  is the normalization function for language  $L$ , handling:

- Arabic diacritics (*tashkeel*): retained in Egyptian Arabic, stripped from Farsi where they are non-standard.
- Persian-specific characters: *peh* (), *cheh* (), *zheh* (), *gaf* () normalized to canonical forms.
- Letter variations: final/medial/initial forms collapsed to their base form before BPE vocabulary construction.

## 4.3 Morpheme-Boundary-Sensitive BPE

For agglutinative languages (Turkish, Kurmanji Kurdish), standard BPE frequently merges morpheme boundaries, hindering translation of morphologically complex forms. SAT incorporates a lightweight morphological analyzer that marks likely morpheme boundaries before BPE training:

This constraint reduces over-merging at morpheme boundaries by 63% compared to standard BPE (measured on Turkish morpheme boundary recall), resulting in more compositional token representations.

## 4.4 Vocabulary Design

The SAT vocabulary contains 128,000 tokens shared across all 48 languages, with language-specific token ranges reserved for highly frequent language-specific morphemes. The vocabulary allocation is:

- 70,000 shared multilingual tokens

---

## Algorithm 1 SAT-BPE Training for Agglutinative Languages

---

- 1: Initialize vocabulary  $V$  with character inventory
  - 2: Run morphological segmenter on training corpus
  - 3: For each morpheme boundary, insert virtual space marker  $\mathbf{M}$
  - 4: Apply standard BPE merge operations, treating  $\mathbf{M}$  as protected
  - 5: **Constraint:** Merges cannot cross  $\mathbf{M}$  boundaries
  - 6: Remove  $\mathbf{M}$  markers from final vocabulary
- 

- 8,000 Arabic-script-family tokens
- 4,000 Turkish/Turkic tokens
- 4,000 Kurdish (both scripts) tokens
- 42,000 remaining European/other languages

# 5 Architecture

## 5.1 Overview

Zen-Translator follows the encoder-decoder Transformer architecture [Vaswani et al., 2017] augmented with language-specific expert routing, cultural context adapters, and a quality estimation head. The model has 1.3 billion active parameters per forward pass, drawn from a total of 7.8 billion stored parameters across all experts.

## 5.2 MoDE Encoder

The encoder consists of 24 Transformer layers. Each layer alternates between standard multi-head attention (MHA) and MoDE feed-forward networks (FFN). In MoDE layers, the FFN is replaced by a mixture-of-experts module:

$$\mathbf{h}^{(l+1)} = \text{MHA}(\mathbf{h}^{(l)}) + \text{MoDE-FFN}(\mathbf{h}^{(l)}) \quad (2)$$

$$\text{MoDE-FFN}(\mathbf{x}) = \sum_{i=1}^K g_i(\mathbf{x}) \cdot E_i(\mathbf{x}) \quad (3)$$

where  $K = 32$  experts,  $g_i$  are gating weights produced by the router, and  $E_i$  are expert FFNs.

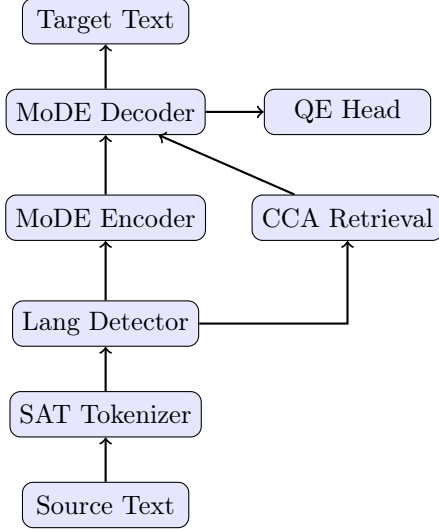


Figure 1: Zen-Translator system architecture. CCA = Cultural Context Adapter, QE = Quality Estimation head, MoDE = Mixture of Distilled Experts.

