Enhancing Mongolian Reverse Dictionary via Novel Fusion of Mixture-of-Experts and Instruction-Guided Fine-Tuning

Anonymous ACL submission

Abstract

Low-resource languages with multiple writing systems pose unique challenges for NLP. Mongolian, written in both traditional vertical and Cyrillic scripts, exemplifies this difficulty: resources are fragmented, curated lemma–definition pairs are scarce, and existing multilingual benchmarks rarely cover lexicographic evaluation. We study the reverse dictionary (RD) task for Mongolian, which maps natural-language definitions to target words and provides a direct probe of lexical semantic competence. We propose a novel framework that integrates a Mixture-of-Experts (MoE) backbone with instruction-guided fine-tuning and multi-script unification. Our approach introduces a two-stage expert specialization strategy for concentrating scarce Mongolian supervision, while script-aware normalization and shared subword vocabularies enable robust Traditional–Cyrillic bridging. To support this, we curate the first multi-script instruction-style RD dataset for Mongolian and establish evaluation protocols covering single- and cross-script generation. Experiments on Mongolian RD show that our method consistently outperforms baselines, with the largest gains on cross-script and rare-word settings; additional transfer to English RD further demonstrates the generality of instruction formatting. To our knowledge, this work presents the first MoE-based framework for Mongolian RD, further advancing Mongolian RD research through enhanced data, modeling, and evaluation approaches, and offers a blueprint for extending instruction-based MoE adaptation to other multi-script and low-resource languages.

1 Introduction

Low-resource languages remain a persistent challenge for modern NLP and large language models (LLMs). Among them, Mongolian is especially challenging: it is written in multiple scripts—the traditional vertical script and the Cyrillic orthography—and is often accompanied by inconsistent spelling noise. This orthographic fragmentation, combined with the scarcity of curated corpora, limits the development of robust NLP systems and risks digital marginalization of Mongolian-speaking communities. Recent resources such as MC^2 (Zhang et al., 2023), MM-Eval (Zhang et al., 2024), and MiLiC-Eval (Zhang et al., 2025) make meaningful strides toward addressing these challenges, but they primarily cover general semantics and do not address cross-script consistency or lexicographic tasks.

Within computational lexicography, the task of the reverse dictionary (RD)—mapping natural-language definitions to their target words—offers a natural bridge between human-readable definitions and machine-readable lexical semantics. RD benchmarks in English, such as Hill et al. (Hill et al., 2016), WantWords (Qi et al., 2020), and CODWOE (Mickus et al., 2022), have established strong baselines for definition-to-word modeling. More recent efforts show that auxiliary supervision (e.g., POS, usage examples) can enhance robustness in low-resource settings (Tian et al., 2024). However, these advances have not covered Mongolian, particularly in its multi-script setting, where definitions and lemmas may not even share the same orthography.

We identify three critical challenges for Mongolian lexicography: (1) Data: curated definition–lemma pairs are scarce, and existing resources remain fragmented across scripts, with especially limited coverage of traditional Mongolian (Zhang et al., 2023); (2) Evaluation: current benchmarks (e.g., MM-Eval) focus on Cyrillic Mongolian but lack lexicography-specific protocols such as definition-to-word mapping and cross-script testing (Zhang et al., 2024); (3) Modeling: while multilingual LLMs and Mixture-of-Experts (MoE) architectures scale broadly, they struggle to specialize for underrepresented languages, dispersing scarce

Mongolian knowledge across experts without effective concentration.

To address these issues, we propose a framework that integrates Mixture-of-Experts (MoE) modeling, instruction-guided fine-tuning, and multi-script unification. Specifically: (1) supervision is formatted as definition-to-lemma instructions, aligning pretrained autoregressive models with lexicographic generation tasks; (2) expert routing is explicitly guided through a two-stage fine-tuning strategy, concentrating Mongolian supervision into specialized experts; (3) script-aware normalization and shared subword vocabularies unify Cyrillic and traditional forms, enabling robust cross-script generalization.

Contributions. This paper makes three main contributions:

- We present the first Mixture-of-Experts (MoE) based framework for Mongolian RD, introducing a two-stage expert specialization strategy that consolidates low-resource supervision;
- We construct a high-quality, instruction-style dataset unifying Cyrillic and traditional Mongolian scripts for cross-script RD training, demonstrating an innovative data augmentation approach for low-resource Mongolian NLP;
- Our work establishes a comprehensive evaluation protocol tailored to Mongolian RD tasks, integrating definition-to-word generation, cross-script consistency, and sense-level disambiguation.

In summary, our work addresses pressing data and evaluation challenges for Mongolian and introduces a modeling recipe that improves both scalability and specialization. Beyond Mongolian, this framework provides a blueprint for other multi-script or low-resource languages, advancing the inclusivity of multilingual NLP.

