

CoPiT: Cognitive Pivot Translation for Digraphic Low-Resource Mongolian in the Traditional Script

Burte Bayarsaikhan^{*1} Serynn Kim^{*2} Buru Chang^{†1}

¹Korea University ²Hankuk University of Foreign Studies

{burtebay, buru_chang}@korea.ac.kr

serynn@hufs.ac.kr

Abstract

Low-resource languages remain challenging for machine translation, and Mongolian is a representative case. As a digraphic language, Mongolian is written in both Cyrillic and Traditional scripts, which exhibit a severe imbalance in data availability. While the Cyrillic script is relatively well-resourced, the Traditional script remains extremely data-scarce and orthographically ambiguous, leading to substantial performance degradation in direct translation. We propose *CoPiT*, a cognitively motivated pivot-based translation pipeline that exploits this internal resource hierarchy by routing translation through the Cyrillic script. The pipeline explicitly resolves script-induced ambiguity in the Traditional script before translation, enabling more stable and accurate meaning transfer. Across multiple backbone models and target languages, *CoPiT* generally improves over direct translation, with substantial absolute BLEU improvements and up to 1.5–1.6× gains in COMET. These gains allow strong open-source models to match or outperform GPT-4.1 under comparable evaluation settings. Beyond inference-time improvements, *CoPiT* enables the construction of synthetic parallel data directly from Traditional-script text, mitigating data scarcity in realistic low-resource scenarios. We release a new multi-script parallel dataset covering Mongolian in both scripts alongside English, Korean, and Russian. All datasets and code are publicly available at <https://github.com/bbyba/CoPiT.git>.

1 Introduction

Large language models (LLMs) have evolved from primarily monolingual systems to multilingual systems through large-scale multilingual pre-training. Despite this progress, LLMs continue to perform poorly on low-resource languages (LRLs), largely due to limited training data and sparse linguistic

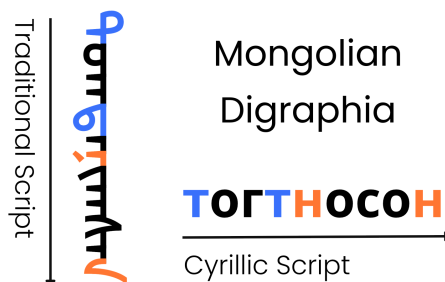


Figure 1: Mongolian digraphia. The same content can be written in Traditional script (left), with multiple surface forms and limited resources, and in better-resourced Cyrillic script (right).

resources. This limitation is particularly evident in machine translation (MT), which relies heavily on large-scale parallel corpora that are often unavailable for LRLs (Raja and Vats, 2025).

These challenges are particularly pronounced in Mongolian, a low-resource language with a unique **digraphic characteristic**, as it is written in both *Cyrillic* and *Traditional* scripts (see Figure 1). The Cyrillic script, which serves as the dominant modern form, is largely phonemic and exhibits a high degree of consistency in the sound representation. In contrast, the Traditional script, an archaic writing system, is written vertically, and its letter forms vary by positional context. Many phonological and morphological distinctions are not explicitly encoded in the script, such that a single written form may correspond to multiple plausible interpretations. This intrinsic ambiguity distinguishes the Traditional script structurally from its Cyrillic counterpart. For example, the Gemini-3-Pro-Preview API flags content written in the Traditional Mongolian script as harmful, as shown in Appendix A.1.

Beyond these structural differences, the two Mongolian scripts are associated with markedly different levels of linguistic and computational resources. The Cyrillic script has long served as the dominant writing system in modern Mongolian society following its institutional adoption through

^{*}Equal contribution.

[†]Corresponding author.

state language policies. As a result, the vast majority of textual data and language technologies are developed for the Cyrillic script. In contrast, the Traditional script, despite representing the same language, is reintroduced into official use through a policy-driven effort to promote traditional writing practices¹. Owing to this delayed revival, the Traditional script remains severely under-resourced, giving rise to a hierarchical resource imbalance within a single LRL. This imbalance leads current NLP systems to exhibit pronounced performance asymmetry between the two scripts, with poorer performance on the Traditional script.

To address this gap, we introduce a novel pipeline, *CoPiT* (*Cognitive Pivot Translation*), for translating Mongolian written in the Traditional script into other languages. Our pipeline is motivated by how fluent Mongolian readers process the Traditional script in practice (Altangerel and Togtokh, 2024). Due to the dominance of the Cyrillic script in modern usage, many native speakers do not read the Traditional script in a purely direct manner. Instead, they implicitly map Traditional script forms to their Cyrillic counterparts and interpret them through this familiar representation, resolving script-dependent ambiguities in a phonologically informed way. Accordingly, we adopt Mongolian written in the Cyrillic script as an intermediate representation. While it remains low-resource, it offers a more phonemically transparent and comparatively better-resourced representation than the Traditional script, and thus aligns more effectively with both human reading strategies and the strengths of current LLMs.

Concretely, we decompose the MT task into two sequential stages: (1) conversion from Mongolian written in the Traditional script to the Cyrillic script, followed by (2) translation from the Cyrillic script to the target language. This design isolates the script-specific challenges of the Traditional script from downstream semantic translation while leveraging the relatively greater, though still limited, resources available for the Cyrillic script.

The Traditional-to-Cyrillic conversion is further factorized into three linguistically motivated steps that align with disambiguation processes used by fluent Mongolian readers: vowel harmony recovery,

Latin-assisted normalization, and Cyrillic normalization. Collectively, these steps transform the underspecified Traditional script into a phonologically specified Cyrillic representation, thereby facilitating subsequent translation into the target language. This explicit resolution of script-level ambiguity enables *CoPiT* to translate Mongolian written in the Traditional script without relying on large-scale parallel corpora. Instead, it requires only a small amount of data for learning script conversion.

Comprehensive experimental results demonstrate that our pipeline significantly improves translation accuracy across multiple target languages, including English, Korean, and Russian. When applied to GPT-4.1, zero-shot *CoPiT* increases BLEU-3 scores from 2.75 to 14.07 for English, and yields an average improvement from 2.0 to 12.05 across target languages compared to direct translation. In addition, fine-tuned *CoPiT* models on open-source backbones match or exceed GPT-4.1 performance under both reference-based and reference-free evaluations. Beyond this, *CoPiT* generates a synthetic parallel corpus of 8,034 sentence pairs aligning Traditional Mongolian with English, Korean, and Russian. This synthetic data proves useful beyond forward translation: it further improves forward translation quality via translation module fine-tuning and enables reverse-direction translation, demonstrating the broader utility of our approach under realistic low-resource constraints.

Our contributions are as follows:

- We propose *CoPiT*, a cognitively motivated multi-stage pipeline that translates Traditional-script Mongolian by resolving script-induced ambiguity through Cyrillic pivoting.
- We release a multi-script parallel corpus aligning Traditional and Cyrillic Mongolian with English, Korean, and Russian, addressing a resource gap for digraphic low-resource MT.
- We show that *CoPiT* generally improves over direct translation across backbones and target languages, with fine-tuned open-source models matching or exceeding GPT-4.1 direct-translation performance.
- We show that *CoPiT*-generated synthetic data enables bidirectional translation, improving both Target→Traditional Mongolian and Traditional Mongolian→Target directions.

¹The Law on the Mongolian Language, adopted on February 12, 2015, and amended with effect from January 1, 2025, mandates the use of both Cyrillic and Traditional scripts in official state and local government affairs; Official text available at <https://legalinfo.mn/mn/detail/10932>.

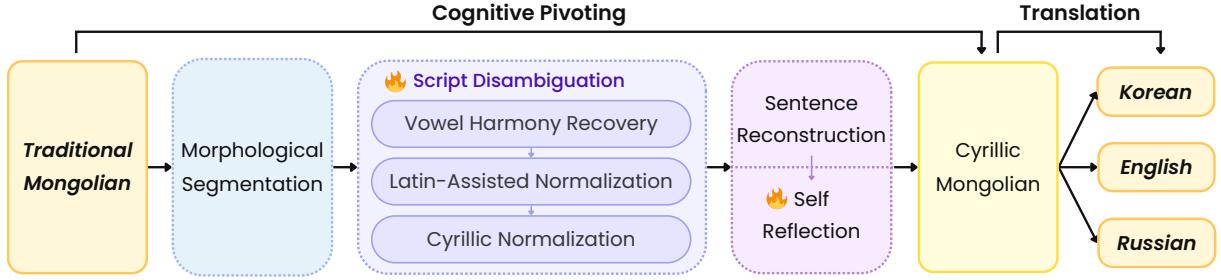


Figure 2: Overview of *CoPiT*, a pipeline translating Mongolian from the Traditional script via the Cyrillic script.