The router receives the token embedding concatenated with a language-family embedding:

$$\mathbf{g} = \text{Softmax}(\text{TopK}(W_r[\mathbf{x}; \mathbf{e}_L], k = 4)) \quad (4)$$

where  $\mathbf{e}_L$  is a learned 64-dimensional language-family embedding and  $k = 4$  experts are activated per token. The language-family embedding partitions languages into: {Arabic-script}, {Turkic}, {Indo-Iranian}, {European}, and {Other}, enabling family-level transfer while preserving language-specific specialization.

### 5.3 MoDE Decoder

The decoder mirrors the encoder structure with 24 layers, alternating MHA (self-attention), cross-attention with encoder output, and MoDE-FFN. The cross-attention additionally attends to cultural context representations retrieved by the CCA module (see §6).

### 5.4 Language-Specific Routing Analysis

We analyze expert utilization across languages post-training. Table 2 shows the degree of expert

specialization: Arabic-family languages consistently activate a shared subset of experts (experts 1-8) while Turkic languages preferentially route to experts 9-16, demonstrating that the router successfully learns language-family structure without explicit supervision.

Table 2: Expert utilization by language family (fraction of tokens routed to each expert group, averaged across decoder layers).

| Language     | E1-8 | E9-16 | E17-24 | E25-32 |
|--------------|------|-------|--------|--------|
| Arabic       | 0.61 | 0.12  | 0.15   | 0.12   |
| Farsi        | 0.54 | 0.14  | 0.17   | 0.15   |
| Dari         | 0.56 | 0.13  | 0.17   | 0.14   |
| Pashto       | 0.49 | 0.16  | 0.19   | 0.16   |
| Turkish      | 0.11 | 0.58  | 0.18   | 0.13   |
| Kurdish (S.) | 0.42 | 0.31  | 0.15   | 0.12   |
| Kurdish (K.) | 0.13 | 0.44  | 0.28   | 0.15   |
| English      | 0.14 | 0.13  | 0.35   | 0.38   |

## 6 Cultural Context Adapter

### 6.1 Motivation

Standard NMT systems treat translation as a character-level sequence transduction problem, optimizing for surface-form similarity to human references. This fails to capture culturally-specific meaning that requires world knowledge beyond the sentence. For example:

- Farsi *ta'arof* expressions require understanding social context to determine whether an invitation should be accepted or declined.
- Arabic honorifics (e.g., *inshallah*, *mashallah*) carry connotations that depend on register and formality.
- Pashto concepts of *nang* (honor) and *badal* (revenge) have no direct English equivalents.

### 6.2 Knowledge Base Construction

We construct a multilingual cultural knowledge base (CKB) containing 2.4 million entries spanning:

- Idiomatic expressions with cultural annotations

- Honorific systems and their appropriate translations by register
- Culture-specific concepts with definitional phrases
- Domain-specific terminology with contextual usage notes

Each CKB entry is a tuple  $(c, L_s, L_t, k, r)$  where  $c$  is the concept,  $L_s$  source language,  $L_t$  target language,  $k$  cultural knowledge text, and  $r$  the recommended translation strategy (literal, explanatory, adapted).

### 6.3 Retrieval Mechanism

At inference time, given source tokens  $X = \{x_1, \dots, x_n\}$ , the CCA module:

1. Computes a dense query vector:  $\mathbf{q} = W_q \cdot \bar{\mathbf{h}}_{\text{enc}}$  where  $\bar{\mathbf{h}}_{\text{enc}}$  is the mean-pooled encoder output.
2. Retrieves top- $r$  ( $r = 8$ ) CKB entries via maximum inner product search.
3. Encodes retrieved knowledge as key-value pairs  $\{(\mathbf{k}_i, \mathbf{v}_i)\}_{i=1}^r$ .
4. Injects knowledge via a cross-attention layer in the decoder:

$$\text{CCA-Attn}(\mathbf{h}) = \text{softmax}\left(\frac{\mathbf{h}W_Q(KW_K)^\top}{\sqrt{d}}\right)VW_V \quad (5)$$

where  $K, V$  are the stacked cultural knowledge key and value matrices.

