Kyrgyz Text Normalization: A Comparative Study of Neural and Rule-Based Approaches

Zarina Uvalieva*, **Bektemir Kumarbai uulu, Adilet Metinov, Tynchtykbek Tashbaltaev, Nurtilek Alibekov**
Airun Intelligence Lab, Bdigital LLC
uv@bdigital.kg, zarina.uvalievaa@gmail.com

Abstract

Text normalization—the task of converting noisy, informal text into a standardized form—is a fundamental preprocessing step for many NLP applications. Despite the growing need for Kyrgyz language processing tools, to the best of our knowledge, no prior work has addressed automatic text normalization for Kyrgyz, a morphologically rich, low-resource Turkic language. In this paper, we present the first systematic study of Kyrgyz text normalization. We collect a dataset of 1.67 million noisy–clean text pairs sourced from YouTube comments, Instagram posts, and Telegram channels, where users frequently write without punctuation, capitalization, or standard spelling. Pairs were annotated with Gemini 3 Pro; the 1,000-example test set was fully verified by two native Kyrgyz speakers with adjudication, and a random subset of the training data was spot-checked, while the full 1.67M training set was not verified exhaustively. For continual pre-training, we additionally use a 538 MB Kyrgyz corpus compiled from news portals and books. We evaluate five systems: a rule-based baseline, zero-shot mT5, a fine-tuned mT5-small model, a continually pre-trained mT5-small followed by fine-tuning, and zero-shot Gemma 4. Our experiments show that fine-tuned mT5-small achieves a CER of 0.0796, outperforming the rule-based baseline (CER 0.2029), zero-shot mT5 (CER 0.9887), and zero-shot Gemma 4 (CER 0.1620), a roughly 32× larger model in a fine-tuned vs. zero-shot setting. Human evaluation by two native Kyrgyz speakers confirms these results, with fine-tuned mT5-small rated as correct in 99.8% of cases. We further analyze why continual pre-training with span corruption does not improve over direct fine-tuning, finding hallucination in 35/40 of the inspected failure cases (87.5%, 95% Wilson CI [74%, 95%]).

* Corresponding author: uv@bdigital.kg

1 Introduction

The Kyrgyz language is spoken by approximately 4.5 million people, primarily in Kyrgyzstan, and is closely related to other Turkic languages such as Kazakh and Uzbek. Despite its significant speaker population, Kyrgyz remains severely underrepresented in NLP research. One of the most common challenges when processing Kyrgyz text from the web is the lack of standardization: users on social media platforms, YouTube, and messaging apps frequently write without punctuation, capitalization, or correct spelling. This informal, noisy text is difficult to process directly with downstream NLP tools such as machine translation, speech synthesis, or information retrieval systems.

Text normalization—converting such noisy text into a clean, standardized form—is therefore an essential first step for building robust Kyrgyz NLP pipelines. Downstream tasks such as machine translation, named entity recognition, and speech synthesis all benefit significantly from normalized input. However, to the best of our knowledge, no prior work has specifically addressed this task for Kyrgyz, leaving a critical gap in the NLP infrastructure for this language.

In this work, our comparison contrasts a fine-tuned mT5 sequence-to-sequence model against an intentionally minimal rule-based lower bound and two zero-shot baselines (zero-shot mT5 and zero-shot Gemma 4). Stronger comparisons—in particular a finite-state-transducer-based rule pipeline tailored to Kyrgyz (Washington et al., 2012), a byte-level neural model such as ByT5 (Xue et al., 2022), and fine-tuned or few-shot variants of large open LLMs—are important and explicitly deferred to future work; the goal of the present study is to establish a first reproducible benchmark and a strong, deployable small-model baseline rather than to map out the full design space. With that scope in mind, we make the following contribu-

Proceedings of the 1st Workshop on Multilinguality in the Era of Large Language Models (MeLLM 2026), pages 52–62

tions:

1. We construct the first large-scale dataset for Kyrgyz text normalization, containing 1.67 million noisy–clean text pairs collected from YouTube, Instagram, and Telegram, annotated with Gemini 3 Pro; the test set is fully human-verified and a random subset of the training data has been spot-checked.

2. We conduct a systematic comparison of five normalization systems: a rule-based baseline, zero-shot mT5, fine-tuned mT5-small, continually pre-trained mT5-small followed by fine-tuning, and Gemma 4 in a zero-shot setting.

3. We demonstrate that a fine-tuned mT5-small model significantly outperforms all baselines, including the substantially larger Gemma 4 model in a zero-shot setting.

4. We analyze the failure modes of continual pre-training with span corruption, finding that hallucination is the dominant cause of degradation.

5. We quantify reference bias on a 50-example probe with an independent human annotator, showing that fine-tuned mT5's CER changes by only 0.012 and system ranking is preserved—evidence that our main conclusions are robust to the choice of reference.