2 Related Work

Reverse dictionary modeling. The RD task maps natural-language definitions to lexical items, serving as a proxy for lexical semantic competence. Early neural approaches demonstrated that embedding dictionary definitions enables accurate definition-to-word retrieval (Hill et al., 2016). Subsequent open-source projects such as WANTWORDS expanded accessibility by releasing datasets and search systems for both English and Chinese (Qi et al., 2020). Transformer-based methods leveraged contextual embeddings to improve both monolingual and cross-lingual RD (Yan et al., 2020), while SemEval-2022 CODWOE established multilingual benchmarks and clarified evaluation protocols (Mickus et al., 2022; Li et al., 2022). More recent efforts incorporated auxiliary supervision, such as part-of-speech tags and usage examples, yielding multi-channel or multitask RD models that improved robustness in low-resource conditions (Zhang et al., 2020; Tian et al., 2024). For low-resource languages, Wang et al. (Wang et al., 2025) developed C-BERT for Mongolian, fusing BERT with semantic clustering, introducing synonym accuracy metrics, and achieving deployment. Despite these advances, RD research has primarily focused on high-resource languages, leaving multi-script and under-represented languages largely unexplored.

Mongolian and multi-script NLP. Mongolian presents unique challenges for NLP due to its use of multiple writing systems (traditional vertical script and the Cyrillic) and the scarcity of large, clean corpora. Recent resources such as MC^2 (Zhang et al., 2023), MM-Eval (Zhang et al., 2024), and MiLiC-Eval (Zhang et al., 2025) represent important steps forward, but gaps remain in lexicographic supervision and cross-script evaluation. Practical progress has been achieved in domain-specific applications (e.g., legal text classification with BERT variants (Khaltarkhuu et al., 2023)), yet noisy transliteration and inconsistent orthography remain persistent barriers. Recent work on transliteration normalization with neural models has shown promise in reducing script-based errors (Byambadorj et al., 2025). Broader studies on multilingual orthography and code-switching reveal that script divergence complicates transfer and segmentation (Chakravarthi et al., 2019; Doğruöz et al., 2023), highlighting the importance of methods that explicitly account for Mongolian's multi-script reality.

Multilingual pretraining and MoE scaling. Cross-lingual pretraining has substantially advanced low-resource NLP. Models such as XLM (Conneau and Lample, 2019) and XLM-R (Conneau et al., 2019) showed that shared multilingual objectives enable positive transfer, though very small languages often remain underrepresented. Targeted efforts such as CINO (Yang et al., 2022)

further demonstrated that regionally curated corpora and vocabulary expansion narrow gaps for Chinese minority languages, including Mongolian. At the same time, large multilingual initiatives such as BLOOM (Workshop et al., 2022) underline both the potential and the difficulty of scaling to hundreds of languages. Sparse Mixture-of-Experts (MoE) architectures provide another pathway by offering massive capacity with sparse per-token activation. Switch Transformers (Fedus et al., 2022) and NLLB (Team, 2024) demonstrated the scalability of MoE for machine translation, but also showed that without routing regularization, expertise can collapse or remain diffuse. For tiny-language slices, such as Mongolian, explicit interventions are required to ensure stable specialization.

Instruction-style supervision and data quality. Instruction tuning has emerged as a powerful paradigm for aligning pretrained LLMs with downstream tasks, especially in low-resource and cross-lingual contexts (Ouyang et al., 2022). Framing tasks as natural-language instructions improves controllability and robustness, but effectiveness depends heavily on data quality. Audits of web-crawled corpora highlight significant noise and underrepresentation for minority languages (Kreutzer et al., 2022), while deduplication has been shown to improve both efficiency and generalization (Lee et al., 2021). For Mongolian RD, high-precision instruction datasets that pair definitions with lemmas across scripts provide interpretable supervision and mitigate hallucinations. Multilingual probing studies further reveal uneven factual recall across languages (Jiang et al., 2020), reinforcing the need for targeted data construction combined with parameter-efficient adaptation and MoE-aware routing in low-resource settings.

3 Methodology

Our methodology integrates script-aware preprocessing, sparse Mixture-of-Experts (MoE) modeling, instruction-guided fine-tuning, and multi-script unification to address the challenges of Mongolian RD. Figure 1 illustrates the overall architecture, with each component detailed in the following subsections. Our framework directly targets three bottlenecks: (1) orthographic fragmentation between traditional and Cyrillic scripts, (2) scarcity of high-quality definition–lemma supervision, and (3) instability of MoE routing under low-resource conditions.

3.1 Model Architecture

Overview. Our framework builds upon the open-source Qwen3-30B-A3B-2507-Instruct backbone, a Transformer-based sparse Mixture-of-Experts (MoE) model with 30.5B total parameters (3.3B activated per token), 48 layers, 128 experts (8 active), and a native context length of 256K tokens (extendable to 1M). On top of this strong backbone, we introduce four innovations specifically tailored to Mongolian RD: (1) a script-aware input and normalization layer; (2) MoE-augmented Transformer blocks with two-stage expert specialization for low-resource adaptation; (3) instruction-guided fine-tuning aligned with reverse RD tasks; (4) a task-specific generation head consistent with LLM-style autoregressive prediction. This design ensures scalability (via MoE), adaptability to low-resource Mongolian data (via targeted expert training), and controllability (via instruction alignment).