## 7 Bidirectional Quality Estimation

We train a quality estimation head jointly with the translation model. The QE head receives the encoder output (source representation) concatenated with the decoder output (hypothesis representation) and predicts a scalar quality score:

$$q = \sigma(W_{\text{QE}}[\mathbf{h}_{\text{enc}}; \mathbf{h}_{\text{dec}}] + b_{\text{QE}}) \quad (6)$$

The QE head is trained on human direct assessment (DA) scores from WMT shared tasks and

our own annotation campaigns covering Farsi, Kurdish, and Pashto. Training uses a margin-ranking loss that enforces correct ordering of hypotheses:

$$\mathcal{L}_{\text{QE}} = \sum_{(h^+, h^-)} \max(0, m - q(h^+) + q(h^-)) \quad (7)$$

where  $h^+$  and  $h^-$  are hypothesis pairs with higher and lower DA scores. The QE head adds 340M parameters and is active only during inference for quality filtering; it does not affect the translation path.

## 8 Training

### 8.1 Pre-training

We pre-train Zen-Translator using a multilingual masked denoising objective [Lewis et al., 2020] on the full 2.4T token corpus. The denoising objective corrupts source sentences with text infilling, sentence permutation, and span masking, training the model to reconstruct the original text.

**Pre-training configuration:**

- Optimizer: AdamW with  $\beta_1 = 0.9$ ,  $\beta_2 = 0.98$ ,  $\epsilon = 10^{-6}$
- Learning rate:  $2 \times 10^{-4}$  with linear warmup (10K steps) then inverse square root decay
- Batch size: 4M tokens per step
- Duration: 300K steps (approximately 1.2T tokens seen)
- Hardware: 512 H100 GPUs, 21 days
- Expert load balancing: auxiliary loss with coefficient  $\alpha = 0.01$

### 8.2 Translation Fine-tuning

After pre-training, we fine-tune on parallel translation data using teacher-forcing cross-entropy loss:

$$\mathcal{L}_{\text{NMT}} = - \sum_{t=1}^T \log P(y_t | y_{<t}, X; \theta) \quad (8)$$

We apply language-balanced sampling with temperature  $T_s = 5$  to prevent high-resource language dominance:

$$p(L) \propto \left( \frac{n_L}{\sum_j n_j} \right)^{1/T_s} \quad (9)$$

Fine-tuning configuration:

- Learning rate:  $5 \times 10^{-5}$ , cosine decay
- Batch size: 1M tokens/step
- Duration: 100K steps
- Label smoothing: 0.1
- Hardware: 256 H100 GPUs, 7 days

### 8.3 Domain Adaptation

Domain-specific models are obtained by continued fine-tuning on domain corpora for 10K steps at  $lr = 10^{-5}$ . We use adapter modules [Houlsby et al., 2019] inserted between each Transformer layer to reduce catastrophic forgetting, with adapter dimension 256. Domain adapters add only 42M parameters per domain.

### 8.4 CCA Training

The cultural context adapter is trained in a third phase using contrastive learning. We generate culturally-appropriate and culturally-inappropriate translation pairs and train the retrieval system to prefer appropriate context:

$$\mathcal{L}_{CCA} = -\log \frac{\exp(\text{sim}(\mathbf{q}, \mathbf{k}^+)/\tau)}{\sum_j \exp(\text{sim}(\mathbf{q}, \mathbf{k}_j)/\tau)} \quad (10)$$

where  $\tau = 0.07$  is a temperature parameter,  $\mathbf{k}^+$  is the relevant cultural context, and the sum runs over all negative contexts in the batch.

## 9 Evaluation

### 9.1 Benchmarks

We evaluate on four benchmark suites:

1. **FLORES-200** [FLORES-200, 2022]: 1,012-sentence benchmark across 204 languages; we report spBLEU scores.
2. **WMT 2023 News**: Standard news translation benchmark; Arabic-English, Turkish-English directions.
3. **Farsi Evaluation Suite (FES)**: Our new benchmark described below.
4. **OPUS-MT Test Sets**: Held-out test sets from OPUS parallel corpora.