Our code, fine-tuned model checkpoints, and a 20,000-pair subset of the training data together with the full human-verified test set are publicly available.[1]

2 Related Work

Text Normalization for Turkic Languages. The closest related work is on Turkish text normalization, which shares morphological and structural similarities with Kyrgyz. Torunoğlu and Eryiğit (2014) proposed a cascaded approach for normalizing Turkish social media text, classifying errors into seven categories. Çolakoğlu et al. (2019) applied neural machine translation approaches to normalize non-canonical Turkish text. Koksal et al. (2020) introduced a benchmark dataset for Turkish text correction on Twitter. For the other Central Asian Turkic languages most closely related to Kyrgyz—Kazakh, Uzbek, and Tatar—we are not aware of a publicly available, social-media-scale normalization dataset comparable to ours, although morphological analyzers and spelling resources exist.

Text Normalization for Low-Resource Languages. Several recent works have applied multilingual pre-trained models to normalization in low-resource settings. Zupon et al. (2021) studied text normalization for eight African languages using sequence-to-sequence models. Lutgen et al. (2025) demonstrated the effectiveness of mT5 and ByT5 for normalizing Luxembourgish, a morphologically rich low-resource language.

Kyrgyz NLP. Prior work on Kyrgyz NLP has been limited. Washington et al. (2012) developed a finite-state morphological transducer for Kyrgyz. Alekseev et al. (2023) introduced a benchmark for multilabel topic classification in Kyrgyz. To our knowledge, no prior work has addressed text normalization for Kyrgyz.

Multilingual Pre-trained Models. mT5 (Xue et al., 2021) is a massively multilingual text-to-text model trained on 101 languages, including Kyrgyz. Its encoder-decoder architecture makes it well-suited for sequence-to-sequence tasks such as text normalization.

Character- and Byte-Level Encoders. Because normalization is largely a character-level task (punctuation insertion, case changes, diacritic and digit–word adjustments), token-free or byte-level models are a natural alternative to subword-based mT5. ByT5 (Xue et al., 2022) operates directly on UTF-8 bytes and has been shown to be competitive with or better than mT5 on noisy, morphologically rich, or spelling-sensitive inputs; Lutgen et al. (2025) report favorable ByT5 results for Luxembourgish normalization. We leave a systematic ByT5 comparison for Kyrgyz to future work and focus here on mT5 as a strong, widely used baseline.

3 Dataset

3.1 Data Collection

We collected noisy Kyrgyz text from three social media sources: YouTube comments (45%), Insta-

[1] Code: https://github.com/Zarina33/Kyrgyz-Text-Normalization-Conference. Models: https://huggingface.co/Zarinaaa/mt5-small-kyrgyz-normalization, https://huggingface.co/Zarinaaa/mt5-small-kyrgyz-normalization-ptft. Dataset: https://huggingface.co/datasets/Zarinaaa/kyrgyz-text-normalization.

gram posts and comments (25%), and Telegram channel messages (30%). These platforms were chosen because Kyrgyz-speaking users write informal text without punctuation, capitalization, or standard orthography, making them a rich source of naturally occurring noisy text. For continual pretraining (§4.4), we additionally compiled a separate 538 MB Kyrgyz text corpus from news portals and books, representing clean formal text.

3.2 Data Annotation

Each noisy comment was paired with a normalized version using Gemini 3 Pro as an automatic annotation tool. Normalization targets include: (1) restoring correct punctuation, (2) fixing capitalization, (3) correcting dialectal and non-standard spellings, and (4) standardizing orthographic variants (e.g., digit–word compounds such as 8жыл $\rightarrow$ 8 жыл).

Verification. Verification was carried out at two levels. *(i) Full test-set review:* all 1,000 test pairs were independently reviewed by two native Kyrgyz speakers; disagreements on reference text were adjudicated in a third pass (conducted by one of the authors) in which a single reference string was fixed per example before any system evaluation. This adjudication step applies only to the construction of the reference test set and is distinct from the independent rating procedure used in the human evaluation of system outputs (§5.3), which deliberately has no arbitration step. *(ii) Training spot-check:* we manually reviewed 400 randomly sampled training pairs from the full 1.67M set. Of these, 336 (84%) were judged to be valid normalizations requiring no further edits, while 64 (16%) contained issues such as minor punctuation choices, partial errors, or over-correction. At $N = 400$, the 84% acceptance rate has a 95% Wilson confidence interval of approximately $[80\%, 87\%]$. We explicitly do not claim that the full 1.67M training set was verified exhaustively—such a claim would be implausible given the data size—and we report the training set as *Gemini-annotated with a 400-example human spot-check*. The implications of this residual label noise (reference bias, noisy supervision) are discussed in the Limitations section.

3.3 Dataset Statistics

The final dataset contains **1,673,715** noisy–clean text pairs. We split the data into a training set (1,672,715 pairs) and a test set (1,000 pairs, held out before training). A representative subset of 20,000 training pairs and the full test set are publicly available.[2] We release only a 20,000-pair subset of the training data rather than the full 1.67M pairs for three reasons: (i) the source platforms (YouTube, Instagram, Telegram) impose restrictions on bulk redistribution of user-generated content; (ii) releasing the full set at scale would heighten the risk of re-identifying individual users from comment content even after handle removal; and (iii) only a sample of the training pairs has been human-verified, so we prefer to release a smaller subset that we can stand behind qualitatively. We additionally release all training code, fine-tuned model checkpoints, and the fully human-verified test set so that reported results are reproducible end-to-end. Specifically, the released checkpoints together with the full test set are sufficient to reproduce all numbers in Tables 3, 4, 7, and 11 *exactly*, without retraining; the 20,000-pair subset enables small-scale fine-tuning experiments and analyses on a representative slice of the training distribution, but cannot reproduce the full 1.67M-pair training condition. Future comparisons that retrain on a smaller subset should report subset size alongside results and treat numbers obtained on subsets as not directly comparable to ours.