2 Related Work

2.1 Low-Resource Language Processing

Prior work on low-resource language processing can be broadly categorized into three lines of research. *Data-centric* methods aim to expand parallel data through web mining (Schwenk et al., 2021a,b), back-translation (Sennrich et al., 2016; Edunov et al., 2018), or LLM-generated synthetic corpora (de Gibert et al., 2025; Waldendorf et al., 2025). *Model-centric* methods leverage multilingual pretrained models (Conneau et al., 2020; Xue et al., 2021), but their shared representation assumptions often break down for languages with divergent or multiple writing systems (Liu et al., 2024). *Linguistically motivated* approaches incorporate explicit knowledge such as transliteration (Ma et al., 2025) or morphology-aware tokenization (Chaudhary et al., 2018; Nzeyimana, 2024). While effective for reducing surface-level variation, these methods do not explicitly model script-level ambiguity arising from digraphy.

Our work lies at the intersection of linguistically motivated modeling and data-centric augmentation, as the proposed pivot not only resolves script-level ambiguity but also enables the creation of synthetic parallel data for low-resource translation.

2.2 Mongolian Language Processing

Prior work on Mongolian language processing can be broadly grouped by the scripts they address: the Cyrillic script, the Traditional script, and conversion between the two. Most studies focus on Mongolian written in the Cyrillic script, including benchmark construction (Zhang et al., 2025a; Tumur-Ochir et al., 2025) and language model pre-training (Na et al., 2024b). These works typically assume input text written in a single standardized script and do not account for challenges arising from digraphia. Research on Mongolian written in the Traditional script has largely focused on classical NLP tasks such as morphological analysis (Liu

et al., 2020; Zhang et al., 2025b), with limited connection to downstream applications or LLM-based frameworks. Work addressing both scripts has primarily framed the problem as script conversion. Early approaches relied on rule-based methods (Li and Sarina, 2011), followed by hybrid systems combining linguistic rules with neural models (Na et al., 2023, 2024a), encoder-decoder architectures (Na et al., 2022), and more recent neural machine translation formulations (Dulamragchaa et al., 2024).

In this work, we treat script conversion as a structured inference step that resolves script-level ambiguity and produces a meaning-consistent intermediate representation, enabling more accurate translation from the Traditional script under digraphic low-resource conditions.

3 CoPiT: Cognitive Pivot Translation

3.1 Overview

Our task is to translate Mongolian text written in the Traditional script into a target language. This is challenging for current LLMs because Mongolian is a digraphic low-resource language with a strong resource imbalance across scripts. The Traditional script is severely under-resourced and orthographically ambiguous, leading to significant performance degradation compared to the Cyrillic script and making direct translation difficult.

To address this challenge, we propose *CoPiT*, a translation pipeline for Mongolian text written in the Traditional script. As illustrated in Figure 2, *CoPiT* is cognitively motivated by how fluent Mongolian readers process the Traditional script: rather than interpreting it directly, they implicitly map it to the Cyrillic script to resolve script-induced ambiguities. *CoPiT* consists of two stages: (1) Traditional-to-Cyrillic pivoting, which includes morphological segmentation, multi-step script disambiguation, and sentence-level self-reflection; and (2) translation from the disambiguated Cyrillic representation to the target language. More details, including the

Before Reflection : **Олс** бүр өөрсдийн далбаатай.
(Each **rope** has its own flag.)

After Reflection : **Улс** бүр өөрсдийн далбаатай.
(Each **country** has its own flag.)

Figure 5: Sentence-level self-reflection for global semantic and grammatical consistency.

chine translation (Wang et al., 2024; Chen et al., 2024). The model reviews its earlier local decisions and revises them only when they result in semantic or grammatical inconsistencies at the sentence level, as shown in Figure 5. Through this process, the model selects a single globally coherent analysis for the sentence, resolving distinctions such as lemma choice, case marking, or tense interpretation that cannot be reliably determined from the surface form of the Traditional script alone. Because many ambiguities in the Traditional script cannot be resolved locally, this reflection component plays a critical role in producing a coherent and internally consistent Cyrillic representation suitable for downstream translation.

3.3 Translation from the Disambiguated Representation

Finally, the disambiguated Cyrillic representation is translated into the target languages—English, Korean, and Russian—using a prompt-tuned LLM. Crucially, the MT system no longer operates under script-induced uncertainty: the ambiguities inherent in the Traditional script have been resolved upstream, reducing error propagation into the translation output. As a result, the translation model can focus on cross-lingual meaning transfer rather than script interpretation, yielding translations that better reflect the intended semantics of the original Traditional-script sentence. Compared to direct translation, this pipeline yields more stable translation behavior without requiring additional training, by explicitly resolving script-level ambiguity and enabling intermediate consistency checks.

3.4 Component-wise Training

We train each LLM-based component independently using supervised fine-tuning with parameter-efficient adaptation (Hu et al., 2022). Each component is trained with dedicated prompts and objectives, enabling effective specialization in its intended function within the overall pipeline.

The Vowel Harmony Recovery, Latin-Assisted Normalization, and Cyrillic Normalization components are trained using *word-level* supervision

from a Traditional–Cyrillic parallel lexical dataset comprising 14,125 entries. This dataset is derived from a dictionary that aligns Traditional and Cyrillic word forms with phonetic transliterations, with vowel harmony labels automatically assigned and verified by a native Mongolian speaker. Because the data consist of dictionary-level word pairs rather than sentence-level parallel corpora, they can be collected with substantially lower cost and effort, which is particularly advantageous in low-resource settings. The proposed components are explicitly designed to leverage this word-level lexical supervision and apply it compositionally to longer input sequences in the end-to-end pipeline.

The sentence-level self-reflection component is trained separately using 2,061 synthetic Cyrillic revision pairs. To construct this data, sentences written in the Traditional script are first converted into Cyrillic by a base LLM to produce imperfect or noisy hypotheses, which are then paired with their original Cyrillic counterparts. This data construction process is relatively straightforward in Mongolian due to its digraphic nature, where equivalent content is commonly available in both scripts. Training on such revision pairs enables the model to correct locally plausible but globally inconsistent analyses at the sentence level. All datasets used for component-wise training are publicly released to support reproducibility and future research.

4 Experiments

We conduct our experiments to address the following three research questions, which collectively examine translation quality, component-level effectiveness, and the potential of the proposed pipeline to alleviate low-resource challenges.

RQ1. Does our pipeline improve end-to-end machine translation performance over baselines?

RQ2. How does each component affect overall translation quality?

RQ3. Can the pipeline contribute to mitigating low-resource data challenges?

4.1 Experimental Setups

Datasets. Due to the severe data scarcity of Mongolian, it is impractical to collect parallel corpora that are well aligned across languages or to obtain fully human written reference translations. We therefore construct two datasets for reference-based and reference-free evaluation, respectively.

