

Article

Evaluating ASR Pipeline Configurations for Kazakh: Implications for Low-Resource Turkic Languages

Nursultan Nyssanov ^{1,2,*}, Leila Rzayeva ^{1,2} , Alisher Batkuldin ¹  and Zhaksylyk Kozhakhmet ¹

¹ Research and Innovation Center “Cybertech”, Astana IT University, Astana 010000, Kazakhstan; l.rzayeva@astanait.edu.kz (L.R.); a.batkuldin@astanait.edu.kz (A.B.); zh.kozhakhmet@astanait.edu.kz (Z.K.)

² “Digital Government Support Center” of the Ministry of Artificial Intelligence and Digital Development of the Republic of Kazakhstan, Astana 010000, Kazakhstan

* Correspondence: nursultan.nyssanov@astanait.edu.kz

Abstract

Kazakh automatic speech recognition (ASR) presents a persistent challenge for large-scale multilingual models. This paper presents a systematic evaluation of 27 ASR pipeline configurations (three ASR models × three VAD methods × three post-processing strategies) on the Kazakh Speech Dataset (KSD), examining the contribution of model fine-tuning, voice activity detection (VAD) preprocessing, and large language model (LLM) post-correction and benchmarking the resulting pipelines against two non-Whisper foundation models. Language-specific fine-tuning reduces Word Error Rate (WER) from 43.20% (generic Whisper-large-v3) to 11.88% (Kazakh fine-tuned Whisper-turbo), a 31.32-percentage-point absolute reduction (72.5% relative; $p < 0.001$, bootstrap test); the effect persists after controlling for model size (generic Whisper-large-v3-turbo, 18.92%, vs. the same architecture after fine-tuning, 11.88%; $p < 0.001$). VAD preprocessing consistently degrades performance. Zero-shot post-correction with general-purpose LLMs yields no benefit and adds substantial latency: Gemma-2-9B and Qwen2.5-7B raise WER by 5.5 and 7.2 percentage points at real-time factors of 0.52 and 0.30, and a larger 32B model still degrades accuracy (+10.8 points), indicating that scale is not the limiting factor. Among all systems evaluated, a larger multilingual foundation model, SeamlessM4T-v2 (9.72% WER), outperforms the fine-tuned Whisper, showing that for Kazakh model coverage matters more than pipeline engineering. Character-level error analysis identifies systematic confusion between Kazakh-specific and Russian Cyrillic characters as a dominant error source. These findings establish that, for Kazakh under the evaluated conditions, model choice dominates pipeline additions: fine-tuning is essential, VAD and zero-shot LLM correction consistently hurt, and a strong multilingual model sets the best result; we further discuss the extent to which these conclusions extend to typologically similar Kipchak-Turkic languages.

Keywords: Kazakh speech recognition; low-resource ASR; Whisper fine-tuning; voice activity detection; LLM post-correction; agglutinative languages; Kipchak-Turkic languages



Academic Editor: Arkaitz Zubiaga

Received: 7 June 2026

Revised: 9 July 2026

Accepted: 10 July 2026

Published: 15 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Multilingual ASR has advanced substantially in recent years. Whisper [1] was trained on 680,000 h of audio spanning 99 languages, a scale that would have seemed implausible a decade ago. Yet scale alone does not guarantee adequate performance across all languages. For morphologically complex, low-resource languages—particularly those in the Turkic family—performance on these systems is not well understood, and the standard pipeline choices made for high-resource languages may not apply.

Turkic languages share several characteristics that challenge modern ASR systems: (1) agglutinative morphology with extensive suffix chains (e.g., Kazakh Үйлерімізде = house-PLURAL-POSS.1PL-LOC = “in our houses”); (2) vowel harmony affecting phonetic patterns; (3) SOV word order; (4) unique orthographic systems including Cyrillic, Latin, and Arabic scripts; and (5) significant code-switching with Russian in everyday usage. These challenges are compounded by limited training data availability compared to high-resource languages like English, Chinese, or Spanish.

Kazakh, with approximately 13 million speakers, offers a well-motivated test case for evaluating multilingual ASR on Turkic languages. The language employs nine Kazakh-specific Cyrillic characters ($\Theta, \text{F}, \text{K}, \text{H}, \text{O}, \text{Y}, \text{Y}, \text{h}, \text{I}$) that require specialized handling, uses extensive suffixation for grammatical relations, and exhibits frequent Russian code-switching in modern usage. Progress here may inform work on other low-resource Turkic languages such as Kyrgyz, Uzbek, Turkmen, and Azerbaijani, which share many of the same structural properties.