Split	Examples	Avg. Input	Avg. Target
Train	1,672,715	131.5	136.2
Test	1,000	135.8	140.6

Table 1: Dataset statistics. Lengths are in characters.

3.4 Normalization Analysis

We analyzed the types of normalization required across the dataset. As shown in Figure 1 and Table 2, the dominant transformation is **punctuation restoration**, required in 84.9% of test examples. Nearly all examples (99.8%) differ between input and target, confirming that the dataset captures real normalization needs.

4 Systems

4.1 Rule-Based Baseline

Our rule-based system applies three transformations: (1) capitalizing the first character, (2) collapsing multiple whitespace characters, and (3) ap-

[2] https://huggingface.co/datasets/Zarinaaa/kyrgyz-text-normalization

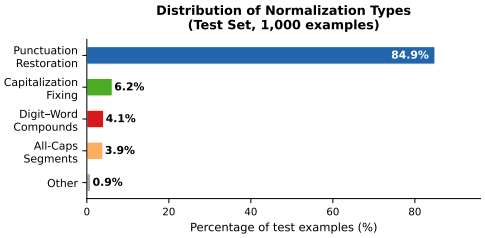

Figure 1: Distribution of normalization types in the test set (1,000 examples).

Normalization Type	Train (%)	Test (%)
Punctuation restoration	84.5	84.9
Capitalization fixing	9.6	6.2
All-caps segments	3.0	3.9
Digit–word compounds	3.9	4.1
Other	0.5	0.9
Input $\neq$ Target	99.7	99.8

Table 2: Types of normalization required in the dataset. Categories are not mutually exclusive; one example may require multiple types.

pending a period if the input lacks sentence-final punctuation. Characters following sentence-final punctuation are also capitalized. This baseline is intentionally minimal and is meant to serve as a *lower bound*, isolating how much of the test-set CER can be recovered by trivial orthographic heuristics. Stronger rule-based pipelines—for example, the finite-state morphological transducer of Washington et al. (2012) combined with dictionary-based substitution of frequent dialectal forms and digit–word compounds—would likely close a substantial portion of the gap to fine-tuned mT5, and we leave such a competitive rule-based system to future work.

4.2 Zero-Shot mT5

We evaluate the pre-trained `google/mt5-small` model (Xue et al., 2021) without any task-specific fine-tuning, prompted with the prefix `"correct: "` followed by the input text.

4.3 mT5 Fine-Tuned

We fine-tune `google/mt5-small` on our training set using the prefix `"correct: "`. Training details: effective batch size 64 (physical batch size 4, gradient accumulation over 16 steps), learning rate 3×10^{-4} with cosine schedule and 500 warmup steps, 5 epochs, random seed 42, on a single NVIDIA RTX 5080 (16 GB VRAM). Training data are split 95/5 into train/validation, and the checkpoint with the lowest validation loss is used for test-set evaluation. The test set is held out before any training and is disjoint from both the train and validation splits. We additionally verified explicitly that no noisy input from the 1,000-example test set appears in the 1.67M training set (0/1,000 exact-match overlap, also 0/1,000 under case-insensitive matching), so the reported test CER is not inflated by train–test leakage. We chose mT5-small (300M parameters) as it provides the best trade-off between model capacity and training efficiency for our dataset size; larger variants such as mT5-base and mT5-large have significantly higher computational requirements and would benefit from substantially more training data than is currently available for Kyrgyz.

4.4 mT5 Continually Pre-Trained + Fine-Tuned

We first apply continual pre-training of `google/mt5-small` on a 538 MB Kyrgyz corpus compiled from news portals and books using T5-style span corruption (mask rate 0.15, mean span length 3) for 3 epochs, with a 98/2 train/validation split and random seed 42; the checkpoint with the lowest validation loss is retained. We then fine-tune the resulting model using the same procedure as §4.3.

4.5 Gemma 4 Zero-Shot

We evaluate Gemma 4 (`gemma4:e4b`, 9.6 B parameters, 4-bit quantized) in a zero-shot setting via Ollama.[3] The prompt instructs the model to normalize the input text and return only the corrected output.

5 Experiments

5.1 Evaluation Metrics

We evaluate all systems using three automatic metrics: **CER** (Character Error Rate), **WER** (Word Error Rate), and **EM** (Exact Match). Lower CER and WER indicate better performance; higher EM is better.

5.2 Automatic Evaluation Results

Fine-tuned mT5-small achieves the best performance across all metrics. All differences against Fine-Tuned are statistically significant at $p < 0.001$ *except* Pre-Train+FT, which yields $p = 0.06$.

[3]We use an English-language prompt as it yielded more consistent outputs than a Kyrgyz-language prompt in preliminary experiments.