The first dataset contains 1,031 sentence-level

Backbone	Method	English			Korean			Russian		
		BLEU-3/4↑	chrF/chrF++↑	COMET↑	BLEU-3/4↑	chrF/chrF++↑	COMET↑	BLEU-3/4↑	chrF/chrF++↑	COMET↑
Qwen-3 (4B)	<i>Direct</i>	0.54/0.21	13.34/11.73	0.373	0.74/0.35	2.88/2.75	0.351	0.35/0.10	5.03/4.65	0.266
	<i>CoPiT (Zero-shot)</i>	0.94/0.46	15.57/13.73	0.422	0.92/0.41	3.50/3.36	0.372	0.59/0.31	11.42/9.37	0.512
	<i>CoPiT (Fine-tuned)</i>	9.40/5.70	29.70/27.55	0.628	10.78/6.65	11.15/10.91	0.639	3.45/1.99	18.60/16.24	0.544
Qwen-3 (30B)	<i>Direct</i>	1.25/0.48	19.73/17.51	0.443	1.28/0.59	3.94/3.85	0.421	0.33/0.12	9.55/8.33	0.331
	<i>CoPiT (Zero-shot)</i>	1.86/0.92	21.75/19.28	0.493	2.95/1.43	6.08/5.87	0.498	0.83/0.39	13.41/11.54	0.388
	<i>CoPiT (Fine-tuned)</i>	7.61/4.56	29.25/26.94	0.593	9.31/5.81	10.91/10.65	0.595	2.83/1.55	19.51/17.00	0.487
Ministral-3 (3B)	<i>Direct</i>	1.21/0.46	23.19/20.21	0.500	0.39/0.25	0.65/0.68	0.236	0.60/0.20	14.08/12.03	0.404
	<i>CoPiT (Zero-shot)</i>	1.09/0.22	22.66/19.82	0.497	0.81/0.30	3.61/3.42	0.393	0.69/0.26	16.34/14.15	0.439
	<i>CoPiT (Fine-tuned)</i>	4.81/2.66	29.38/26.21	0.619	5.59/3.17	9.46/8.88	0.585	2.21/1.00	21.92/19.08	0.547
Ministral-3 (14B)	<i>Direct</i>	1.79/0.77	21.44/19.13	0.467	1.44/0.65	4.53/4.30	0.435	0.92/0.34	16.70/14.44	0.473
	<i>CoPiT (Zero-shot)</i>	3.02/1.40	26.02/23.30	0.549	3.37/1.80	6.93/6.72	0.528	2.31/1.12	21.70/19.05	0.558
	<i>CoPiT (Fine-tuned)</i>	14.16/9.58	38.24/35.41	0.707	14.64/9.82	16.01/15.30	0.714	9.04/5.84	32.49/29.12	0.728
GPT-4.1	<i>Direct</i>	2.75/1.35	19.59/17.66	0.468	2.67/1.39	5.30/5.04	0.454	1.57/0.73	15.86/14.00	0.516
	<i>CoPiT (Zero-shot)</i>	14.07/9.83	37.29/24.59	0.663	13.37/9.04	14.59/14.05	0.667	8.71/5.63	31.84/28.53	0.694

Table 1: Machine translation performance across languages under reference-based evaluation.

Backbone	Method	English			Korean			Russian		
		COMETKiwi↑	Adeq.↑	Fluen.↑	COMETKiwi↑	Adeq.↑	Fluen.↑	COMETKiwi↑	Adeq.↑	Fluen.↑
Qwen-3 (4B)	<i>Direct</i>	0.078	1.33	1.89	0.204	1.00	1.33	0.167	1.00	1.00
	<i>CoPiT (Zero-shot)</i>	0.183	1.00	2.00	0.160	1.00	1.44	0.124	1.00	1.00
	<i>CoPiT (Fine-tuned)</i>	0.352	2.44	3.00	0.411	2.56	3.00	0.280	1.22	1.33
Qwen-3 (30B)	<i>Direct</i>	0.204	1.11	2.33	0.322	1.00	1.78	0.156	1.00	1.00
	<i>CoPiT (Zero-shot)</i>	0.250	2.11	3.00	0.376	2.00	3.00	0.211	1.56	2.67
	<i>CoPiT (Fine-tuned)</i>	0.360	2.33	3.44	0.419	2.67	2.89	0.301	1.89	2.22
Ministral-3 (3B)	<i>Direct</i>	0.315	1.22	2.67	0.344	1.00	1.44	0.228	1.11	1.33
	<i>CoPiT (Zero-shot)</i>	0.291	1.22	2.89	0.272	1.00	1.22	0.237	1.00	1.00
	<i>CoPiT (Fine-tuned)</i>	0.402	2.78	3.22	0.413	2.33	2.56	0.338	1.89	2.22
Ministral-3 (14B)	<i>Direct</i>	0.360	1.56	2.44	0.391	1.00	1.44	0.360	1.00	1.33
	<i>CoPiT (Zero-shot)</i>	0.309	2.11	2.33	0.400	1.89	1.89	0.349	1.89	2.11
	<i>CoPiT (Fine-tuned)</i>	0.461	4.22	4.44	0.490	3.56	3.33	0.449	3.67	3.89
GPT-4.1	<i>Direct</i>	0.270	1.44	2.67	0.249	1.44	1.56	0.429	1.44	2.67
	<i>CoPiT (Zero-shot)</i>	0.470	4.33	4.56	0.472	3.44	3.33	0.399	4.00	4.44

Table 2: Machine translation performance across languages under reference-free evaluation.

web-crawled Traditional–Cyrillic parallel pairs with human-verified reference translations. English references are translated from the Cyrillic side and subsequently validated by bilingual speakers. Korean and Russian references are then generated from the verified English sentences and independently reviewed by fluent speakers of each target language. This dataset enables reliable reference-based evaluation across multiple target languages.

The second dataset comprises 380 news articles, each containing about five sentences on average, in the Traditional and Cyrillic scripts collected from the official Mongolian president’s website.² As no reference translations are available, this dataset is used exclusively for reference-free evaluation.

Backbone Models. We adopt Qwen-3 (4B, 30B) (Yang et al., 2025) and Ministral-3 (3B, 14B) (Mistral AI, 2025) as state-of-the-art multilingual open-source backbones, and GPT-4.1 (Achiam et al., 2023; OpenAI, 2024) as representative commercial LLMs, enabling a comprehensive evaluation across model scale, and real-world deployment settings.

Evaluation Metrics. We employ different evaluation metrics depending on the availability of reference translations. For datasets with references, we use BLEU-3/4 (Papineni et al., 2002)

and chrF/chrF++ (Popović, 2015, 2017) to measure lexical overlap, along with COMET (Rei et al., 2022) to assess semantic adequacy.

For datasets without reference translations, we adopt COMETKiwi (Rei et al., 2023), a reference-free quality estimation metric that scores translation quality from source–hypothesis pairs. We additionally conduct human evaluation with six fluent bilingual speakers of Mongolian and each target language. Translations are evaluated along two criteria: **Adequacy (Adeq.)**: Faithful meaning preservation with respect to the source text. **Fluency (Fluen.)**: Grammaticality and overall naturalness of the translation in the target language.

Implementation Details. Open-source backbone models are fine-tuned with LoRA, as described in Section 3.4. GPT-4.1 is accessed via its APIs without additional fine-tuning. Additional implementation details, including hyperparameters, train/validation/test dataset split information, and inference settings, are provided in Appendix A.2.

4.2 Experimental Results

4.2.1 Machine Translation Evaluation

Table 1 presents reference-based evaluation results across English, Korean, and Russian. Across all backbones, fine-tuned *CoPiT* outperforms di-

²<https://president.mn/en/>

Backbone	Method	English			Korean			Russian		
		BLEU-3/4↑	chrF/chrF++↑	COMET↑	BLEU-3/4↑	chrF/chrF++↑	COMET↑	BLEU-3/4↑	chrF/chrF++↑	COMET↑
Qwen-3 (4B)	<i>Direct</i>	0.54/0.21	13.34/11.73	0.373	0.74/0.35	2.88/2.75	0.351	0.35/0.10	5.03/4.65	0.266
	<i>CoPiT w/o VHR</i>	8.08/5.00	29.57/29.37	0.633	6.79/4.17	10.50/10.23	0.638	2.00/1.12	17.31/15.17	0.546
	<i>CoPiT w/o LAN</i>	3.06/1.55	23.34/21.28	0.568	3.42/1.93	7.25/7.09	0.558	0.74/0.39	13.04/11.20	0.456
	<i>CoPiT w/o SR</i>	1.44/0.74	18.22/16.36	0.550	1.60/0.88	5.19/4.99	0.540	0.44/0.20	10.50/8.94	0.415
	<i>CoPiT</i>	9.40/5.70	29.70/27.55	0.628	10.78/6.65	11.15/10.91	0.639	3.45/1.99	18.60/16.24	0.544
Ministral-3 (3B)	<i>Direct</i>	1.21/0.46	23.19/20.21	0.500	0.39/0.25	0.65/0.68	0.236	0.60/0.20	14.08/12.03	0.404
	<i>CoPiT w/o VHR</i>	5.02/2.71	29.92/26.72	0.625	4.57/2.66	9.17/8.59	0.594	2.03/1.07	20.30/17.62	0.542
	<i>CoPiT w/o LAN</i>	3.86/1.95	28.64/25.47	0.607	3.40/1.89	7.78/7.30	0.563	1.60/0.84	18.60/16.12	0.507
	<i>CoPiT w/o SR</i>	2.46/1.16	26.41/23.21	0.577	1.68/0.82	5.84/5.41	0.510	0.58/0.22	15.17/13.04	0.442
	<i>CoPiT</i>	4.81/2.66	29.38/26.21	0.619	5.59/3.17	9.46/8.88	0.585	2.21/1.00	21.92/19.08	0.547

Table 3: Ablation study of individual components in the proposed *CoPiT* across target languages under reference-based evaluation, where *VHR*, *LAN*, and *SR* denote Vowel Harmony Recovery, Latin-Assisted Normalization, and Self-Reflection, respectively.

rect translation. Even without training, the *CoPiT* pipeline yields measurable gains, particularly for Korean and Russian, indicating the effectiveness of script-level pivoting for typologically distant languages. For example, on Ministral-3 (3B), zero-shot *CoPiT* improves COMET by +0.157 for Korean and +0.035 for Russian, while yielding no improvement for English. With component fine-tuning (check, gains become substantial across all languages; on Ministral-3 (14B), COMET increases by +0.240 (English), +0.279 (Korean), and +0.255 (Russian). Similar trends are observed for BLEU and chrF/chrF++, where fine-tuned *CoPiT* consistently achieves higher scores across languages. Overall, larger backbones attain higher absolute performance, and fine-tuned *CoPiT* on open-source models matches or exceeds GPT-4.1 under direct translation.