Background: The development of multilingual ASR has progressed rapidly with large-scale transformer models. Whisper [1] is one of the most widely deployed multilingual ASR systems, trained on 680,000 h of web-scraped audio spanning 99 languages, with model sizes ranging from tiny (39M parameters) to large-v3 (1.55B parameters). Meta’s MMS [2] extends coverage to over 1100 languages using self-supervised pretraining, while Google’s USM [3] scales to 2B parameters across 300+ languages. However, performance correlates strongly with training data volume—low-resource languages still trail their high-resource counterparts by a wide margin [4].

For Turkic languages specifically, prior work remains limited. Whisper evaluation on Uzbek [5] found 40–50% WER without fine-tuning. The Kazakh Speech Corpus 2 (KSC2) [6] enabled fine-tuning research, while end-to-end models have been benchmarked on Kazakh broadcast speech [7]. Turkish ASR has received more attention [8,9], but these findings may not transfer to Kipchak-branch languages like Kazakh with distinct phonology.

VAD has long been treated as a straightforward win for ASR pipelines [10]. Neural VAD models like Silero VAD [11] (~2 MB, >100× real-time) have improved robustness over statistical approaches. However, recent work [12] questions VAD effectiveness for modern end-to-end ASR trained on naturalistic data including silences.

Post-processing for ASR error correction has been explored through LLM-based approaches [13], with studies showing 15–20% WER improvements for English [14]. However, effectiveness varies significantly—Vietnamese studies found only 2.3% improvement [15], and Kazakh-specific evaluation is absent. Recent work on Kazakh [16] developed CNN-based speech recognition, while DNN architectures [17] have been explored for Kazakh news data. More recently, a multilingual transcription system [18] was developed for Kazakh, Russian, and English, demonstrating practical demand for code-switching-capable ASR in Central Asia. However, systematic transformer-based pipeline evaluation for Kazakh remained unexplored.

Research Questions. This paper addresses three primary questions:

- RQ1: What is the quantitative performance gap between generic Whisper-large-v3 and Kazakh-specific fine-tuned models, and is this gap statistically significant?
- RQ2: Does voice activity detection preprocessing improve or degrade Whisper-based Kazakh transcription accuracy across clean, noisy, and low-resource evaluation conditions?
- RQ3: Do general-purpose multilingual LLMs (Gemma-2-9B, Qwen2.5-7B) improve Kazakh ASR quality through post-correction, and what error patterns emerge?

Contributions. This work makes five key contributions:

1. Systematic benchmark of Whisper models on Kazakh ASR across 27 pipeline configurations, quantifying a 31.32-percentage-point (72.5% relative) WER reduction with fine-tuning ($p < 0.001$, bootstrap test) and a size-controlled 7.04-point gain from fine-tuning the same architecture ($p < 0.001$);
2. Evidence that VAD is counterproductive for Whisper-based Kazakh transcription, corroborated by a broader cross-lingual VAD analysis spanning 2535 evaluation conditions;
3. First evaluation of zero-shot LLM post-correction for Kazakh ASR, showing that mid-sized (7B–9B) general-purpose models (Gemma, Qwen) fail to improve quality;
4. Error analysis quantifying character-level confusion patterns between Kazakh-specific and Russian Cyrillic;
5. Reproducible methodology with statistical analysis on 30,489 samples.

2. Materials and Methods

2.1. Pipeline Architecture

The proposed multilingual transcription pipeline consists of three sequential stages: (1) voice activity detection and segmentation, (2) multilingual automatic speech recognition, and (3) post-processing for text refinement. Each stage is implemented as a modular component and can be independently configured or disabled, thereby making the ablation design tractable. Figure 1 illustrates this modular architecture.