System	CER↓	WER↓	EM↑	vs. FT
Rule-Based	0.2029 ±0.006	0.5659	0.0040	$p<0.001^{***}$
Zero-Shot mT5-small	0.9887	0.9981	0.0000	$p<0.001^{***}$
Gemma 4 Zero-Shot	0.1620 ±0.004	0.4320	0.0150	$p<0.001^{***}$
mT5-small Fine-Tuned	**0.0796** ±0.003	**0.1978**	**0.1860**	—
mT5-small Pre-Train+FT	0.0825 ±0.004	0.2017	0.1840	$p=0.06$

Table 3: Automatic evaluation results on the test set (1,000 examples). CER ± std. from bootstrap resampling (n=10,000); Zero-Shot mT5 bootstrap std. omitted (near-degenerate outputs). Column "vs. FT" reports paired bootstrap *two-sided* p-values against mT5-small Fine-Tuned (10,000 resamples). Zero-Shot mT5 is retained as a sanity check to show that mT5 without task-specific fine-tuning does not perform the task; it is not intended as a competitive baseline. Best results in bold.

We emphasize that $p = 0.06$ is *insufficient evidence to reject the null of no difference* at the conventional threshold, not a positive demonstration of equivalence; with $n = 1{,}000$ test examples we cannot rule out a small real effect in either direction, and we therefore do not claim that the two fine-tuned variants are statistically indistinguishable. Fine-tuned mT5-small substantially outperforms zero-shot Gemma 4 in CER (0.0796 vs. 0.1620), despite Gemma 4 having roughly 32× more parameters. This should be read as a *fine-tuned vs. zero-shot* comparison rather than a direct capacity comparison: a fine-tuned or few-shot Gemma 4 could plausibly close or reverse this gap, and we leave such a comparison to future work.

5.3 Human Evaluation

To complement automatic metrics, we conducted a human evaluation on 200 randomly sampled test examples. Two native Kyrgyz-speaking annotators independently rated each system output as correct (1) or incorrect (0). Annotators were fluent adult speakers of Kyrgyz recruited from the authors' professional network; they participated as uncompensated volunteers after being informed of the study purpose and data source, and were free to withdraw at any time. No formal IRB review was conducted for this rating study, which we consider minimal-risk: annotators reviewed short, publicly available social-media text snippets and provided binary quality judgements only. The sample size of 200 per system was chosen to be consistent with prior low-resource NLP human evaluation (Lutgen et al., 2025) and to remain tractable for careful volunteer review. Instructions were provided in Kyrgyz and asked annotators to mark an output as correct if it reads as natural, well-punctuated Kyrgyz with no significant spelling or grammatical errors; outputs with repetitions, hallucinations, or unnatural phrasing were marked as incorrect. Annotators worked independently without seeing each other's ratings and did not confer on individual examples. Zero-Shot mT5-small was excluded from human evaluation, as its outputs consist entirely of sentinel tokens (e.g., `<extra_id_0>`) and are not suitable for meaningful annotation. The mean human scores are reported in Table 4.

System	Score↑	95% Wilson CI
Rule-Based	0.500	[0.451, 0.549]
Gemma 4 Zero-Shot	0.793	[0.750, 0.829]
mT5-small Fine-Tuned	**0.998**	[0.986, 0.9996]
mT5-small Pre-Train+FT	**0.998**	[0.986, 0.9996]

Table 4: Human evaluation results (mean score over 2 annotators, 200 examples; $n = 400$ ratings per system). Score = proportion of outputs rated as correct. Wilson 95% confidence intervals computed over the pooled 400 ratings.

Human evaluation strongly confirms the automatic results. Fine-tuned mT5-small was rated as correct in **99.8%** of cases, compared to 79.3% for Gemma 4 and only 50.0% for the rule-based baseline. The 95% confidence intervals for the two fine-tuned variants do not overlap with those of Gemma 4 or the rule-based baseline, indicating that the qualitative ranking is robust to sampling variance at this scale, even though we cannot resolve the fine-tuned variants from each other at the 99.8% ceiling.

Why is human accuracy (99.8%) so much higher than EM (18.6%)? This gap is expected and is consistent with the reference-bias analysis (§5.4). EM is a strict character-level match against a single fixed reference, so it is penalized by any deviation, including legitimate alternatives that a native speaker would still rate as correct—e.g., a comma placed before vs. after a discourse particle, an em-dash vs. comma, or one of several acceptable spellings of a borrowed word. The reference-

agreement CER of 0.12 between two valid Kyrgyz references (Table 7) provides a direct measure of this surface-form variability: any single fixed reference captures only one of several valid normalizations, so EM is a lower bound on output quality and human accuracy is a complementary, less surface-sensitive estimate. We can quantify this directly on the human evaluation data: of the 199 fine-tuned mT5 outputs that both annotators agreed were correct, **162 (81.4%)** are not character-identical to the Gemini reference. A representative example is shown in Table 5: the model output differs from the reference in comma placement (after болбой vs. before өзү) and final punctuation (? vs. .), yet both annotators rated it correct. The gap should therefore be read as a property of the metric pair under reference variability, not as an inconsistency in the evaluation.