Table 2 reports reference-free evaluation results using COMETKiwi and human adequacy and fluency ratings. Consistent with reference-based findings, fine-tuned *CoPiT* improves translation quality across most backbones over direct method. On Ministral-3 (14B), fine-tuned *CoPiT* improves COMETKiwi by +0.101 (English), +0.099 (Korean), and +0.089 (Russian), accompanied by gains in human adequacy and fluency across most conditions. When applied to GPT-4.1, zero-shot *CoPiT* also yields improvements, increasing COMETKiwi by +0.200 (English) and +0.223 (Korean), highlighting the effectiveness of pivot-based inference in reference-scarce settings. Across both reference-based and reference-free evaluations, *CoPiT* improves translation quality across model backbones and target languages. Applying the *CoPiT* pipeline alone already provides gains in most cases, while fine-tuning the pipeline components yields the best overall performance. These results demonstrate

that script-level pivoting is an effective and practical approach for improving low-resource digraphic machine translation and for substantially reducing the performance gap between open-source and proprietary models. Detailed annotation protocols with inter-annotator agreement analyses and few-shot prompting baselines are provided in Appendix A.3 and A.4, respectively.

4.2.2 Ablation Study

Table 3 presents an ablation study of individual components in the proposed *CoPiT* under reference-based evaluation. Across both Qwen-3 (4B) and Ministral-3 (3B), the full pipeline achieves the strongest overall performance, attaining the best or second-best results in most settings. This indicates that backbones benefit from the complete pipeline rather than from any single component in isolation. At the same time, performance does not consistently decrease when individual components are removed, suggesting non-trivial interactions among components and backbone-dependent sensitivities.

The impact of Vowel Harmony Recovery, in particular, varies across backbones and languages. For Qwen-3 (4B), its removal reduces BLEU scores, while chrF++ and COMET remain comparable or slightly improved. In contrast, for Ministral-3 (3B), removing Vowel Harmony Recovery improves English performance across all metrics and yields competitive or higher COMET scores for Korean, suggesting partial redundancy with the backbone’s internal phonological representations. In comparison, Self-Reflection emerges as the most consistently important component: its removal leads to substantial performance degradation across both backbones, all target languages, and all evaluation metrics. Latin-Assisted Normalization provides complementary but moderate gains, indicating a

Backbone	Method	English→TM		Korean→TM		Russian→TM	
		BLEU-3/4↑	chrF/chrF++↑	BLEU-3/4↑	chrF/chrF++↑	BLEU-3/4↑	chrF/chrF++↑
Qwen-3 (4B)	Base	0.07/0.03	0.10/0.12	0.10/0.04	0.12/0.14	0.07/0.02	0.10/0.12
	Fine-tuned	5.04/3.22	26.03/22.61	6.68/4.26	29.94/26.21	4.43/2.68	25.08/21.78
Ministral-3 (3B)	Base	0.05/0.02	1.02/0.80	0.04/0.01	3.18/2.42	0.04/0.02	0.19/0.18
	Fine-tuned	5.31/3.28	23.81/20.64	6.99/4.39	28.38/24.88	5.42/3.36	24.34/21.11

Table 4: Reverse-direction translation performance across languages using *CoPiT*-generated synthetic parallel data.

supporting role in orthographic stabilization rather than a dominant contribution. Overall, these findings show that the full *CoPiT* pipeline improves performance across different model backbones and all target languages. The single- and pairwise- component ablation analyses for Traditional-to-Cyrillic pivoting stage is provided in Appendix A.6.

4.2.3 Low-Resource Mitigation

Synthetic Data Generation. Using the proposed *CoPiT* pipeline with Ministral-3 (14B), we construct a new parallel corpus of 8,034 sentence pairs aligning Mongolian written in the Traditional script with English, Korean, and Russian. We clarify that this dataset is not purely synthetic; while the target-language translations are generated by our pipeline, the Traditional Mongolian source texts are curated from real-world corpora. This resource addresses a critical data gap for digraphic low-resource languages. We validate its usefulness by reversing the dataset for reverse-direction translation and training the translation module separately. All datasets are publicly released to facilitate future research. A failure mode analysis of the *CoPiT*-generated synthetic corpus is provided in Appendix A.7.

Data Scaling Analysis. Figure 6a presents translation performance as a function of synthetic data scale. Specifically, models were fine-tuned with Ministral-3 (3B) across increasing data scales (1K–7K) and evaluated on a fixed reference test set. Across all target languages, performance generally improves as more synthetic data is added. From the smallest to the largest data scale, COMET scores increase by approximately +0.09 for English, +0.36 for Korean, and +0.21 for Russian. At smaller scales, English shows a slight performance decrease from Base to 1K, with a reduction of approximately −0.03, while Korean exhibits nearly negligible changes between 1K and 3K. Such behavior is characteristic of unstable learning dynamics in extremely low-resource regimes (Sennrich and Zhang, 2019). Beyond this range, performance improvements remain stable across languages, despite typological differences.

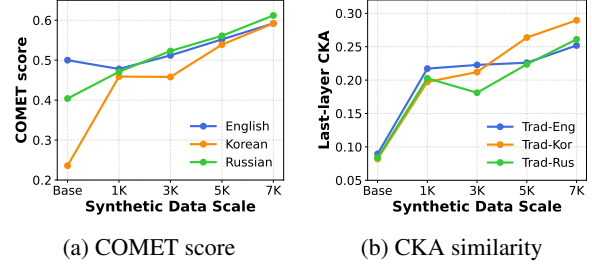


Figure 6: Evaluation of Ministral-3 (3B) trained with varying amounts of *CoPiT*-generated synthetic data.

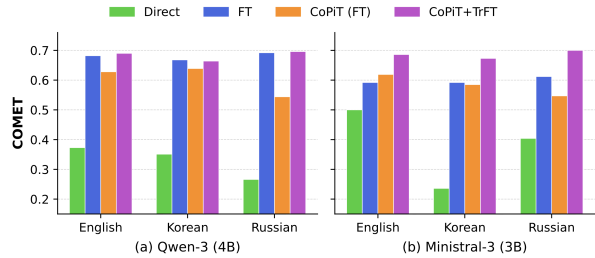


Figure 7: COMET scores under pipeline configurations with and without translation module fine-tuning.

To relate performance trends to internal representations, we analyze last-layer representations using centered kernel alignment (CKA) (Kornblith et al., 2019), which quantifies how representations induced by Traditional-script Mongolian align across different target languages as training data scale increases, as shown in Figure 6b. Overall, representations evolve smoothly as training data increases, with CKA similarity rising by roughly +0.16–0.20 across all target languages from the lowest to the highest scale. A mild decrease in CKA similarity is observed for Russian between 1K and 3K, with an approximate reduction of −0.02, aligning with the low-data regime where performance gains remain limited for other target languages. Together with the performance trends in Figure 6a, this behavior is consistent with a transitional adaptation phase under extremely limited data. In this regime, representation and performance do not yet evolve synchronously across languages. As more data is introduced, both trends stabilize, demonstrating that *CoPiT*-generated data provides reliable supervision for low-resource machine translation.