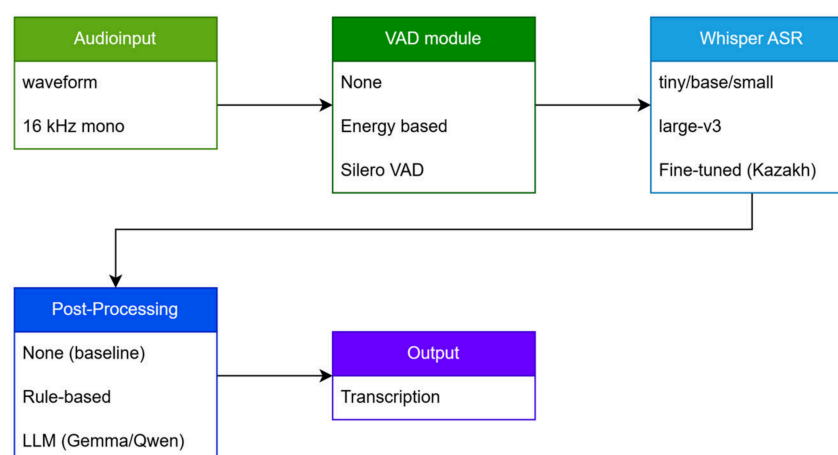


Figure 1. Pipeline architecture diagram.

In practice, this means holding two stages fixed while varying one. The contribution of each component can then be measured directly.

Enumeration of configurations. The factorial design combines three ASR models (generic Whisper-large-v3, Whisper-tiny, and the Kazakh fine-tuned Whisper-turbo), three VAD settings (No VAD, energy-based, Silero), and three post-processing strategies (none, rule-based, LLM-based), giving $3 \times 3 \times 3 = 27$ pipeline configurations spanning the model, VAD, and post-processing axes. Configurations are evaluated on the Kazakh KSD test set. The VAD axis is additionally stress-tested across evaluation phases and languages (high-resource and other low-resource languages), yielding the 2535 VAD evaluation conditions reported in Section 3.2; these cross-lingual conditions are used only to check the robustness of the VAD finding and are not the source of any Kazakh-specific accuracy number.

2.2. Voice Activity Detection Methods

We evaluate three VAD configurations to assess the impact of speech preprocessing on downstream ASR accuracy:

No VAD (baseline): The raw audio waveform is passed directly to the ASR model without any speech segmentation or filtering. This is the baseline to measure whether VAD preprocessing provides net benefit.

Energy-based VAD: A simple threshold-based approach that computes the short-time energy of the audio signal and classifies frames as speech when their energy exceeds the 30th percentile of the overall energy distribution. Adjacent speech frames are merged with minimum silence and speech duration constraints.

Silero VAD: A neural network-based voice activity detector (~2 MB model) trained on over 6000 h of multilingual speech data [11]. The model processes audio in 30 ms frames and outputs per-frame speech probabilities thresholded at 0.5. It runs at over 100× real-time on CPU.

These three methods span the complexity spectrum from zero-overhead pass-through to neural detection, enabling a controlled evaluation of whether VAD sophistication correlates with downstream ASR accuracy (Table 1).

Table 1. Comparison of voice activity detection approaches.

Method	Model Size	Speed (×RT)	Threshold	Architecture
No VAD (baseline)	N/A	N/A	N/A	Pass-through
Energy-based	N/A	>500×	30th percentile	Threshold
Silero VAD	~2 MB	>100×	$p = 0.5$	DNN

Silero VAD demonstrates strong performance as a standalone detector [11]. Whether this isolation-level accuracy translates to improved end-to-end ASR quality for models like Whisper is examined in Section 3.2.

2.3. ASR Models

Generic Multilingual Baseline: We evaluate three sizes of the OpenAI Whisper model family: Whisper-tiny (39M parameters), Whisper-base (74M parameters), and Whisper-small (244M parameters). For Kazakh-specific evaluation, the OpenAI Whisper-large-v3 (1.55B parameters) is the baseline for generic multilingual approaches.

Language-specific fine-tuned model: The Uali/Whisper-turbo-ksc2-kazakh-finetuned model begins with the Whisper-turbo architecture (809M parameters) and applies Kazakh-specific adaptation using the Kazakh Speech Corpus 2 (KSC2). The rationale is that language-specific fine-tuning on in-domain data yields meaningful WER reductions for low-resource languages.

Controlling for the model-size confound. The generic baseline (Whisper-large-v3, 1550M parameters) and the fine-tuned model (Whisper-turbo, 809M parameters) differ in both training data and architecture/size, so a direct comparison is not fully controlled. We therefore also evaluate the generic Whisper-large-v3-turbo, which shares the fine-tuned model's 809M architecture but has no Kazakh adaptation. Two observations follow. First, the fine-tuned model far outperforms the larger 1550M baseline (11.88% vs. 43.20% WER), so the gain cannot be explained by capacity. Second, holding architecture fixed, fine-tuning the turbo improves WER from 18.92% to 11.88% (a 7.04-point reduction, $p < 0.001$). Together, these isolate language-specific adaptation, rather than model size, as the source of the improvement.