Input	Эркеги ит экенда жаман катындай болбой ОЗУ эле чечпейби
Ref	Эркеги ит экен да. Жаман катындай болбой өзү эле чечпейби.
FT output	Эркеги ит экен да. Жаман катындай болбой, өзү эле чечпейби?
A1, A2	both rated *correct*

Table 5: Representative case where the fine-tuned mT5 output is rated correct by both annotators but differs from the reference (comma placement, final punctuation). 162 of 199 both-rated-correct FT outputs (81.4%) deviate from the reference at the character level.

Interpreting $\kappa = 0.25$. Inter-annotator agreement was measured using Cohen's κ, yielding an overall pooled value of $\kappa = 0.25$. This number is not in itself a verdict on the quality of the evaluation. Cohen's κ is well known to collapse when the marginal distribution is highly skewed: when one category (here, "correct") dominates, the expected-agreement term is large, and even near-perfect observed agreement produces a low κ. Feinstein and Cicchetti (1990) formalize this as the *first kappa paradox*. To give a more complete picture of reliability we additionally report PABAK and Gwet's AC1, both of which are known to be more robust than Cohen's κ under prevalence skew (Table 6).

The pattern is exactly what the kappa-paradox literature predicts. On the fine-tuned systems, percent agreement is at ceiling (99.5%) and both annotators rate essentially every output as correct; the marginals are so skewed that Cohen's κ collapses to zero, while PABAK and Gwet's AC1 correctly report near-perfect agreement (≥ 0.99).

System	%agr.	κ	PABAK	AC1
Rule-Based	48.0	0.012	−0.040	−0.040
Gemma 4	62.5	−0.030	0.250	0.441
mT5 Fine-Tuned	99.5	0.000	0.990	0.995
mT5 PT+FT	99.5	0.000	0.990	0.995
Overall (pooled)	77.4	0.246	0.548	0.680

Table 6: Per-system inter-annotator agreement on the human evaluation ($n = 200$ examples, 2 annotators). %agr. = percent example-level agreement; κ = Cohen's kappa; PABAK = prevalence-adjusted and bias-adjusted kappa $= 2 \cdot p_{obs} - 1$; AC1 = Gwet's first-order agreement coefficient. Under heavy prevalence skew (fine-tuned systems, ≈99.5% observed agreement), κ collapses to near zero by the *first kappa paradox*; PABAK and AC1 correctly indicate near-perfect agreement.

On Gemma 4, where the prevalence is more balanced, AC1 = 0.44 indicates moderate agreement, while $\kappa = -0.03$ is again misleadingly pessimistic due to a small marginal mismatch between annotators. On the Rule-Based system, all three metrics agree that reliability is low: annotators differ substantively on which marginal outputs (partial punctuation, trailing periods) count as correct. We therefore read the human evaluation as well-supported on the fine-tuned vs. non-fine-tuned ranking (where PABAK and AC1 are high), and we explicitly flag low reliability on absolute Rule-Based scores.

Scale of the study. The human evaluation involves two annotators, 200 examples per system, and does not include a third-party arbitration step. It should therefore be read as a *consistency check* on the automatic metrics—providing evidence that the large CER gap between the fine-tuned models and the other systems is perceived by native speakers—rather than as a definitive, high-powered assessment of absolute output quality. A larger study with additional annotators and arbitration would be needed to give tight confidence intervals on the 99.8% number, and we flag this as a limitation.

5.4 Reference Bias Analysis

Because the test references were constructed on top of Gemini 3 Pro outputs (§3.2), the reported CER gap could in principle reflect *Gemini-style* rather than *objective correctness*. To quantify this effect, we asked a second native Kyrgyz speaker—not involved in the original test-set review—to

write independent normalization references *from scratch* for 50 randomly selected test examples (indices 100–149), without access to either the Gemini references or any system output, and recomputed CER against both reference sources (Table 7).

Three observations follow. *(1) There is a non-trivial reference noise floor:* the two valid references disagree at CER = 0.1200, so a CER below this value with respect to one reference means the system output is closer to that reference than the references are to each other. *(2) Fine-tuned mT5 is not inflated by Gemini-style bias:* its CER changes by only 0.012 when switched to the independent reference (0.0913 $\rightarrow$ 0.0793), and 0.0793 is *below* the 0.1200 noise floor, suggesting the model has learned a general normalization function rather than matching Gemini-specific stylistic choices. *(3) Simpler systems are evaluated more harshly against Gemini:* Rule-Based and zero-shot Gemma 4 show larger drops (-0.09, -0.07) against the independent reference, indicating that Gemini carries systematic preferences (comma placement, dialect forms) that a fresh annotator does not fully share. Crucially, system ranking (Rule-Based $<$ Gemma 4 $<$ fine-tuned variants) is preserved under both references, so our main conclusions are robust to this choice. With $N = 50$ and a single independent annotator we cannot rule out sampling variance in any individual estimate, but the consistent pattern across all five systems—simpler, Gemini-mismatched systems showing the largest Δ—suggests a systematic rather than incidental effect; a fuller study would use a larger independent set and multiple annotators.