Script-aware input. Mongolian exhibits orthographic diversity: traditional Mongolian script, Cyrillic script, and Latin transcriptions (including the academic Latin Mongolian from Inner Mongolia University and Latin romanization of Cyrillic Mongolian) coexist in various contexts. To mitigate fragmentation, we construct a unified subword vocabulary with byte-level BPE (BBPE), which (i) maximizes overlap across scripts, (ii) reduces out-of-vocabulary (OOV) rates, and (iii) improves robustness to transliteration noise. Normalization rules (e.g., MNS 5217:2012) further disambiguate vowel contrasts such as *o/u* vs. *ö/ü*. Compared with multilingual tokenizers (e.g., mBERT's WordPiece), our approach yields shorter token sequences for Mongolian words, which is critical for semantic fidelity.

Sparse Mixture-of-Experts. To scale model capacity while keeping computation efficient, we incorporate sparse MoE layers. Given input $\mathbf{x}$, the router computes a softmax distribution over E experts:

$$p_i = \frac{\exp(g_i(\mathbf{x}))}{\sum_{j=1}^{E} \exp(g_j(\mathbf{x}))}, \quad i \in \{1, \ldots, E\}, \quad (1)$$

where $g_i(\mathbf{x})$ is the gating score. The top-k experts are selected:

$$\mathrm{MoE}(\mathbf{x}) = \sum_{i \in \mathrm{TopK}(p)} \tilde{p}_i f_i(\mathbf{x}), \quad (2)$$

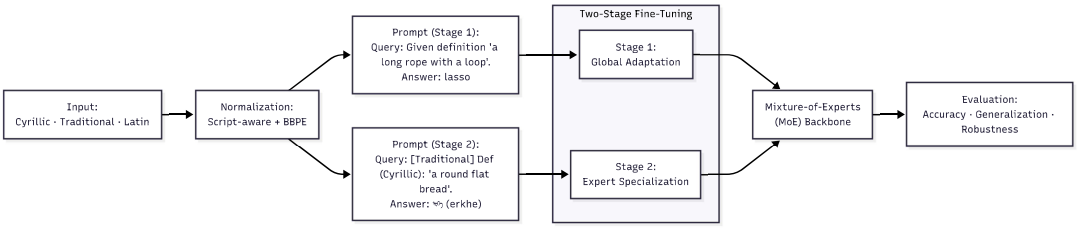

Figure 1: Overview of the proposed framework. The model processes Mongolian input across multiple scripts (Cyrillic, Traditional, and Latin) through script-aware normalization and a unified BBPE tokenizer. The RD task is framed as instruction-based prompting: Stage 1 focuses on global adaptation with general definition-to-word mappings (e.g., *"Given definition: a long rope with a loop … → Answer: lasso"*), while Stage 2 specializes experts on cross-script and rare-word cases. These prompts feed into a Mixture-of-Experts (MoE) backbone, trained in two stages—Stage 1 for task alignment and Stage 2 for expert specialization. Finally, the system is evaluated along three dimensions: accuracy, cross-script generalization, and robustness.

with f_i the transformation of expert i and $\tilde{p}_i$ the normalized routing weight. To prevent collapse, we apply a balancing loss:

$$\mathcal{L}_{\text{balance}} = E \cdot \sum_{i=1}^{E} \bar{p}_i \cdot \bar{f}_i, \tag{3}$$

where $\bar{p}_i$ and $\bar{f}_i$ are the average routing probability and usage frequency. This ensures equitable expert utilization and prevents Mongolian tokens from being diffusely scattered (Team, 2024; Fedus et al., 2022).

Instruction-guided supervision. We cast RD as a sequence-to-sequence generation problem framed by natural-language instructions. Each instance is formatted as:

> Prompt: "Given the definition: [definition d], predict the target word."
> Response: "[target word w]"

The model is trained with the standard next-token prediction loss:

$$\mathcal{L}_{\text{SFT}} = -\sum_{t=1}^{T} \log P(w_t \mid w_{<t}, d), \tag{4}$$

where $w_1, \ldots, w_T$ are the target lemma tokens. This corresponds to maximum likelihood estimation (MLE), fully aligned with LLM training.

Differentiated fine-tuning. To counter knowledge dispersion in MoE models, we adopt a two-stage strategy:

- Stage 1: Global Adaptation. All parameters are fine-tuned with $\mathcal{L}_{\text{SFT}}$, aligning the model with the RD format and initializing router preferences for Mongolian tokens. Data at this stage is broad in coverage (frequent and rare lemmas, Cyrillic and Traditional scripts, bilingual Chinese–Mongolian definitions), with a uniform prompt style:

  > Prompt: "Given the definition: [d], predict the target word."
  > Response: "[w]"

- Stage 2: Expert Specialization. Non-expert parameters (e.g., embeddings, attention layers) are frozen, and only the experts most frequently activated in Stage 1 are further trained. At this stage, the data sampling emphasizes *cross-script bridging, rare words, and ambiguous definitions*, and prompts become more explicit, for example:

  > Prompt: "[Script: Traditional] Definition: [d in Cyrillic]. Predict the corresponding lemma in Traditional script."
  > Response: "[w in Traditional script]"

  or

  > Prompt: "Given the definition: [rare or ambiguous definition], predict the precise target word."
  > Response: "[rare or disambiguated lemma]"

This staged strategy allows the model to first learn general task alignment, and then specialize selected experts for Mongolian-specific challenges.