Improving Forward Translation. To test whether CoPiT-generated synthetic data improves forward translation, we fine-tune the translation module on the Cyrillic–Target portion of the corpus and combine it with the pivot mechanism (*CoPiT+TrFT*). We compare with *Direct* (direct translation), *FT* (direct translation fine-tuned on the synthetic data), and *CoPiT (FT)* (component fine-tuned *CoPiT* pipeline without translation module fine-tuning). Since FT and CoPiT+TrFT are trained on identical synthetic data, performance differences generally reflect the structural contribution of the *CoPiT* pipeline rather than data exposure.

As shown in Figure 7, CoPiT+TrFT tends to achieve strong COMET scores across most languages and backbones, suggesting that *CoPiT*-generated synthetic data can provide effective supervision for translation; full results across BLEU-4 and chrF++ are provided in Appendix A.8. In most settings, CoPiT+TrFT outperforms FT despite using the same training data, offering evidence that the pivot mechanism contributes independently of the data; CoPiT (FT) also performs comparably to FT in several settings, suggesting that the structural benefits of pivoting are substantial even without translation module fine-tuning with parallel data. We note that this comparison represents an upper-bound analysis, as CoPiT is by design trained using only Traditional Mongolian–Cyrillic data without access to target-language references.

Reverse-Direction Translation. To further demonstrate the practical utility of *CoPiT*, we examine whether the synthetic parallel data it generates can enable *Target* $\rightarrow$ *Traditional Mongolian* translation, a direction that is otherwise infeasible or limited due to the lack of parallel resources. We construct reverse-direction training data by flipping each pair in the *CoPiT*-generated synthetic corpus and fine-tune the models on this reversed data.

The fine-tuned models yield generally consistent improvements over base models across languages and backbones. As shown in Table 4, BLEU and chrF scores increase steadily for ENG $\rightarrow$ TM, KOR $\rightarrow$ TM, and RUS $\rightarrow$ TM, suggesting that the *CoPiT*-generated synthetic data can provide effective supervision for generating Traditional Mongolian for both backbones.

We note that evaluation for Traditional Mongolian remains challenging due to the absence of specialized metrics and the limited representation of the script in existing multilingual models. As a result, semantic metrics such as COMET are unre-

liable for this language. Following prior work on low-resource and morphologically rich languages, we therefore report BLEU and chrF/chrF++ scores as reference-based evaluation metrics. While our forward-direction results are verified via human evaluation, we restrict our reverse-direction analysis to the aforementioned metrics, leaving comprehensive human validation of generated Traditional Mongolian script to future work.

5 Contribution

In this paper, we introduce *CoPiT*, a cognitively motivated translation pipeline for digraphic low-resource languages, focusing on Mongolian in the Traditional script. Rather than treating Traditional-to-Cyrillic conversion as standalone transcription, *CoPiT* formulates it as structured disambiguation, using Cyrillic as a meaning-consistent intermediate representation for downstream translation. This design mirrors native reading practices and decouples script-level ambiguity from semantic transfer. Our pipeline integrates linguistically grounded components, vowel harmony recovery, intermediate normalization, and sentence-level reflection, each targeting a distinct source of underspecification in the Traditional script. Extensive experiments across multiple backbones and target languages show consistent improvements, including in zero-shot settings. Beyond forward translation, *CoPiT*-generated synthetic data enables reverse-direction translation and further improves forward translation quality via translation module fine-tuning, demonstrating the practical utility of our approach for digraphic low-resource machine translation.

Limitations

The effectiveness of the proposed *CoPiT* pipeline depends in part on the quality of the downstream Cyrillic-to-target translation stage. This reliance is partly mitigated by the relatively higher availability of Traditional–Cyrillic parallel data and by the steady increase in resources, which support continued improvement of the pivoting and synthetic data generation process. As shown in Section 4.2, the resulting *CoPiT*-generated synthetic data can further improve downstream translation quality through fine-tuning. In addition, the multi-stage design introduces higher inference latency compared to single-stage models, as multiple conversion and translation components are sequentially applied. Detailed latency analysis is provided in

Appendix A.5. Although this limits applicability to strict real-time settings, the pipeline is more suitable for offline document translation and batch-oriented low-resource MT scenarios. Finally, the core *CoPiT* inference pipeline is designed for out-bound translation from Traditional Mongolian into non-Mongolian target languages. While our experiments show that *CoPiT*-generated synthetic data can support preliminary Target→Traditional Mongolian translation across multiple language pairs and backbones, a fully integrated bidirectional pipeline with dedicated evaluation metrics and human validation for Traditional Mongolian remains future work.

Acknowledgements

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No. IITP-2026RS-2020-II201819, the ICT Creative Consilience Program, 30%; No. RS-2026-25585074, Development of Agentic AI Technology for Enterprise-wide Business Innovation Based on Goal-Oriented Long-term Context Reasoning, 50%; No. RS-2026-25507543, Development of AI Co-Scientist based on Scientific Causal World Model, 10%). This work was also partly supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. RS-2025-24533089, 10%).

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. *GPT-4 Technical Report*. *arXiv preprint arXiv:2303.08774*.
- Ganchimeg Altangerel and Otgontuul Togtokh. 2024. *Classical Mongolian: A Textbook for Students, Scholars, and Everyone Interested in Mongolian*. Reichert Verlag, Wiesbaden, Germany.
- Aditi Chaudhary, Chunting Zhou, Lori Levin, Graham Neubig, David R. Mortensen, and Jaime Carbonell. 2018. *Adapting word embeddings to new languages with morphological and phonological subword representations*. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3285–3295, Brussels, Belgium. Association for Computational Linguistics.
- Andong Chen, Lianzhang Lou, Kehai Chen, Xuefeng Bai, Yang Xiang, Muyun Yang, Tiejun Zhao, and Min Zhang. 2024. *DUAL-REFLECT: Enhancing large language models for reflective translation through dual learning feedback mechanisms*. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 693–704, Bangkok, Thailand. Association for Computational Linguistics.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. *Unsupervised cross-lingual representation learning at scale*. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- Ona de Gibert, Joseph Attieh, Teemu Vahtola, Mikko Aulamo, Zihao Li, Raúl Vázquez, Tiancheng Hu, and Jörg Tiedemann. 2025. *Scaling low-resource MT via synthetic data generation with LLMs*. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27662–27680, Suzhou, China. Association for Computational Linguistics.
- Uuganbaatar Dulamragchaa, Nomuundalai Batbileg, Sainbaya Batbaatar, Purevsuren Tumurbaatar, Bilguutei Narantuya, Baatar Nyamkhuu, Tungalagtamir Bold, and Bayarchimeg Enkhtaivan. 2024. *Joint nmt models for text conversion between traditional mongolian script and cyrillic mongolian: a comparative study*. *Journal of Institute of Mathematics and Digital Technology*, 6(1):99–108.
- Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. *Understanding back-translation at scale*. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. *LoRA: Low-rank adaptation of large language models*. In *International Conference on Learning Representations*.
- Juha A. Janhunen. 2012. *Mongolian*. John Benjamins Publishing Company.
- Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. *Similarity of neural network representations revisited*. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3519–3529. PMLR.
- Klaus Krippendorff. 2019. *Content Analysis: An Introduction to Its Methodology*, fourth edition. SAGE Publications, Inc., Thousand Oaks, CA.
- Hao Li and Bao Sarina. 2011. *The study of comparison and conversion about traditional mongolian and*