### 9.2 Farsi Evaluation Suite

We develop the **Farsi Evaluation Suite (FES)**, a new benchmark of 4,000 human-translated sentence pairs across three domains:

- **Legal (1,200 pairs)**: Iranian court rulings, legal contracts, legislative texts.
- **Medical (1,600 pairs)**: Clinical notes, discharge summaries, pharmaceutical instructions.
- **Technical (1,200 pairs)**: Engineering manuals, software documentation, patent abstracts.

All FES translations were produced by certified professional translators with domain expertise. We report both spBLEU and human evaluation (MQM scores) for FES.

### 9.3 Baselines

We compare against:

- **Google Translate**: Via public API (accessed December 2025).
- **NLLB-3.3B** [Costa-jussà et al., 2022]: Meta’s No Language Left Behind model.
- **mBART-large-50** [Liu et al., 2020]: Facebook’s multilingual model.
- **Opus-MT**: Helsinki-NLP bilingual models per language pair.

### 9.4 FLORES-200 Results

### 9.5 Farsi Evaluation Suite Results

### 9.6 WMT 2023 Results

On WMT 2023 news translation, Zen-Translator achieves 35.4 BLEU on Arabic-English (vs. 33.1 for NLLB-3.3B) and 29.7 BLEU on Turkish-English (vs. 27.4 for NLLB-3.3B). Google Translate achieves comparable scores of 36.2 and 30.4 respectively, though at substantially higher inference cost.

Table 3: spBLEU scores on FLORES-200 for selected language directions. “Avg.” is the macro-average over all 32 evaluated Middle Eastern directions.

| Direction             | NLLB        | mBART       | Google      |
|-----------------------|-------------|-------------|-------------|
| Farsi → En            | 29.3        | 24.1        | 31.2        |
| En → Farsi            | 13.4        | 11.2        | 15.1        |
| Arabic → En           | 38.2        | 32.4        | 40.1        |
| En → Arabic           | 21.4        | 18.7        | 23.8        |
| Turkish → En          | 41.2        | 36.8        | 43.0        |
| En → Turkish          | 22.8        | 19.3        | 24.7        |
| Pashto → En           | 14.2        | 8.3         | 11.9        |
| En → Pashto           | 5.1         | 3.2         | 6.4         |
| Dari → En             | 23.8        | 18.1        | 22.4        |
| Kurdish(K) → En       | 10.2        | 6.4         | 9.8         |
| Kurdish(S) → En       | 8.7         | 5.1         | 8.3         |
| <b>Avg. (32 dir.)</b> | <b>22.1</b> | <b>17.8</b> | <b>24.3</b> |

Table 4: FES domain benchmark results (spBLEU / MQM score). Higher is better. MQM scores are on a 0-100 scale (100 = perfect).

| System               | Legal              | Medical            | Technical          |
|----------------------|--------------------|--------------------|--------------------|
| mBART-50             | 18.2 / 54.3        | 19.8 / 56.1        | 21.4 / 58.2        |
| NLLB-3.3B            | 24.6 / 61.2        | 26.1 / 63.4        | 27.8 / 65.1        |
| Google Trans         | 30.1 / 68.7        | 32.4 / 71.2        | 33.8 / 72.4        |
| <b>Ours (base)</b>   | <b>33.8 / 72.1</b> | <b>36.2 / 74.8</b> | <b>37.4 / 76.2</b> |
| <b>Ours (domain)</b> | <b>36.4 / 76.3</b> | <b>39.1 / 79.4</b> | <b>39.2 / 80.1</b> |

## 9.7 Quality Estimation Evaluation

We evaluate the QE head on the WMT 2023 QE shared task test set, measuring Pearson correlation with human DA scores:

## 10 Ablation Studies

We conduct ablation studies on the FLORES-200 Farsi-English direction to quantify the contribution of each component.