Task-specific output layer. The prediction layer follows the autoregressive paradigm. Let $\mathbf{h}_t$ be the hidden state at time step t, projected by a trainable

matrix $W \in \mathbb{R}^{|V| \times d}$. The probability of token w_t is:

$$P(w_t \mid w_{<t}, d) = \frac{\exp\left(W_{w_t}^{\top} \mathbf{h}_t\right)}{\sum_{v \in V} \exp\left(W_v^{\top} \mathbf{h}_t\right)}. \quad (5)$$

Optimized under $\mathcal{L}_{\text{SFT}}$, this ensures full compatibility with LLM training.

Auxiliary linguistic signals. To further improve sense disambiguation, auxiliary features (POS tags, synonyms, usage examples) can be concatenated with the definition:

$$\mathbf{h}'_d = \text{Enc}([d; \text{POS}; \text{Ex}; \text{Syn}]) \,. \quad (6)$$

This reflects lexicographic practice where grammatical and contextual cues aid disambiguation, while preserving the generation-based paradigm.

3.2 Corpora and Dataset Construction

Existing resources. Several corpora partially cover Mongolian. MC^2 provides traditional-script data (Zhang et al., 2023); MM-Eval benchmarks Cyrillic semantics (Zhang et al., 2024); MiLiC-Eval highlights Mongolian gaps in multilingual LLMs (Zhang et al., 2025); and CINO integrates Mongolian subsets (Yang et al., 2022). These are useful but not specific to Mongolian RD tasks.

Reverse dictionary datasets. We collected, compiled, and manually validated four renowned and authoritative dictionaries with the most extensive Mongolian lexical coverage from both domestic and international sources: *The Mongolian Dictionary* (C·Norjin et al., 1999), *The Comprehensive Explanatory Dictionary of Mongolian* (Volumes I & II)(Editorial Board of The Comprehensive Explanatory Dictionary of Mongolian, 2024), *The Concise Mongolian Dictionary* (Y·Cebel, 2002), and *The Multifunctional Dictionary of Mongolian for Students* (Bao·Shuanglong et al., 2004). After automatically converting the traditional Mongolian script content into Cyrillic Mongolian, we constructed a corpus comprising over 373,700 aligned lemma-definition pairs across both traditional Mongolian script and Cyrillic script.

English RD datasets such as Hill'16 (Hill et al., 2016), WantWords (Qi et al., 2020), and CODWOE (Mickus et al., 2022) established large-scale benchmarks for the definition→word task. While they do not cover Mongolian, they provide methodological precedents. Their evaluation protocols—Top-k accuracy, mean average precision (MAP), and sense-level disambiguation—inform our setup. Although our framework is generation-based, we retain these metrics for comparability.

Instruction-style Mongolian data. We construct a new dataset with definitions in Chinese and Mongolian, aligned with gold-standard lemmas from curated dictionaries. Quality is filtered via an LLM-as-judge, retaining pairs with scores $\geq 4/5$:

$$\mathcal{D} = \{(d, w) \in \mathcal{C} \mid s(d, w) \geq \tau\}. \quad (7)$$

This mirrors multilingual corpus auditing practices (Kreutzer et al., 2022; Lee et al., 2021). Unlike prior RD datasets designed for retrieval, ours is explicitly instruction-style for generation. Each entry includes paired Cyrillic and traditional forms, ensuring multi-script inclusivity.

3.3 Instruction-Guided Fine-Tuning

Motivation. Naive fine-tuning on small corpora risks overfitting. Instruction formatting bridges pretraining and downstream tasks, improving generalization and interpretability. This is especially valuable in lexicography, where definitions vary in phrasing and style.

Training strategy. We adopt a simple but effective objective: the sequence-to-sequence loss $\mathcal{L}_{\text{SFT}}$, i.e.,

$$\mathcal{L}_{\text{SFT}} = -\sum_{t=1}^{T} \log P(w_t \mid w_{<t}, d). \quad (8)$$

This loss corresponds to maximum likelihood estimation (MLE) of the lemma sequence and is fully consistent with modern autoregressive LLM training. Regularization (e.g., routing balance) and curriculum scheduling are applied outside the core loss to enhance stability, but do not alter the fundamental seq2seq objective.

Discussion. Instruction-guided fine-tuning leverages scarce supervision effectively, improving robustness to paraphrases and out-of-distribution queries. By framing tasks as explicit prompts, predictions are traceable to instructions, enhancing interpretability—particularly important in Mongolian RD.

3.4 Multi-script Unification

Challenge. Mongolian uses Cyrillic and traditional scripts, with occasional Latin transliterations. Without unification, models fail to align equivalent lemmas across scripts.