6 Analysis

6.1 Why Does Continual Pre-Training Hurt?

Despite our expectation that continual pre-training would improve normalization, the Pre-Train+FT model performs numerically worse than direct fine-tuning (CER 0.0825 vs. 0.0796; $p = 0.06$, which we treat as insufficient evidence to reject the null, not as equivalence). As a qualitative probe into *what kinds* of errors drive any numerical gap, we analyzed 40 examples where fine-tuned mT5 outperforms Pre-Train+FT by more than 0.05 CER and categorized the error types. This sample is deliberately small and is intended to surface failure *modes*, not to estimate their population-level prevalence.

Hallucination is the dominant failure mode within this inspected sample (35/40, 87.5%, 95% Wilson CI $[74\%, 95\%]$): the pre-trained model repeats or copies fragments of the input text, producing outputs that are longer than expected and diverge from the reference. Tables 9 and 10 show representative examples.

We consider two non-exclusive hypotheses for this behavior. *(H1) Copy bias from span corruption.* T5-style span corruption trains the decoder to reconstruct spans of the input verbatim, which may reinforce a tendency to copy and repeat input content; this is benign for generic language modeling but directly harmful for a normalization objective whose target is usually shorter than—or at least not a superset of—the input. *(H2) Distributional shift between pre-training and fine-tuning.* Our continual pre-training corpus consists of *clean, formal* Kyrgyz from news portals and books, whereas the fine-tuning targets are normalized versions of *noisy, informal* social-media text. The register gap between the two stages may push the model toward producing fluent formal continuations (which look like hallucinations with respect to the input) rather than minimal normalization edits. We do not decisively separate H1 from H2 here and flag this as a direction for future work (e.g., continual pre-training on in-domain noisy text, or on denoising objectives closer to the normalization target). The two models produce similar outputs in 92.5% of examples, confirming that the difference is small and concentrated in specific failure modes. In 3.5% of examples, Pre-Train+FT is numerically better than Fine-Tuned mT5, with a few cases involving number normalization (e.g., 14жыл $\rightarrow$ Он төрт жыл) or word reordering; as discussed in §6.5 at the category level, the sample sizes are too small to treat these as robust advantages.

6.2 Zero-Shot mT5 Failure

Zero-shot mT5-small fails catastrophically at the normalization task (CER 0.9887, i.e., nearly total character-level disagreement with the reference). Inspection of the outputs reveals that the model almost exclusively produces sentinel tokens (e.g., `<extra_id_0>`), which are artifacts of its span-corruption pre-training objective. Without task-specific fine-tuning, the model has no understanding of the normalization task and defaults to its pre-training output format. This underscores the necessity of fine-tuning even small models for low-resource languages.

System	CER (Gemini ref.)	CER (Independent ref.)	Δ
Rule-Based	0.2129	0.1217	−0.0912
Zero-Shot mT5-small	0.9858	0.9865	+0.0007
Gemma 4 Zero-Shot	0.1693	0.1023	−0.0670
mT5-small Fine-Tuned	0.0913	**0.0793**	−0.0120
mT5-small Pre-Train+FT	0.0836	0.0806	−0.0029
Reference-agreement CER (Gemini reference vs. independent reference): 0.1200			

Table 7: CER of each system evaluated against the Gemini-based test reference vs. an independent from-scratch reference ($N = 50$ examples, test indices 100–149). Δ = CER(Indep.) − CER(Gemini). The *reference-agreement CER* (last row) is computed as CER(Gemini reference; independent reference), i.e. treating the Gemini reference as the hypothesis and the independent reference as the reference; it provides a noise floor for this evaluation.

Error Type	Count	%†
Hallucination (repetition)	35	87.5
Punctuation errors	8	20.0
Over-correction	2	5.0
Other	4	10.0

†Multiple labels per example ($\Sigma > 100\%$).

Table 8: Error categories in cases where Pre-Train+FT underperforms Fine-Tuned mT5 (40 examples).

Input	Фергана ороонунун тору Аксы Ала бука эн кооз аймактар.
Ref	Фергана өрөөнүнүн төрү Аксы, Ала-Бука эң кооз аймактар.
FT	Фергана өрөөнүнүн төрү Аксы, Ала-Бука эң кооз аймактар.
PT+FT	Фергана өрөөнүнүн төрү Аксы, Ала-Бука эң кооз аймактар. Фергана өрөөнүнүн төрү, Аксы... *(repeated)*

Table 9: Hallucination: Pre-Train+FT repeats the entire sentence.

6.3 Gemma 4 vs. Fine-Tuned mT5

Zero-shot Gemma 4 (CER 0.1620) substantially outperforms both the rule-based baseline and zero-shot mT5, suggesting that large multilingual LLMs have acquired some knowledge of Kyrgyz during pre-training. Fine-tuned mT5-small (300M parameters) in turn outperforms zero-shot Gemma 4 (9.6B parameters) by roughly a factor of 2× in CER. We stress that this is a *fine-tuned vs. zero-shot* comparison, not a direct capacity comparison: it shows that task-specific fine-tuning on a domain-matched dataset is highly effective for low-resource normalization (Zupon et al., 2021), but it does *not* show that a 300M model is fundamentally more capable than a 9.6B one at this task. A fine-tuned or few-shot Gemma 4 could plausibly match or surpass fine-tuned mT5-small, and we explicitly flag this comparison as future work. The practical implication remains relevant: in deployment-constrained low-resource settings, a 300M fine-tuned model is a strong, cheap, and reproducible option.