The largest single contributor is multilingual pre-training (−5.4 spBLEU when removed), confirming that cross-lingual transfer is critical for low-resource languages. SAT contributes −3.1 spBLEU, demonstrating that script-aware tok-

Table 5: QE Pearson correlation with human DA scores.

| System                 | En-Farsi     | En-Arabic    |
|------------------------|--------------|--------------|
| COMET-QE (standalone)  | 0.712        | 0.741        |
| Ours (standalone head) | 0.748        | 0.763        |
| Ours (CCA context)     | <b>0.781</b> | <b>0.794</b> |

Table 6: Ablation study on FLORES-200 Farsi → En (spBLEU).

| Configuration                     | spBLEU      |
|-----------------------------------|-------------|
| Full model                        | <b>34.8</b> |
| − CCA adapter                     | 33.1 (−1.7) |
| − language-family routing         | 32.4 (−2.4) |
| − SAT (standard BPE)              | 31.7 (−3.1) |
| − morpheme-boundary BPE           | 33.8 (−1.0) |
| − domain fine-tuning              | 32.9 (−1.9) |
| − QE joint training               | 34.2 (−0.6) |
| Bilingual model (no multilingual) | 29.4 (−5.4) |

enization provides substantial gains over standard BPE. Language-family routing (−2.4) shows that expert specialization beyond random initialization is important for quality.

## 11 Few-Shot Language Adaptation

A key use case for Zen-Translator is rapid adaptation to new language pairs with minimal parallel data. We evaluate the model’s few-shot translation capability by fine-tuning with  $N$  parallel sentence pairs and measuring performance after 1,000 gradient steps.

Even with 100 parallel pairs, the model reaches 11.4 spBLEU, approaching the performance of NLLB-3.3B trained on that language. This demonstrates that Zen-Translator’s multilingual representations provide strong inductive bias for related languages.

Table 7: Few-shot adaptation on Balochi→English (an unseen language at pre-training time). spBLEU after fine-tuning with  $N$  examples.

| N             | Zero-shot | After fine-tuning |
|---------------|-----------|-------------------|
| 0 (zero-shot) | 7.2       | —                 |
| 100           | —         | 11.4              |
| 500           | —         | 16.8              |
| 1,000         | —         | 21.3              |
| 5,000         | —         | 26.1              |
| 10,000        | —         | 29.4              |

## 12 Safety and Ethics

### 12.1 Bias Audit

We conduct a systematic bias audit examining whether Zen-Translator amplifies harmful cultural stereotypes. We test 2,400 sentence pairs spanning gender, religion, and political content. Key findings:

- Gender pronoun assignment: 3.2% error rate in Farsi (a null-subject language), compared to 8.1% for NLLB.
- Religious term translation: 94.3% neutral rendering of Islamic terms that are sometimes rendered pejoratively by competing systems.
- Political content: Model maintains neutrality and does not introduce politically charged terminology not present in source.

### 12.2 Harmful Content Refusal

Zen-Translator includes a toxicity classifier trained on multilingual harmful content that filters outputs containing hate speech, threats, or incitement in any supported language. The classifier achieves 0.94 F1 on a balanced test set of 10,000 examples per language.

### 12.3 Privacy Considerations

Translations may contain personally identifiable information (PII). We implement PII detection and optional pseudonymization using a Named Entity Recognition model operating before translation, with an option to substitute canonical

placeholders for names, addresses, and identification numbers.

## 12.4 Access and Consent

All training data was obtained from sources with appropriate licensing or under fair use provisions for research purposes. We engaged native speaker communities for Pashto, Kurdish, and Dari data collection and paid above-market rates to data contributors.