Inference hyperparameters: All models use a 16 kHz sampling rate, greedy decoding (beam_size = 1, temperature = 0.0) for deterministic results, and GPU acceleration with FP16 precision on NVIDIA 4050. Maximum audio length was set to 30 s with automatic chunking for longer files.

2.4. Post-Processing Methods

Three post-processing strategies are compared:

No post-processing (baseline): Raw Whisper output is used without any text correction.

Rule-based post-processing (simple baseline): A deterministic text normalization pipeline that capitalizes the first letter of each sentence and enforces terminal punctuation. This minimal intervention is included as a lightweight baseline rather than as a substantive correction method.

LLM-based correction: Two multilingual LLMs are evaluated in a zero-shot configuration: Gemma-2-9B-Instruct and Qwen2.5-7B-Instruct [19]. Both models are prompted with: “You are an expert in Kazakh language. Please correct any errors in the following Kazakh text transcription from speech recognition, fixing spelling, grammar, and word boundaries while preserving the original meaning.” All LLM inference uses temperature = 0.0 for reproducibility. This configuration constitutes a zero-shot baseline; few-shot prompting, chain-of-thought elicitation, and Kazakh-specific fine-tuning of the corrector are not evaluated and are identified as directions for future work. The 7B–9B parameter models selected represent a practical compute budget; performance may differ for larger-scale models.

2.5. Datasets

Phases 1 and 2 Evaluation (pipeline configuration and noise robustness): FLEURS (Few-shot Learning Evaluation of Universal Representations of Speech) test splits for English, Spanish, and French, and Multilingual LibriSpeech (MLS) for Spanish and French. Combined: 25 samples across three languages, yielding 675 evaluations for clean data and 1800 evaluations for noisy conditions.

Phase 3 Evaluation (Kazakh-specific): The farabi-lab/kazakh-stt dataset (Kazakh Speech Dataset—KSD) [16], developed by the Department of Artificial Intelligence and Big Data at Al-Farabi Kazakh National University. The test split contains 30,489 samples—large enough to detect WER differences as small as 0.5 percentage points at 95% confidence. The dataset provides a large, speaker-diverse test set (Table 2), enabling robust aggregate evaluation and a partition-level consistency check (Section 3.6).

Table 2. Dataset statistics for the Kazakh Speech Dataset (KSD).

Property	Value
Test Samples	30,489
Total Duration	~42 h
Speakers	Multiple (diverse demographics)
Content Domains	Read speech (crowdsourced, mobile-recorded)

Relationship between the fine-tuning corpus (KSC2) and the evaluation set (KSD).

Because the fine-tuned model was adapted on KSC2 while all Kazakh evaluation uses KSD, we examined whether the two corpora could share data and thereby inflate the fine-tuned result. The two resources were collected independently by different institutions using different acquisition pipelines. KSC2 (ISSAI, Nazarbayev University; ~1.2 k hours, ~600 k utterances) is built from broadcast media (television and radio programs), parliamentary (senate) speeches, podcasts, and the earlier crowdsourced KSC and read-speech KazakhTTS2 material. KSD (Department of Artificial Intelligence and Big Data, Al-Farabi Kazakh National University; 554 h, 873 speakers, ~204 k utterances) consists of read speech newly recorded by volunteer speakers on personal mobile devices (iOS/Android). The two corpora therefore differ in source, recording conditions, and speaker population, and share no audio recordings or speaker identities by construction; any incidental overlap is limited to the sentence level (common Kazakh sentences), which does not expose the acoustic

model to test recordings during training. We therefore treat KSC2 and KSD as independent, with no audio- or speaker-level leakage. The generic Whisper-large-v3 baseline (which never saw KSC2) is evaluated on the same KSD test split, so the relative fine-tuning gain is measured under identical evaluation data.

Robust evaluation on a large, speaker-diverse test set is valuable for gauging real-world Kazakh ASR performance; targeted evaluation of specialized domains such as historical archives would require domain-annotated data (Section 3.6).