Input	Подезд журуп башка иштерди кылганга билимдин деле кереги жок
Ref	Кире бериштерди жууп, башка иштерди кылганга билимдин деле кереги жок.
FT	Подъезд жүрүп башка иштерди кылганга билимдин деле кереги жок.
PT+FT	Подъезд жүрүп башка иштерди кылганга билимдин деле кереги жок. Билимдин деле кереги жок. *(repeated)*

Table 10: Hallucination: Pre-Train+FT appends a repeated phrase.

6.4 Rule-Based Baseline Analysis

The rule-based system achieves a surprisingly competitive CER (0.2029) given its simplicity. This is directly explained by the dataset statistics: punctuation restoration accounts for 84.9% of all required normalizations (§3.4), and the rule-based system partially addresses this by appending sentence-final punctuation. However, it fails entirely on word-level corrections, dialect normalization, and sentence restructuring, resulting in poor WER (0.5659) and near-zero EM (0.0040). This highlights that while punctuation is the most frequent normalization type, word-level corrections are equally important for overall text quality.

6.5 Per-Category Error Breakdown

To understand where each system succeeds or fails, we compute CER separately for each normalization category (Table 11). Categories are not mutually exclusive, as examples may require multiple transformation types.

Category	N	Rule	G4	FT	PT+FT
Punctuation	849	0.199	0.162	**0.078**	0.081
Capitalization	62	0.167	0.140	**0.084**	0.085
All-Caps	39	0.852	0.168	0.084	**0.083**
Digit–Word	41	0.218	0.199	0.076	**0.067**

Table 11: Per-category CER. Rule=Rule-Based, G4=Gemma 4 Zero-Shot, FT=mT5 Fine-Tuned, PT+FT=mT5 Pre-Train+FT. Zero-Shot mT5 omitted. Best per row in bold.

The breakdown reveals two notable patterns. First, the rule-based system collapses on All-Caps text (CER 0.852): while it capitalizes the first character of a sentence, it has no mechanism to convert fully capitalized input (e.g., ШАИР БИР ЕРКЕКЧИЛИК) to standard mixed case, leaving the remaining characters unchanged. In contrast, both fine-tuned models and Gemma 4 handle this category well (CER ≈0.08–0.17), as they have learned case normalization from data. Second, Pre-Train+FT is numerically slightly better than Fine-Tuned mT5 on Digit–Word compounds (0.067 vs. 0.076). With only $N = 41$ examples in this category and no paired significance test at this scale, we do not treat this difference as a robust advantage of continual pre-training; it is suggestive but requires a larger, category-targeted evaluation to confirm.

7 Conclusion

We presented the first systematic study of Kyrgyz text normalization, introducing a large-scale dataset of 1.67M noisy–clean text pairs collected from YouTube, Instagram, and Telegram. Our automatic and human evaluations show that fine-tuned mT5-small substantially outperforms all baselines—including zero-shot Gemma 4—achieving a CER of 0.0796; both fine-tuned variants achieved a human accuracy of 99.8%. We showed that continual pre-training with span corruption does not improve over direct fine-tuning, primarily due to hallucination, and provided detailed error analysis of this failure mode.

We hope this work serves as a foundation for future Kyrgyz NLP research. Future directions include: (1) evaluating on out-of-domain text such as news or spoken language transcripts; (2) scaling to larger mT5 variants as hardware permits; (3) incorporating Kyrgyz-specific morphological knowledge into the rule-based baseline; and (4) applying the normalization pipeline as a preprocessing step for downstream tasks such as machine translation and speech synthesis.

Limitations

Domain of the test set. Our test set is drawn from the same sources as the training data (YouTube, Instagram, Telegram). This ensures a representative evaluation of in-domain social-media normalization, but it also means we do not measure out-of-domain generalization. Kyrgyz news articles, speech transcripts, and formal government text have different noise profiles (e.g., ASR-style errors, stylistic punctuation conventions), and performance there is not guaranteed by our numbers. Out-of-domain evaluation is an important avenue for future work.

Reference bias. Because the training references are produced by Gemini 3 Pro, the models we fine-tune are optimized toward a single LLM's idiolect of "normalized Kyrgyz". See §5.4 for a quantitative probe on 50 examples showing that reference bias does not alter system ranking and changes fine-tuned mT5's CER by only 0.012; the probe is limited to a single independent annotator and a fuller study remains future work.

Annotation verification at scale. Only the 1,000-example test set was fully reviewed by two native Kyrgyz speakers with adjudication; the 1.67M training pairs were produced by Gemini 3 Pro and only spot-checked. We therefore cannot rule out that a non-trivial fraction of the training set contains annotation errors that our models may have learned. The spot-check provides an *estimate* of per-example quality, not a guarantee of it.

Data release. For the reasons given in §3.3 (platform redistribution restrictions, re-identification risk, incomplete verification), we release only a 20,000-pair subset of the training data rather than the full 1.67M. This is a deliberate trade-off in favor of user privacy and verification quality at the cost of full reproducibility of the dataset scale. We release the full human-verified test set, all training code, and fine-tuned model checkpoints so that reported results can be reproduced end-to-end modulo the training data subset.