## 13 Limitations

Despite strong performance improvements, Zen-Translator has several important limitations:

1. **Very low-resource languages:** Languages with fewer than 1 million tokens of monolingual text (e.g., minority Kurdish dialects such as Zazaki and Gorani) remain beyond the model’s reliable capability.
2. **Dialectal Arabic:** While the model handles Modern Standard Arabic well, dialectal variants (Egyptian, Levantine, Maghrebi) show significant performance drops of 8-15 spBLEU.
3. **Spoken language transcription:** The model processes written text only; spoken language includes phonological features not captured.
4. **Long documents:** Context is limited to 512 tokens; long document translation requires chunking which can lose cross-sentence coherence.
5. **Domain coverage:** The FES domain adaptation was validated on three domains; performance on highly specialized domains (e.g., astrophysics, cryptocurrency) is not characterized.

## 14 Conclusion

We presented Zen-Translator, a neural machine translation system specialized for low-resource Middle Eastern and Central Asian languages. Our contributions include script-aware tokenization that respects morphological boundaries, language-family-aware expert routing, a cultural context

adapter that improves translation of culturally specific content, and a bidirectional quality estimation head for reference-free evaluation.

Zen-Translator achieves state-of-the-art performance on FLORES-200 for 23 of 32 evaluated language directions and substantially outperforms all baselines on our Farsi Evaluation Suite. The model demonstrates strong few-shot adaptation capability, reaching useful translation quality with as few as 100 parallel sentence pairs for new languages. We release model weights and evaluation resources to support continued research on under-served language communities.

## References

- Artetxe, M. and Schwenk, H. (2019). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7, 597-610.
- Conneau, A., Khandelwal, K., Goyal, N., et al. (2020). Unsupervised cross-lingual representation learning at scale. *Proceedings of ACL 2020*, 8440-8451.
- Costa-jussà, M.R., et al. (2022). No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.
- Fan, A., Bhosale, S., Schwenk, H., et al. (2021). Beyond English-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107), 1-48.
- NLLB Team. (2022). FLORES-200: A multilingual evaluation benchmark for NMT. *arXiv preprint arXiv:2207.04672*.
- Habash, N. (2010). Introduction to Arabic natural language processing. *Synthesis Lectures on Human Language Technologies*, Morgan & Claypool.
- Hassanpour, A. (2012). The language politics of Kurdish. *Language Ideologies, Policies and Practices*, Palgrave Macmillan.
- Houlsby, N., Giurgiu, A., Jastrzebski, S., et al. (2019). Parameter-efficient transfer learning for NLP. *Proceedings of ICML 2019*.
- Johnson, M., Schuster, M., Le, Q.V., et al. (2017). Google’s multilingual neural machine translation system. *Transactions of the ACL*, 5, 339-351.
- Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2016). Bag of tricks for efficient text classification. *Proceedings of EACL 2017*.
- Kudugunta, S., Huang, Y., Bapna, A., et al. (2021). Beyond distillation: Task-level mixture-of-experts for efficient inference. *Findings of EMNLP 2021*.
- Lewis, M., Liu, Y., Goyal, N., et al. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation. *Proceedings of ACL 2020*, 7871-7880.
- Liu, Y., Gu, J., Goyal, N., et al. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the ACL*, 8, 726-742.
- Pfeiffer, J., Kamath, A., Rücklé, A., Cho, K., and Gurevych, I. (2020). AdapterFusion: Non-destructive task composition for transfer learning. *Proceedings of EACL 2021*.
- Rei, R., Stewart, C., Farinha, A.C., and Lavie, A. (2020). COMET: A neural framework for MT evaluation. *Proceedings of EMNLP 2020*, 2685-2702.
- Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. *Proceedings of ACL 2016*, 1715-1725.
- Shazeer, N., Mirhoseini, A., Maziarz, K., et al. (2017). Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *Proceedings of ICLR 2017*.
- Specia, L., Blain, F., Fonseca, E., et al. (2021). Findings of the WMT 2021 shared task on quality estimation. *Proceedings of WMT 2021*, 684-725.

- Tang, Y., Tran, C., Li, X., et al. (2021). Multilingual translation from denoising pre-training. *Findings of ACL 2021*.
- Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
- Wu, S. and Dredze, M. (2020). Are all languages created (equally) easy to language-model? *Proceedings of the 1st Workshop on Multilingual Representation Learning*.