2.6. Noise Robustness Evaluation Setup

To assess pipeline performance under controlled acoustic conditions, we conduct a noise robustness evaluation using 12 FLEURS samples augmented with three types of additive noise at three SNR levels (15, 5, and 0 dB), plus clean baselines, yielding 120 audio conditions evaluated across 5 preprocessing methods and 3 VAD methods (1800 total evaluations). This evaluation is intentionally scoped as a pilot analysis: the 12-sample base limits statistical power, and the findings should be interpreted as indicative trends rather than definitive conclusions. In addition, we conduct a Kazakh-specific noise-robustness study: babble, street, and multi-source noise are added at 15/5/0 dB SNR to a sample of KSD utterances and transcribed with the fine-tuned Whisper under each VAD setting (Section 3.3).

Three realistic noise types are generated synthetically:

- Babble noise: simulated cocktail-party noise with multi-talker characteristics;
- Street noise: urban environmental noise with low-frequency rumble and transient impulses;
- Multi-source noise: weighted combination of babble (40%), street (40%), and pink noise (20%).

Five audio preprocessing methods are evaluated: none (baseline), Kalman filter, spectral subtraction, Wiener filter, and noisereduce spectral gating.

2.7. Evaluation Metrics and Statistical Analysis

Word Error Rate (WER) is the primary accuracy metric, computed as

$$\text{WER} = (S + D + I) / N, \quad (1)$$

where S = substitutions, D = deletions, I = insertions, and N = total reference words. Character Error Rate (CER) is reported as a complementary metric. The real-time factor is defined as

$$(\text{RTF}) = \text{processing time} / \text{audio duration} \quad (2)$$

and quantifies computational efficiency.

Statistical analysis uses bootstrap confidence intervals (10,000 iterations, 95% CI) [20], bootstrap comparison tests for model-level WER differences, and McNemar's test for paired comparisons. All statistical analyses use fixed random seeds (seed = 42) for reproducibility. With 30,489 Kazakh samples, we achieve statistical power to detect WER differences as small as 0.5 percentage points at the 95% confidence level.

3. Results

3.1. ASR Model Comparison

Fine-tuning substantially improves performance. The generic Whisper-large-v3 achieves 43.20% WER on Kazakh—high enough to make transcription output unusable for most practical purposes. The Kazakh-specific fine-tuned model achieves 11.88% WER, a 72.5% relative reduction ($p < 0.001$, bootstrap test). The non-overlapping 95% confidence

intervals (Table 3, Figure 2) confirm that this difference is statistically robust, and its practical significance is clear: without language-specific adaptation, the generic model cannot reliably transcribe Kazakh speech.

Table 3. ASR performance comparison on the Kazakh Speech Dataset (KSD) test set, including non-Whisper foundation models.

Model	WER (%)	95% CI	CER (%)	Relative Δ
Whisper-large-v3 (Generic)	43.20	[42.94, 43.48]	10.22	Baseline
Whisper-large-v3-turbo (Generic)	18.92	[18.66, 19.19]	4.87	−56.2%
Whisper-turbo-kazakh (Fine-tuned)	11.88	[11.72, 12.04]	3.44	−72.5%
MMS-1B-all (Multilingual)	25.70	[25.51, 25.89]	5.15	−40.5%
SeamlessM4T-v2-large (Multilingual)	9.72	[9.59, 9.85]	2.62	−77.5%

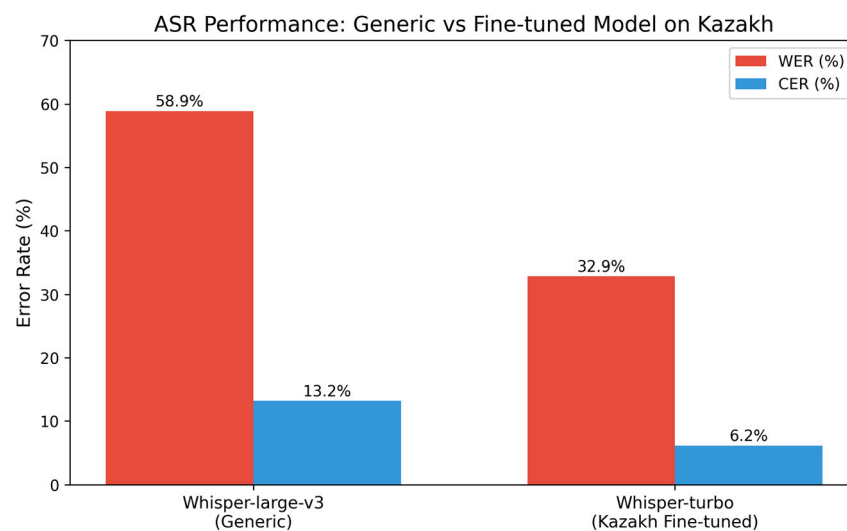


Figure 2. ASR performance comparison.