- cyrillic mongolian. In *2011 4th International Conference on Intelligent Networks and Intelligent Systems*, pages 199–202.
- Na Liu, Xiangdong Su, Haoran Zhang, Guanglai Gao, and Feilong Bao. 2020. [Incorporating inner-word and out-word features for Mongolian morphological segmentation](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4638–4648, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Yihong Liu, Chunlan Ma, Haotian Ye, and Hinrich Schuetze. 2024. [TransliCo: A contrastive learning framework to address the script barrier in multilingual pretrained language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2476–2499, Bangkok, Thailand. Association for Computational Linguistics.
- Chunlan Ma, Yihong Liu, Haotian Ye, and Hinrich Schuetze. 2025. [Exploring the role of transliteration in in-context learning for low-resource languages written in non-Latin scripts](#). In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 397–410, Suzhou, China. Association for Computational Linguistics.
- Mistral AI. 2025. [Introducing mistral 3](#). <https://mistral.ai/news/mistral-3>. Accessed: 2025-12-26.
- Muhan Na, Feilong Bao, Weihua Wang, Guanglai Gao, and Uuganbaatar Dulamragchaa. 2023. [Traditional mongolian-to-cyrillic mongolian conversion method based on the combination of rules and transformer](#). In *2023 IEEE 9th International Conference on Cloud Computing and Intelligent Systems (CCIS)*, pages 440–445.
- Muhan Na, Feilong Bao, Weihua Wang, Guanglai Gao, and Uuganbaatar Dulamragchaa. 2024a. [Cyrillic mongolian-to-traditional mongolian conversion method based on the transformer](#). *Journal of Institute of Mathematics and Digital Technology*, 6(1):120–129.
- Muhan Na, Rui Liu, Feilong Bao, and Guanglai Gao. 2022. [A deep investigation of RNN and self-attention for the cyrillic-traditional mongolian bidirectional conversion](#). In *Neural Information Processing - 29th International Conference, ICONIP 2022, Virtual Event, November 22-26, 2022, Proceedings, Part VI*, volume 1793 of *Communications in Computer and Information Science*, pages 612–625. Springer.
- Muhan Na, Rui Liu, Feilong Bao, and Guanglai Gao. 2024b. [Pre-training language model for mongolian with agglutinative linguistic knowledge injection](#). In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.
- Antoine Nzeyimana. 2024. [Low-resource neural machine translation with morphological modeling](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 182–195, Mexico City, Mexico. Association for Computational Linguistics.
- OpenAI. 2024. Gpt-4.1. <https://openai.com/index/gpt-4-1/>. Accessed: 2025-01-10.
- 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.
- Maja Popović. 2017. [chrF++: words helping character n-grams](#). In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.
- Rahul Raja and Arpita Vats. 2025. [Parallel corpora for machine translation in low-resource Indic languages: A comprehensive review](#). In *Proceedings of the Eighth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2025)*, pages 129–143, Albuquerque, New Mexico, U.S.A. Association for Computational Linguistics.
- Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. [COMET-22: Unbabel-IST 2022 submission for the metrics shared task](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Ricardo Rei, Nuno M. Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André F. T. Martins. 2023. [Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore. Association for Computational Linguistics.
- Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021a. [Wiki-Matrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.
- Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021b. [CCMatrix: Mining billions of high-quality parallel sentences on the web](#). In *Proceedings of the*

- 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 6490–6500, Online. Association for Computational Linguistics.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. [Improving neural machine translation models with monolingual data](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.
- Rico Sennrich and Biao Zhang. 2019. [Revisiting low-resource neural machine translation: A case study](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Florence, Italy. Association for Computational Linguistics.
- Jan-Olof Svantesson. 2024. [Vowel harmony in mongolic languages](#). In *The Oxford Handbook of Vowel Harmony*, pages 774–779. Oxford University Press.
- Dorjnyam Tumur-Ochir, Fei Cheng, Yugo Murawaki, and Chenhui Chu. 2025. [Evaluating large language models in mongolian](#). In *Proceedings of the 31st Annual Meeting of the Association for Natural Language Processing*, pages 368–373.
- Jonas Waldendorf, Barry Haddow, Alexandra Birch, and Mateusz Klimaszewski. 2025. [Multilingual data filtering using synthetic data from large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 9317–9334, Suzhou, China. Association for Computational Linguistics.
- Yutong Wang, Jiali Zeng, Xuebo Liu, Fandong Meng, Jie Zhou, and Min Zhang. 2024. [TasTe: Teaching large language models to translate through self-reflection](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6144–6158, Bangkok, Thailand. Association for Computational Linguistics.
- Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. [mT5: A massively multilingual pre-trained text-to-text transformer](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Mengyuan Zhang, Ruihui Wang, Bo Xia, Yuan Sun, and Xiaobing Zhao. 2025a. [MM-Eval: A Hierarchical Benchmark for Modern Mongolian Evaluation in LLMs](#). In *Intelligent Multilingual Information Processing*, pages 352–363, Singapore. Springer Nature Singapore.
- Zhonghao Zhang, Jiajia Ma, Na Liu, Nier Wu, Yatu Ji, and Guiping Liu. 2025b. [Dependent syntactic analysis of mongolian based on semi-supervised self-training](#). In *2025 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Language	#Annotators	#Systems	Krippendorff's α
			Adeq. / Fluen.
English	6	14	0.630 / 0.263
Korean	6	14	0.639 / 0.274
Russian	6	14	0.618 / 0.635

Table 6: Inter-annotator agreement (Krippendorff's α)

Language	Combined α	B1 α / B2 α	Pairwise $ \Delta \leq 1$
English	0.263	0.378 / 0.200	73%
Korean	0.274	0.578 / 0.135	78%
Russian	0.635	0.553 / 0.738	81%

Table 7: Fluency agreement statistics.

Adequacy shows moderate agreement across languages ($\alpha = 0.618$ – 0.639), indicating reasonable consistency in meaning preservation judgments. Fluency exhibits greater variability across languages (English: $\alpha = 0.263$; Korean: $\alpha = 0.274$; Russian: $\alpha = 0.635$), reflecting the subjective nature of fluency assessment, as perceptions of naturalness vary across annotators. To further examine this fluency pattern, we conducted a per-batch analysis. Agreement shown in Table 7 varies across groups: some batches show moderate consistency (e.g., Korean B1; Russian B1 and B2), while others are lower (e.g., English B2; Korean B2). This variability suggests differences in scoring tendencies, where certain annotators systematically use lower or higher portions of the rating scale.

Importantly, absolute difference analysis indicates that disagreements are predominantly small: 73% (English), 78% (Korean), and 81% (Russian) of pairwise ratings are exact matches or differ by only one level, while extreme disagreements are rare. This pattern suggests boundary-level variation in fluency judgments.

Overall, although fluency assessment involves inherent subjectivity and scale variation, the structured disagreement pattern and limited extreme divergence support the use of the annotations for comparative system analysis.

A.4 Few-shot Prompting Evaluation

Table 8 presents machine translation performance under few-shot prompting. Few-shot baselines are additionally evaluated for Ministral-3 (3B) and Qwen-3 (4B) under both Direct and *CoPiT* settings; results for other conditions are carried over from Table 1 for reference. Both *CoPiT* and Direct improve over their respective zero-shot counterparts with few-shot prompting; however, fine-tuned

CoPiT still achieves the best performance across languages and backbones, indicating that supervised adaptation drives the primary gains rather than prompt-level context alone.

A.5 Inference Latency Analysis

Although *CoPiT* improves translation quality, it incurs additional inference latency due to its multi-stage pipeline. Table 9 shows that *CoPiT* requires approximately 1.9 seconds more per sample than the direct baseline, primarily due to the Vowel Harmony Recovery, Latin-Assisted Normalization, and Self-Reflection components. In contrast, the direct approach often produces substantially lower-quality translations, whereas *CoPiT* consistently generates higher-quality outputs across model backbones and target languages, as reflected in Table 1. This additional latency may be acceptable in offline or batch-oriented translation settings where translation quality is prioritized.

A.6 Component Ablation Analysis

We present two complementary ablation analyses targeting different stages of the *CoPiT* pipeline. The first evaluates the contribution of individual components to end-to-end translation quality, covering the full pipeline including the translation stage in Section 4.2.2. The second focuses exclusively on the Traditional-to-Cyrillic pivoting stage, examining how component interactions affect script conversion quality prior to translation.

Traditional-to-Cyrillic Stage Ablation Study. This analysis isolates the Traditional-to-Cyrillic pivoting stage. Tables 10 and 11 present single- and pairwise-component ablations, evaluated using chrF/chrF++ and Character Error Rate (CER).

Across both backbones, jointly removing Latin-Assisted Normalization and Self-Reflection results in the worst performance, yielding the lowest chrF/chrF++ scores and the highest CER. This suggests that these two components provide complementary contributions to pivoting performance. In contrast, removing Vowel Harmony Recovery together with Latin-Assisted Normalization leads to comparatively better scores, suggesting that Latin-Assisted Normalization provides limited additional benefit when phonological regularization from Vowel Harmony Recovery is already absent.