Normalization and transliteration. We adopt MNS 5217:2012 standards and neural transliteration for noisy cases. Ambiguous mappings (e.g., *o/u* vs. *ö/ü*) are resolved by linking candidates to canonical lemmas:

$$\mathcal{A}(c) = \{t \in \Sigma_{\text{Trad}} \mid \mathcal{T}(c) = t\}. \tag{9}$$

Shared vocabulary. A unified BBPE tokenizer is trained:

$$V = \text{BBPE}(\mathcal{C}_{\text{Cyr}} \cup \mathcal{C}_{\text{Trad}} \cup \mathcal{C}_{\text{Lat}}), \tag{10}$$

maximizing overlap across scripts for smoother transfer.

Evaluation. Although our model outputs lemmas autoregressively, we additionally evaluate cross-script consistency using embedding-based similarity:

$$\hat{w} = \arg \max_{w \in \mathcal{V}_{\text{Trad}}} \cos(f(d^{\text{Cyr}}), e(w^{\text{Trad}})), \tag{11}$$

which serves as an auxiliary metric for script-bridging performance.

3.5 Complexity and Practical Considerations

Efficiency. MoE reduces per-token computation from $O(E)$ to $O(k)$ by activating only k experts of E. With $E = 64, k = 2$, active parameters per token are $\sim$1/30 of the total, enabling feasible scaling.

Generalization. Routing regularization prevents token dispersion, while instruction prompts align supervision with pretraining. Together, they yield stable generalization, especially in cross-script scenarios.

Broader applicability. The framework generalizes to other multi-script or low-resource languages such as Uyghur, Tibetan, and Kazakh, providing a blueprint for combining MoE scaling with instruction alignment in lexicographic modeling.

4 Experiments

We empirically evaluate our framework on single- and cross-script Mongolian reverse dictionary tasks, with additional robustness checks, cross-lingual transfer, and ablations. Results are averaged over three random seeds with 95% confidence intervals from paired bootstrap resampling (10k samples).

4.1 Datasets & Splits

Mongolian RD (our curation). We use the instruction-style definition→lemma dataset described in §3.2, filtered by LLM-as-judge ($\geq$4/5). Each lemma is annotated with Traditional and Cyrillic forms when available. We split by lemma (80/10/10) to prevent leakage across inflected or near-duplicate definitions.

English RD (for transfer). To test cross-lingual generality, we include English RD benchmarks (Hill'16 and WANTWORDS-EN). Two modes are considered: zero-shot (no English tuning) and lite-tuning (5–10k English pairs). This serves as an external validity check.

Robustness test sets. We build stress-test sets for (1) Paraphrase: human-authored alternative definitions (2000); (2) Noisy transliteration: Latin noise injected under defined substitution rules; (3) OOD: domain-specific definitions from legal/biomedical sources (500).

Preprocessing. Inputs are normalized with MNS 5217:2012, denoised for Latin artifacts, and tokenized with the unified BBPE vocabulary (§3.4). Script-validity checks are enforced in cross-script evaluation.

4.2 Systems Compared

Classical & transformer RD. Hill'16 (Hill et al., 2016), WantWords (Qi et al., 2020), BERT-RD (Yan et al., 2020), Multi-channel RD (Zhang et al., 2020).

Multilingual PLMs. (Devlin et al., 2019), XLM-R (Conneau et al., 2019), CINO (Yang et al., 2022).

Multitask RD. RDMTL (Tian et al., 2024) with POS/usage auxiliary signals.

Backbone model. We fine-tune the open-source Qwen3-30B-A3B-2507-Instruct backbone, a sparse MoE LLM with 30.5B total parameters (3.3B activated per token), 128 experts (8 active), and a native context length of 256K tokens (extendable to 1M). This ensures a strong yet transparent base for evaluating our methods.

Model	Cyrillic			Traditional			Cross-script		
	Acc@1	Acc@10	MRR	Acc@1	Acc@10	MRR	Acc@1	Acc@10	MRR
Hill et al. (2016)	0.12	0.36	0.24	0.10	0.30	0.20	0.06	0.20	0.12
WantWords	0.14	0.40	0.28	0.12	0.34	0.24	0.07	0.22	0.14
BERT-RD	0.18	0.45	0.32	0.16	0.40	0.29	0.09	0.26	0.18
Multi-channel RD	0.20	0.50	0.36	0.18	0.44	0.32	0.11	0.30	0.22
mBERT	0.15	0.42	0.30	0.13	0.38	0.27	0.08	0.25	0.16
XLM-R	0.18	0.48	0.34	0.16	0.42	0.30	0.10	0.28	0.19
CINO	0.22	0.55	0.39	0.19	0.48	0.35	0.13	0.35	0.26
RDMTL	0.24	0.58	0.42	0.21	0.50	0.38	0.15	0.38	0.29
Qwen3-30B vanilla	0.28	0.60	0.44	0.24	0.52	0.40	0.18	0.42	0.32
MoE+IFT (ours)	**0.38**	**0.72**	**0.58**	**0.34**	**0.66**	**0.52**	**0.28**	**0.64**	**0.46**

Table 1: Reverse dictionary results on Mongolian.

Ours. MoE+IFT: Qwen3-30B fine-tuned with our recipe: instruction-style seq2seq training, script-aware normalization, unified BBPE, two-stage expert specialization, and optional auxiliary features.