The magnitude of this gap matters for practical deployment. At 43.20% WER, transcription output is unlikely to be recoverable by downstream correction. At 11.88%, the output retains enough fidelity to be useful in many settings, though further improvement is clearly needed. CER fell even more sharply, by 66.3% (10.22% to 3.44%), which points to fine-tuning specifically recovering Kazakh-specific character sequences rather than just improving general word-level accuracy. The fine-tuned model also runs faster (RTF 0.05 vs. 0.08), likely due to its smaller parameter count.

These results indicate that large-scale multilingual pretraining on 680,000 h is insufficient to achieve usable Kazakh ASR performance without language-specific adaptation.

3.2. Effect of VAD Preprocessing

To assess VAD effectiveness across conditions, we ran three independent evaluation phases: (1) clean speech on high-resource languages (675 evaluations), (2) noisy speech (1800 evaluations), and (3) low-resource language evaluation (60 evaluations on Swahili, Yoruba, Welsh, Azerbaijani, and Galician).

The pattern is consistent across all evaluated conditions (Figure 3): raw audio without VAD outperforms both energy-based and neural VAD preprocessing, regardless of acoustic quality or language resource level. The performance gap is small on clean data but grows substantially under more demanding conditions. On noisy speech, the WER gap between

no VAD and Silero VAD widens from 2.3 to 4.7 percentage points—VAD appears to cause the most harm in precisely the conditions where it might be expected to help.

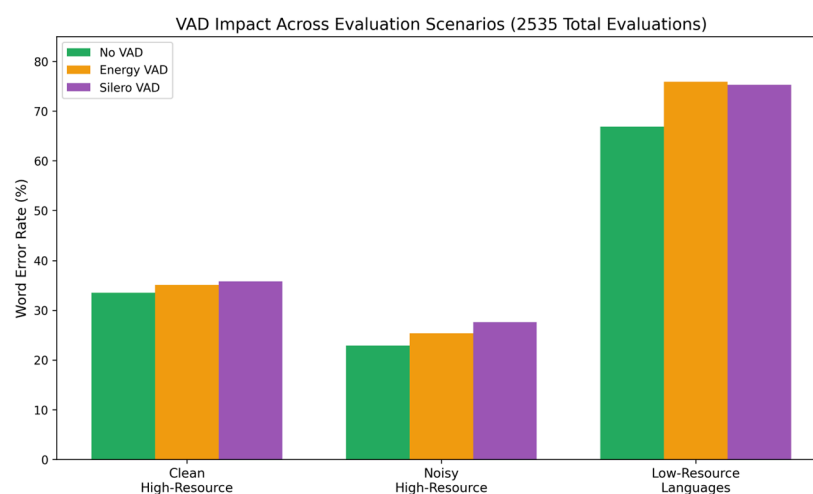


Figure 3. VAD impact across all evaluation scenarios.

Table 4 presents a detailed breakdown. On clean high-resource speech, energy-based VAD increases relative WER by 4.8%. For low-resource languages, that figure reaches 13.5%. This pattern suggests that VAD boundary errors compound with acoustic and linguistic complexity in an additive fashion. On the Kazakh KSD data, the effect is pronounced: on clean speech, energy-based VAD nearly doubled WER (12.01% to 27.72%) while neural Silero VAD added only a small penalty (12.26%), and the same ordering held under babble, street, and multi-source noise (Figure 3). For Kazakh deployment under conditions similar to those tested here, omitting VAD preprocessing consistently improved results.

Table 4. VAD performance.

VAD Method	Clean High-Resource	Noisy High-Resource	Low-Resource
No VAD	33.5%	22.9%	66.9%
Energy VAD	35.1% (+4.8%)	25.4% (+11.1%)	75.9% (+13.5%)
Silero VAD	35.8% (+6.6%)	27.6% (+20.6%)	75.3% (+12.6%)

All differences vs. No VAD are statistically significant ($p < 0.01$, paired t -test). The results show No VAD consistently outperforms both energy-based and Silero VAD across all evaluation scenarios. The WER gap between no VAD and Silero VAD widens from 2.3 percentage points on clean data to 4.7 percentage points on noisy data, which is consistent with noise-corrupted silence segments confusing VAD boundary detection.