Compared with single-component ablations in Table 10, pairwise removals in Table 11 consistently exhibit substantially lower performance, confirming that Traditional-to-Cyrillic pivoting quality

Backbone	Method	English			Korean			Russian		
		BLEU-3/4↑	chrF/chrF++↑	COMET↑	BLEU-3/4↑	chrF/chrF++↑	COMET↑	BLEU-3/4↑	chrF/chrF++↑	COMET↑
Qwen-3 (4B)	<i>Direct</i>	0.54/0.21	13.34/11.73	0.373	0.74/0.35	2.88/2.75	0.351	0.35/0.10	5.03/4.65	0.266
	<i>Direct (Few-shot)</i>	1.78/0.78	19.45/17.10	0.491	3.25/1.59	4.79/4.66	0.505	0.72/0.34	13.95/11.65	0.486
	<i>CoPiT (Zero-shot)</i>	0.94/0.46	15.57/13.73	0.422	0.92/0.41	3.50/3.36	0.372	0.59/0.31	11.42/9.37	0.312
	<i>CoPiT (Few-shot)</i>	1.70/0.78	20.01/17.58	0.463	1.89/0.84	4.78/4.60	0.434	0.91/0.39	15.42/13.07	0.397
	<i>CoPiT (Fine-tuned)</i>	9.40/5.70	29.70/27.55	0.628	10.78/6.65	11.15/10.91	0.639	3.45/1.99	18.60/16.24	0.544
Ministral-3 (3B)	<i>Direct</i>	1.21/0.46	23.19/20.21	0.500	0.39/0.25	0.65/0.68	0.236	0.60/0.20	14.08/12.03	0.404
	<i>Direct (Few-shot)</i>	1.15/0.43	16.72/14.39	0.456	2.56/1.05	4.65/4.32	0.474	0.96/0.39	17.77/14.94	0.511
	<i>CoPiT (Zero-shot)</i>	1.09/0.22	22.66/19.82	0.497	0.81/0.30	3.61/3.42	0.393	0.69/0.26	16.34/14.15	0.439
	<i>CoPiT (Few-shot)</i>	1.12/0.38	23.54/20.37	0.514	2.33/0.99	5.68/5.38	0.491	0.89/0.17	21.45/18.34	0.517
	<i>CoPiT (Fine-tuned)</i>	4.81/2.66	29.38/26.21	0.619	5.59/3.17	9.46/8.88	0.585	2.21/1.00	21.92/19.08	0.547

Table 8: Machine translation performance under few-shot prompting, evaluated under reference-based metrics.

Method	VHR, LAN	Self-Reflection	Translation	Total
<i>Direct</i>	—	—	3.68 s	3.68 s
<i>CoPiT</i>	4.29 s	0.67 s	0.64 s	5.60 s

Table 9: Inference latency comparison between *Direct* and *CoPiT*. *VHR*, *LAN* stand for Vowel Harmony Recovery, Latin-Assisted Normalization, respectively.

Model	Method	chrF / chrF++↑	CER↓
Qwen-3 (4B) + <i>CoPiT</i>	<i>w/o VHR</i>	81.01/77.20	0.10
	<i>w/o LAT</i>	61.85/55.71	0.23
	<i>w/o SR</i>	51.20/43.71	0.27
Ministral-3 (3B) + <i>CoPiT</i>	<i>w/o VHR</i>	78.98/74.66	0.11
	<i>w/o LAT</i>	72.50/ 67.58	0.16
	<i>w/o SR</i>	51.92/ 44.31	0.26

Table 10: Ablation study of single components for Traditional-to-Cyrillic script pivoting, where *VHR*, *LAN*, and *SR* denote Vowel Harmony Recovery, Latin-Assisted Normalization, and Self-Reflection, respectively.

depends on multiple components operating jointly rather than independently.

A.7 Assessment of Synthetic Dataset Noisiness

To assess the noisiness of the synthetic dataset of N=8,034 (Section 4.2), we conducted a targeted inspection as an indirect measure of data quality, complementing the downstream performance improvements observed in our experiments. Specifically, we examined three interpretable failure modes: *prompt artifact leakage*, *sentence-mode drift*, and *code-mixing* across three target languages.

Table 12 reports the observed error rates. Overall, artifact leakage occurs most frequently across languages (4.10%), followed by code-mixing (3.67%) and sentence-mode drift (1.93%). For Korean, code-mixing appears at a noticeably higher rate than the other error types, while for Russian, artifact leakage constitutes the dominant failure mode. For English, artifact leakage accounts for 2.74% of outputs, followed by sentence-mode drift

Model	Method	chrF / chrF++↑	CER↓
Qwen-3 (4B) + <i>CoPiT</i>	<i>w/o VHR & LAT</i>	60.87/54.75	0.24
	<i>w/o VH & SR</i>	50.94/43.40	0.28
	<i>w/o LAT & SR</i>	44.57/37.20	0.33
Ministral-3 (3B) + <i>CoPiT</i>	<i>w/o VHR & LAT</i>	70.09/65.14	0.18
	<i>w/o VHR & SR</i>	51.91/44.26	0.27
	<i>w/o LAT & SR</i>	49.33/41.96	0.29

Table 11: Ablation study of pairwise components for Traditional-to-Cyrillic script pivoting, where *VHR*, *LAN*, and *SR* denote Vowel Harmony Recovery, Latin-Assisted Normalization, and Self-Reflection, respectively.

(1.02%) and code-mixing (0.39%).

Despite these issues, the overall error rates remain relatively low, and this analysis suggests that *CoPiT* and its generated datasets are still useful in low-resource mitigation scenarios despite some noisiness levels.

A.8 Improving Forward Translation.

Figure 9 reports full results across BLEU-4, chrF++, and COMET for both backbones. The trends observed in the main paper are largely consistent across all metrics: *CoPiT*+TrFT achieves the strongest results in most settings, with particularly pronounced gains in BLEU-4 and chrF++ for English and Korean across both backbones. FT and *CoPiT* (FT) show mixed relative performance depending on the language and metric; however, we note that this comparison is not data-matched, as FT fine-tunes the translation module directly with *CoPiT*-generated synthetic data while *CoPiT* (FT) applies only component-level fine-tuning without translation module fine-tuning. The more informative comparison is between FT and *CoPiT*+TrFT, which are trained on identical synthetic data but differ in whether the pivot mechanism is applied: *CoPiT*+TrFT consistently outperforms FT across most languages and metrics, providing cleaner evidence that the pivot mechanism contributes independently of the data. We note that this comparison

Language	Artifact Leakage (%)	Sentence-mode Drift (%)	Code Mixing (%)
English	2.74	1.02	0.39
Korean	0.17	0.54	2.91
Russian	1.19	0.37	0.37
Total (N = 24,102)	4.10	1.93	3.67

Table 12: Error rates (%) across translation outputs in the *CoPiT*-generated synthetic dataset.

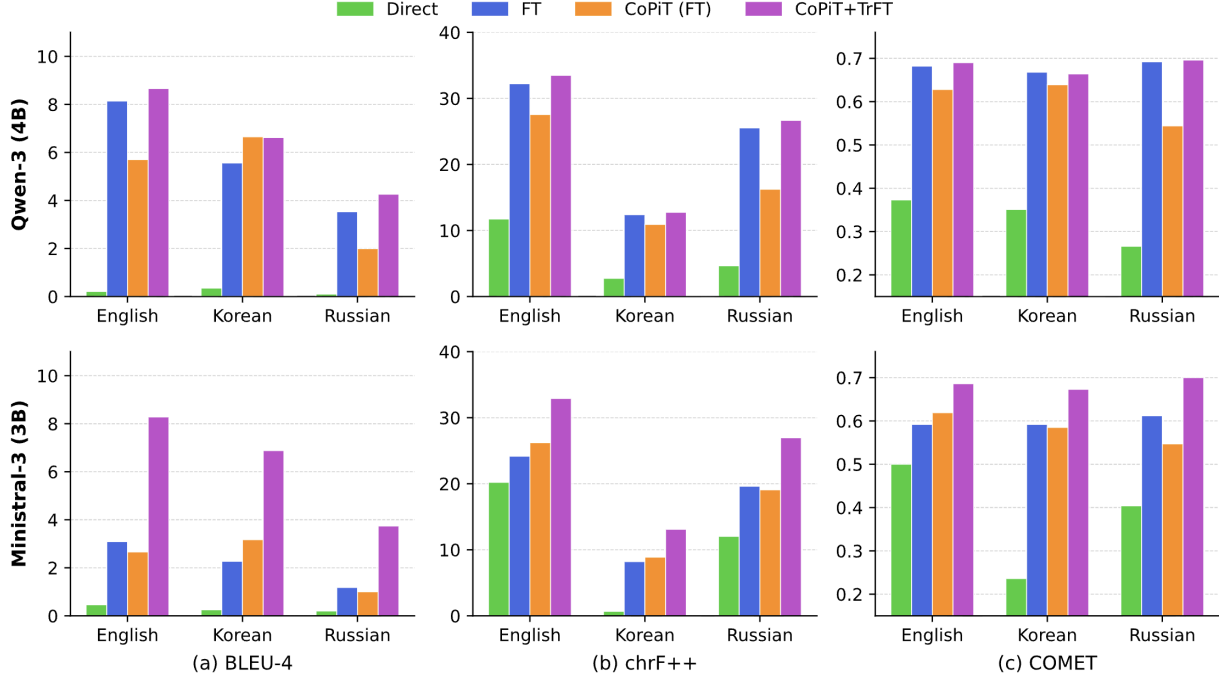


Figure 9: Full results across BLEU-4, chrF++, and COMET under pipeline configurations with and without translation module fine-tuning, where TrFT denotes fine-tuning of the translation module on synthetic parallel data.

operates outside the original low-resource problem setting, in which no Traditional Mongolian–Target parallel data is assumed to be available, and therefore represents an upper-bound analysis of what becomes possible once the pipeline itself is used to generate synthetic resources.