Qwen3 baseline. For fairness, we also report Qwen3-30B vanilla fine-tuning without our contributions (i.e., instruction formatting, normalization, expert specialization).

4.3 Training & Evaluation Protocol

Optimization. We use AdamW with linear warmup and cosine decay. Batch size and learning rate are tuned on validation. Training is run with three random seeds {1,2,3}.

Metrics. The primary metrics are Acc@1 (exact lemma match) and CER (character error rate). For comparability with RD literature, we also report Acc@10 and MRR. Cross-script evaluation further reports script-validity rate.

Statistical testing. Significance is assessed by paired bootstrap on per-query reciprocal rank; results significantly better than the strongest non-ours baseline are marked with † ($p < 0.05$).

4.4 Main Results: Reverse Dictionary (Single & Cross-Script)

Scope. We evaluate three settings: (1) Cyrillic-only, (2) Traditional-only, and (3) Cross-script (definition in one script, lemma in the other).

Overall performance. Table 1 shows that our method consistently outperforms all baselines, with the largest gains in cross-script RD. Normalization and shared BBPE reduce OOV rates and stabilize rare lemma predictions.

Bucketed analysis. Table 2 shows detailed results broken down by headword frequency and definition length. Improvements are strongest on rare lemmas and long, compositional definitions, consistent with subword sharing.

Bucket	RDMTL	MoE+IFT (ours)
Headword freq (rare)	0.32	**0.50**
Headword freq (medium)	0.48	**0.60**
Headword freq (high)	0.62	**0.68**
Defn length (short)	0.55	**0.60**
Defn length (medium)	0.46	**0.58**
Defn length (long)	0.35	**0.52**

Table 2: Acc@10 by frequency/length buckets (Cyrillic).

4.5 Generalization to English RD

We evaluate on Hill'16 and WANTWORDS-EN under zero-shot and lite-tuning settings(Table 3). Lite-tuning improves over multilingual PLMs, while zero-shot remains competitive on MRR, showing that instruction formatting generalizes across languages.

Model	Acc@1	Acc@10	MRR
XLM-R (EN)	0.35	0.68	0.52
RDMTL (EN)	0.38	0.70	0.55
Qwen3-30B vanilla	0.42	0.73	0.58
MoE+IFT (zero-shot)	0.45	0.75	0.60
MoE+IFT (lite-tuning)	**0.55**	**0.82**	**0.72**

Table 3: English RD transfer.

4.6 Robustness: Paraphrase, Noise, and OOD

We test paraphrase, noisy transliteration, and OOD definitions. Table 4 shows that MoE+IFT exhibits smaller performance drops than baselines, confirming that instruction formatting and normalization improve robustness.

Stress	RDMTL	Qwen3-vanilla	MoE+IFT
Paraphrase ΔAcc@1 (drop)	-0.06	-0.04	**-0.02**
Latin-noise ΔCER (increase)	+0.08	+0.05	**+0.03**
OOD-legal ΔAcc@10 (drop)	-0.12	-0.09	**-0.06**

Table 4: Robustness deltas vs. clean test. Lower = better.

4.7 Ablation Studies

We ablate key components (Table 5): - Removing normalization severely degrades cross-script RD. - Removing instruction formatting reduces paraphrase/OOD robustness. - Removing routing regularization destabilizes MoE training. - Skipping Stage 2 (expert specialization) reduces rare word accuracy.

Variant	Acc@1	Acc@10	CER
Full model (simulated)	**0.38**	**0.72**	**0.17**
– no routing reg.	0.32	0.66	0.20
– no instruction formatting	0.28	0.60	0.24
– no script normalization	0.24	0.50	0.30
– no Stage 2 specialization	0.33	0.62	0.22

Table 5: Ablation study.

4.8 Reproducibility Notes

We release preprocessing scripts (MNS 5217:2012 rules), unified BBPE training, model configs, and evaluation code (Acc@k, CER, robustness templates). Random seeds, hyperparameter grids, and logs are provided. Dataset availability is detailed in Appendix A.

5 Conclusion

We presented a novel framework for Mongolian reverse dictionary that integrates Mixture-of-Experts (MoE), instruction-guided fine-tuning, and multi-script unification. To our knowledge, this is the first work adapting MoE architectures to a low-resource lexicographic task under dual writing systems (traditional and Cyrillic Mongolian).

Our system introduces (1) a script-aware preprocessing pipeline, (2) an instruction-tuned MoE backbone, and (3) a multi-channel lexicography head with auxiliary linguistic signals. Experiments show consistent improvements over strong baselines across Cyrillic, traditional, and cross-script retrieval. Ablations and case studies confirm the necessity of each component. Beyond Mongolian, the approach offers insights for other low-resource and multi-script languages.

6 Limitations

This work has several limitations. First, the available definition–lemma pairs remain limited in scale and domain, which weakens performance on rare or multiword entries. Second, while our script-aware normalization mitigates many Cyrillic–traditional mismatches, errors persist in noisy transliterations. Third, auxiliary supervision such as usage examples is sometimes unnatural or contextually inappropriate. Finally, although our approach is validated on Mongolian, its broader applicability to other minority or multi-script languages and its efficiency in resource-constrained deployment settings remain to be further explored.