Model scale. We evaluate only mT5-small, as larger mT5 variants and byte-level alternatives such as ByT5 (Xue et al., 2022) require sub-

stantially more computational resources than were available to us. Our results do not speak to how much further headroom exists with mT5-base, mT5-large, or a ByT5 of comparable parameter count. A systematic scan across model families and sizes—including fine-tuned open LLMs such as Gemma 4—is deferred to future work.

Rule-based baseline. Our rule-based baseline is intentionally minimal and does not incorporate Kyrgyz-specific morphological knowledge; it serves as a lower bound rather than a competitive system. A stronger rule-based pipeline built around the finite-state transducer of Washington et al. (2012) and hand-curated dialect/digit dictionaries would likely be meaningfully more competitive, particularly on the 84.9% of the test set where the dominant transformation is punctuation restoration.

Statistical power. Our test set contains $n = 1{,}000$ examples and our pre-training-failure analysis inspects 40 examples. At these sample sizes, the gap between Fine-Tuned and Pre-Train+FT (CER 0.0796 vs. 0.0825, $p = 0.06$) is underpowered: we cannot confidently declare either a true difference or equivalence. The 87.5% hallucination rate from the 40-example inspection comes with a wide 95% Wilson CI of $[74\%, 95\%]$, and we do not claim that this rate generalizes to the full population of failures. Larger test sets and larger targeted failure samples would allow stronger conclusions.

Human evaluation. Human evaluation involved two native Kyrgyz-speaking annotators rating 200 examples per system with no third-party arbitration step. The sample size is consistent with prior work in low-resource NLP human evaluation (Lutgen et al., 2025), but it is small in absolute terms. We report Wilson 95% confidence intervals on per-system scores (Table 4) and three complementary inter-annotator agreement coefficients per system (Cohen's κ, PABAK, Gwet's AC1; Table 6) so that reliability can be assessed under the prevalence skew of the fine-tuned systems. A larger annotation study with explicit arbitration would still be needed to give tight confidence intervals on the Rule-Based score (where annotators substantively disagree) and to disentangle the two fine-tuned variants from each other at the 99.8% ceiling.

Ethics Statement

The dataset consists of publicly accessible text from YouTube, Instagram, and Telegram, collected for the purpose of non-commercial research on a severely under-resourced language. We acknowledge that redistribution of user-generated content scraped from social-media platforms sits in a legal and ethical gray area, even when the original content is publicly viewable; individual users did not consent to inclusion in an NLP dataset, and platform terms of service vary in how they treat research use. To reduce the impact of these concerns we took three concrete measures: (i) we release only a 20,000-pair subset of the training data rather than the full 1.67M, (ii) we strip user handles, @-mentions, and URLs from all released examples, and (iii) we do not release any associated media, profile information, or platform metadata. The normalization task itself does not involve sensitive content generation. Individuals who believe their content has been included and who wish for it to be removed may request removal through the dataset contact address; we will honor such requests in future releases. Our models are intended for Kyrgyz NLP preprocessing and do not pose foreseeable ethical risks beyond those inherent in general language modeling.

LLMs Usage Statement

During the preparation of this manuscript, we made limited use of large language models (LLMs), specifically Anthropic's Claude, to assist with language refinement and organization of some sections. All technical content, equations, derivations, and experimental design were developed entirely by the authors. The LLM was not used for ideation of methods, data analysis, or generation of results.

Acknowledgments

We thank the annotators who participated in the human evaluation study.

References

Anton Alekseev, Sergey I. Nikolenko, and Gulnara Kabaeva. 2023. Benchmarking multilabel topic classification in the Kyrgyz language. *arXiv preprint arXiv:2308.15952*.

Alvan R Feinstein and Domenic V Cicchetti. 1990. High agreement but low kappa: I. The problems of

two paradoxes. *Journal of Clinical Epidemiology*, 43(6):543–549.

Asiye Tuba Koksal, Ozge Bozal, Emre Yürekli, and Gizem Gezici. 2020. #TurkishTweets: A benchmark dataset for Turkish text correction. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4303–4315. Association for Computational Linguistics.

Anne-Marie Lutgen, Alistair Plum, Christoph Purschke, and Barbara Plank. 2025. Neural text normalization for Luxembourgish using real-life variation data. In *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2025)*, pages 94–106. Association for Computational Linguistics.

Dilara Torunoğlu and Gülşen Eryiğit. 2014. A cascaded approach for social media text normalization of Turkish. In *Proceedings of the Workshop on Language Analysis for Social Media*, pages 62–70. Association for Computational Linguistics.

Jonathan N. Washington, Mirlan Ipasov, and Francis M. Tyers. 2012. A finite-state morphological transducer for Kyrgyz. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 934–940.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

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. Association for Computational Linguistics.

Andrew Zupon, Evan Crew, and Sandy Ritchie. 2021. Text normalization for low-resource languages of Africa. *arXiv preprint arXiv:2103.15845*.

Talha Çolakoğlu, Umut Sulubacak, and Ahmet Cüneyd Tantuğ. 2019. Normalizing non-canonical Turkish texts using machine translation approaches. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 267–272. Association for Computational Linguistics.