A.9 Prompt Templates

This appendix presents the prompts used in our experiments. Figures 10–14 show the prompts used in each stage of the *CoPiT*, while Figure 15 presents the prompt used for the direct translation baseline.

```

system: |-
You are an expert in Traditional Mongolian vowel harmony.
Your task is to classify the vowel harmony type of a single traditional Mongolian word.
Do not provide explanations. Output must follow the required format.

user: |-
### Essential Vowel Harmony Rules
1. Alphabet Structure
- Traditional Mongolian letters change visual shape depending on position (Initial / Medial / Final).
- These shape differences do not change pronunciation. Harmony classification is based solely on vowel identity, not shape.

2. Vowel Harmony Classes
- Masculine:
  - Cyrillic: а, о, у
  - Traditional: ᠠ, ᠨ, ᠣ
  - Rule: Masculine vowels cannot mix with feminine vowels in the same word.
- Feminine:
  - Cyrillic: э, ө, ү
  - Traditional: ᠡ, ᠭ, ᠦ
  - Rule: Feminine vowels cannot mix with masculine vowels in the same word.
- Neutral:
  - Cyrillic: и
  - Traditional: ᠢ
  - Rule: Neutral vowels can appear with either masculine or feminine vowels.

3. Medial Vowels
- Orthographic Shape vs. Pronunciation:
  - Medial forms of ᠠ and ᠡ look the same, medial forms of ᠨ and ᠭ look similar.
  - Therefore, vowel pronunciation and class are determined by the first vowel, not by later medial shapes.
- Pronunciation behavior:
  - If the word is masculine, medial vowels are pronounced as a/o/u.
  - If the word is feminine, medial vowels are pronounced as e/ö/ü.

4. Classification Procedure
- Identify the first vowel of the word.
- That first vowel determines the entire harmony class of the word.
- If the word's first vowel is the neutral vowel look at following vowels and follow their harmony types.
- If a word contains only ᠢ (the neutral vowel), classify it as Feminine.

### Task
Using the rules above, classify the vowel harmony class of the following traditional Mongolian word.

- Traditional Mongolian Word: {trad_word}

### Output Format
Output ONLY one of the following classes, with NO additional text:
- Masculine
- Feminine

### Output

```

Figure 10: Prompt used for vowel harmony recovery

system: |-
 You are an expert in modern Cyrillic Mongolian orthography.
 Your task is to convert the Latin transliteration of a Traditional Mongolian word into the correct modern Cyrillic form.
 Do not provide explanations. Output must follow the required format.

user: |-
 ### Essential Cyrillic Conversion Rules (Latin transliteration letter: Cyrillic letter)
 1. Map each long vowel pattern as follows:
 aga/iga/uga: aa, ege/ige/üge/öge: ээ, igi: ий, ugu/aga/iga/oga/uga: oo, agu/ugu/igu: yy, öge/ege/ögü: өө, egü/ügü/igü: үү, oo: oo, öö: өө, uu/au/iu: yy, üü/eü/iü: үү, agaga: aa, egege: ээ, eguu: үү

2. Map each Vowel+y+Vowel pattern as follows:
 iya: aa/иа/ийа, iye: ие/иэ

3. Map each diphthong as follows:
 ai: ай, ei: эй, oi: ой, ui: уй, üi: үй, au: ay, ua: ya

4. Map each single vowel as follows:
 a: a, e: э, i: и, o: o, u: y, ö: ө, ü: ү

5. Map each consonant as follows:
 b: б, g: г, γ: г, q: q/h: x, t: т, d: д, n: н, ng: нг/н, m: м, l: л, r: p, s: c, š: ш, z: э, j: ж, č: ч, c: ц, y: й

Task
 Using the rules above, convert the following Latin transliteration result into its correct Cyrillic word.
 Use the vowel harmony class only to resolve ambiguous vowel/consonant mapping.

- Traditional Mongolian Word: {trad_word}
 - Transliteration: {latin_trans}

- Vowel Harmony Class: {vowel_harmony}

Output Format
 Output ONLY the final Cyrillic word, with NO other scripts and NO additional text.

Output

Figure 12: Prompt used for Cyrillic normalization

```
system: |
You are an expert editor of Mongolian writing.
Your task is to revise the given sentence only if it is unnatural, ungrammatical, or incorrect.
Do not add new information, remove any content, or reinterpret the sentence.

user: |
## Task
Review the following Mongolian sentence and produce the most natural and grammatically correct version in standard Mongolian.

### Revising Rules
- Correct grammar, spelling, and orthography only where needed.
- Normalize casing when the sentence is entirely uppercase or entirely lowercase.
- Preserve the original meaning strictly; do NOT add, omit, or reinterpret content.

### Sentence
{cyrillic_sent}

### Output Format
Output ONLY the final corrected Mongolian sentence.
Do NOT include explanations, reasoning, or any additional text.

### Output
```

Figure 13: Prompt used for self-reflection

toEnglish:
system: |
You are an expert in Mongolian translation.
Your task is to translate the Cyrillic Mongolian sentence to English.
Do not provide explanations. Output must follow the required format.

user: |
Task
Translate the following Mongolian sentence to English.

- Mongolian sentence: {cyrillic_sent}

Output Format
Output ONLY the final translation result, with NO other scripts and NO additional text.

Output

toKorean:
system: |
You are an expert in Mongolian translation.
Your task is to translate the Cyrillic Mongolian sentence to Korean.
Do not provide explanations. Output must follow the required format.

user: |
Task
Translate the following Mongolian sentence to Korean.

- Mongolian sentence: {cyrillic_sent}

Output Format
Output ONLY the final translation result, with NO other scripts and NO additional text.

Output

toRussian:
system: |
You are an expert in Mongolian translation.
Your task is to translate the Cyrillic Mongolian sentence to Russian.
Do not provide explanations. Output must follow the required format.

user: |
Task
Translate the following Mongolian sentence to Russian.

- Mongolian sentence: {cyrillic_sent}

Output Format
Output ONLY the final translation result, with NO other scripts and NO additional text.

Output

Figure 14: Prompt used for translation

toEnglish:
system: |
You are an expert in Mongolian translation.
Your task is to translate the traditional Mongolian sentence to English.
Do not provide explanations. Output must follow the required format.

user: |
Task
Translate the following Mongolian sentence to English.

- Mongolian sentence: {trad_sent}

Output Format
Output ONLY the final translation result, with NO other scripts and NO additional text.

Output

toKorean:
system: |
You are an expert in Mongolian translation.
Your task is to translate the traditional Mongolian sentence to Korean.
Do not provide explanations. Output must follow the required format.

user: |
Task
Translate the following Mongolian sentence to Korean.

- Mongolian sentence: {trad_sent}

Output Format
Output ONLY the final translation result, with NO other scripts and NO additional text.

Output

toRussian:
system: |
You are an expert in Mongolian translation.
Your task is to translate the traditional Mongolian sentence to Russian.
Do not provide explanations. Output must follow the required format.

user: |
Task
Translate the following Mongolian sentence to Russian.

- Mongolian sentence: {trad_sent}

Output Format
Output ONLY the final translation result, with NO other scripts and NO additional text.

Output

Figure 15: Prompt used for direct translation