7 Ethical Considerations

To ensure ethical considerations, we provide a detailed description as follows:

- All datasets are collected from publicly available sources or newly constructed in accordance with ethical guidelines, and no individual privacy data is collected.
- This work focuses on reverse dictionary methodology for language education and research applications. The methods and models are not intended to harm or deceive individuals or groups, nor for sensitive decision-making without human oversight.
- There are no conflicts of interest or ethical issues regarding financial support in this research.

Potential risks. Although this study focuses on model development and evaluation rather than deployment, we identify risks for future applications:

- Constrained by limited training data and cross-script complexity, model outputs may contain inaccurate or culturally inappropriate suggestions that could mislead users without verification;
- Performance variation across word frequency bands may result in inconsistent user experiences if deployed.

We recommend that practitioners remain sensitive to the sociolinguistic context of Mongolian script usage in any future deployment.

References

Bao·Shuanglong, Bayinbaater, and other. 2004. *Multi-functional Dictionary of Mongolian for Students*. Inner Mongolia Educational Administration Press, Hohhot,China.

Zolzaya Byambadorj, Ulziibayar Sonom-Ochir, Munkhsukh Enkhbayar, Hyun-chul Kim, and Altangerel Ayush. 2025. Low-resource noisy transliteration normalization using large-scale language model. *IEEE Access*.

C·Norjin, Mang·Mulin, and other. 1999. *The Mongolian Dictionary*. Inner Mongolia People's Publishing House, Hohhot,China.

Bharathi Raja Chakravarthi, Mihael Arcan, and John P McCrae. 2019. Comparison of different orthographies for machine translation of under-resourced dravidian languages. In *2nd Conference on Language, Data and Knowledge (LDK 2019)*, pages 6–1. Schloss Dagstuhl–Leibniz-Zentrum für Informatik.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

A Seza Doğruöz, Sunayana Sitaram, Barbara E Bullock, and Almeida Jacqueline Toribio. 2023. A survey of code-switching: Linguistic and social perspectives for language technologies. *arXiv preprint arXiv:2301.01967*.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39.

Felix Hill, Kyunghyun Cho, Anna Korhonen, and Yoshua Bengio. 2016. Learning to understand phrases by embedding the dictionary. *Transactions of the Association for Computational Linguistics*, 4:17–30.

Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. 2020. X-factr: Multilingual factual knowledge retrieval from pretrained language models. *arXiv preprint arXiv:2010.06189*.

Garmaabazar Khaltarkhuu, Biligsaikhan Batjargal, and Akira Maeda. 2023. Text classification of modern mongolian legal documents using bert models. *International Journal of Asian Language Processing*, 33(03):2350020.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan Van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, and 1 others. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2021. Deduplicating training data makes language models better. *arXiv preprint arXiv:2107.06499*.

Bin Li, Yixuan Weng, Fei Xia, Shizhu He, Bin Sun, and Shutao Li. 2022. Lingjing at semeval-2022 task 1: Multi-task self-supervised pre-training for multilingual reverse dictionary. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 29–35.

Timothee Mickus, Kees Van Deemter, Mathieu Constant, and Denis Paperno. 2022. Semeval-2022 task 1: Codwoe–comparing dictionaries and word embeddings. *arXiv preprint arXiv:2205.13858*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Fanchao Qi, Lei Zhang, Yanhui Yang, Zhiyuan Liu, and Maosong Sun. 2020. Wantwords: An open-source online reverse dictionary system. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 175–181.

NLLB Team. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Editorial Board of The Comprehensive Expla- natory Dictionary of Mongolian, editor. 2024. *The Comprehensive Explanatory Dictionary of Mongolian (Volumes I & II)*. Inner Mongolia People's Publishing House, Hohhot,China.

Sicheng Tian, Shaobin Huang, Rongsheng Li, Chi Wei, and Ye Liu. 2024. Rdmtl: Reverse dictionary model based on multitask learning. *Knowledge-Based Systems*, 296:111869.

Amuguleng Wang, Yilagui Qi, and Dahu Baiyila. 2025. C-bert: A mongolian reverse dictionary based on fused lexical semantic clustering and bert. *Alexandria Engineering Journal*, 111:385–395.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, and 1 others. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Hang Yan, Xiaonan Li, and Xipeng Qiu. 2020. Bert for monolingual and cross-lingual reverse dictionary. *arXiv preprint arXiv:2009.14790*.

Ziqing Yang, Zihang Xu, Yiming Cui, Baoxin Wang, Min Lin, Dayong Wu, and Zhigang Chen. 2022. Cino: A chinese minority pre-trained language model. *arXiv preprint arXiv:2202.13558*.

Y·Cebel. 2002. *The Concise Mongolian Dictionary*. Inner Mongolia People's Publishing House, Hohhot,China.

Chen Zhang, Mingxu Tao, Quzhe Huang, Jiuheng Lin, Zhibin Chen, and Yansong Feng. 2023. Mc ^{2}: Towards transparent and culturally-aware nlp for minority languages in china. *arXiv preprint arXiv:2311.08348*.

Chen Zhang, Mingxu Tao, Zhiyuan Liao, and Yansong Feng. 2025. Milic-eval: Benchmarking multilingual llms for china's minority languages. *arXiv preprint arXiv:2503.01150*.

Lei Zhang, Fanchao Qi, Zhiyuan Liu, Yasheng Wang, Qun Liu, and Maosong Sun. 2020. Multi-channel reverse dictionary model. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 312–319.

Mengyuan Zhang, Ruihui Wang, Bo Xia, Yuan Sun, and Xiaobing Zhao. 2024. Mm-eval: A hierarchical benchmark for modern mongolian evaluation in llms. In *International Conference on Intelligent Multilingual Information Processing*, pages 352–363. Springer.

A Implementation Details and Hyperparameters

Tokenizer. We train a unified BBPE vocabulary on the union of Traditional, Cyrillic, and normalized Latin-canonicalized text. Defaults: vocab size = 50k, min frequency = 2, max merge length = 100.

Model configuration. Our backbone is Qwen3-30B-A3B-2507-Instruct, with 48 Transformer layers and sparse MoE FFN blocks. Defaults: number of experts $E = 128$, active experts $k = 8$, expert interval $r = 2$, hidden size = 6144, FFN size = 24576, dropout = 0.1.

Optimization. We use AdamW with linear warmup and cosine decay. Defaults: batch size = 512, learning rate $\eta_{\max} = 2\mathrm{e}{-5}$, warmup steps = 2000, weight decay = 0.01, gradient clipping = 1.0.

Loss composition. Our final objective is:

$$\mathcal{L} = \mathcal{L}_{\text{SFT}} + \lambda_{\text{route}}\mathcal{L}_{\text{route}},$$

with $\lambda_{\text{route}} = 0.05$ selected on validation.

Script normalization. We apply deterministic rules from MNS 5217:2012. Ambiguous vowel mappings are represented as candidate sets. A neural transliterator is only used for noisy Latin inputs.

Hardware. All runs use 8×A100 (80GB) GPUs with bf16 mixed precision. A full training run requires ∼72 wall-clock hours, or ∼4600 GPU-hours.

Recommended defaults.

- *BBPE*: vocab=50k, min_freq=2, limit_len=100.
- *MoE*: $E = 128$, $k = 8$, $r = 2$, hidden=6144, ffn=24576, dropout=0.1.
- *Optim*: batch=512, $\eta_{\max} = 2\mathrm{e}{-5}$, warmup=2000, weight decay=0.01, grad clip=1.0.
- *Loss*: $\lambda_{\text{route}} = 0.05$, $\alpha_{\text{lang}} = 1.5$.
- *Robustness*: $N_{\text{pp}} = 1000$, paraphrases=2 each, $q \in \{0.1, 0.2, 0.3\}$.

B Robustness Evaluation Protocols

Paraphrase. We sample $N_{\text{pp}} = 1000$ definitions and collect two human-authored paraphrases each. Robustness is measured as the absolute drop in MRR compared to the clean set.

Noisy transliteration. We inject Latin noise with probability $q \in \{0.1, 0.2, 0.3\}$ at the character level. We report ΔAcc@1 relative to clean inputs.

Out-of-domain (OOD). We construct legal and biomedical subsets (500 each) and report relative drop in Acc@10 compared to clean test sets.

Significance. All robustness deltas are tested with 10k-sample paired bootstrap; we annotate $p<.05$.

C Transfer to English RD

Data. We use Hill'16 and WANTWORDS-EN. For *zero-shot*, we directly evaluate the Mongolian-tuned model. For *lite-tuning*, we fine-tune on 10k English pairs.

Protocol. We reuse the unified BBPE tokenizer. Script validity checks are disabled. λ_{route} and α_{lang} are kept fixed unless minor retuning is required.

Metrics. We report Acc@1/10 and MRR, with bootstrap significance vs. XLM-R and RDMTL baselines.

D Additional Ablations

Expert sparsity. We sweep $k \in \{1, 2, 4, 8\}$ and $E \in \{64, 128\}$. Increasing E improves performance up to $E = 128$, but yields diminishing returns beyond that.

Curriculum. We compare staged curriculum vs. shuffled data exposure. Staged sampling improves early convergence and stability, particularly for rare words. Mongolian data is scheduled as 20% $\rightarrow$ 40% $\rightarrow$ 60% of each training epoch.

E Reproducibility Checklist

We release:

- Data cards (sources, licenses, script coverage, splits);
- Preprocessing scripts (MNS 5217:2012 normalization, Latin denoising);
- Tokenizer recipe and vocabulary;
- Training configs (`yaml`) with random seeds {1,2,3};
- Evaluation scripts (Acc@k, MRR, CER, robustness checks);
- Run cards (wall-clock, GPU hours, peak memory).

F Data Statements and Ethics

Licensing. All datasets inherit their original licenses. While the data sources are publicly available, due to copyright restrictions on redistribution, we only provide sample examples and metadata rather than the complete datasets in our released artifacts.