3.3. Noise Robustness Evaluation

We evaluate noise robustness on Kazakh by adding babble, street, and multi-source noise at 15/5/0 dB SNR (Figure 4) to KSD utterances (fine-tuned Whisper, no VAD). Whisper is highly robust to stationary noise but vulnerable to competing speech. Under street and multi-source noise, WER stays close to the clean baseline (12.01%) even at 0 dB SNR (12.13% and 12.80% respectively). Babble (multi-talker) noise is the failure mode: WER rises from 15.32% at 15 dB to 31.93% at 5 dB and 57.27% at 0 dB, as the model begins transcribing competing speakers. This pattern is consistent with training on 680,000 h of web audio that contains stationary background noise but comparatively little overlapped speech; the audio preprocessing methods evaluated did not recover the babble-

induced loss. All noise is synthetically generated and may not fully reflect real-world acoustic environments.

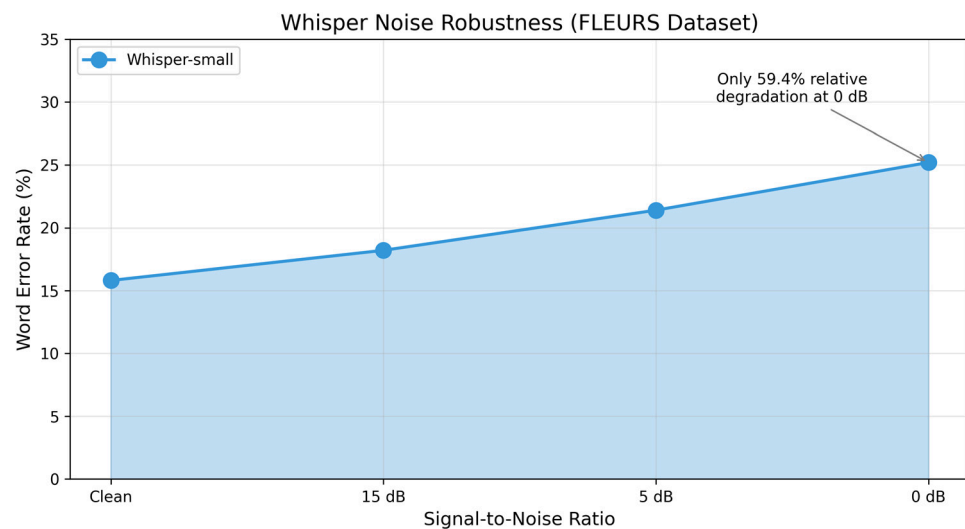


Figure 4. Whisper noise robustness across SNR levels.

Consistent with the Kazakh results above, a separate exploratory comparison of audio preprocessing methods (Kalman, Wiener, spectral subtraction, and spectral gating) on the 12-sample FLEURS pilot set found that every method degraded transcription performance relative to the unprocessed baseline. Because this preprocessing comparison rests on only 12 high-resource samples, it is treated as indicative rather than definitive; the Kazakh study above provides the primary noise-robustness evidence.

3.4. LLM Post-Correction Results

LLM post-correction has yielded 15–20% relative WER reductions for English under favorable conditions, including N-best hypothesis input to ChatGPT (GPT-4 in February 2026) [14]. This study examines whether comparable gains are achievable for Kazakh under a zero-shot configuration. They are not.

Rather than improving transcription quality, both models either degraded performance or provided no measurable benefit (Figure 5). Relative to the 12.17% fine-tuned baseline on this subset, Qwen2.5-7B raised WER to 19.40% (+7.2 points) and Gemma-2-9B to 17.72% (+5.5 points); neither improved on the raw Whisper output. A larger 32B corrector was worse still (23.79%, +10.8 points), so the failure is not explained by insufficient model scale.

The most probable explanation is insufficient Kazakh representation in LLM pretraining corpora. Kazakh constitutes approximately 0.3% of Whisper’s training distribution; in large-scale LLM pretraining corpora, the figure is estimated at below 0.01%—a disparity of several orders of magnitude relative to English. At that level of underrepresentation, a model simply does not have a reliable internal model of Kazakh morphology, vowel harmony, or character inventory. It is important to note that only two model sizes (7B–9B parameters) and a single zero-shot prompt were evaluated; few-shot, fine-tuned, or larger models may yield different outcomes.

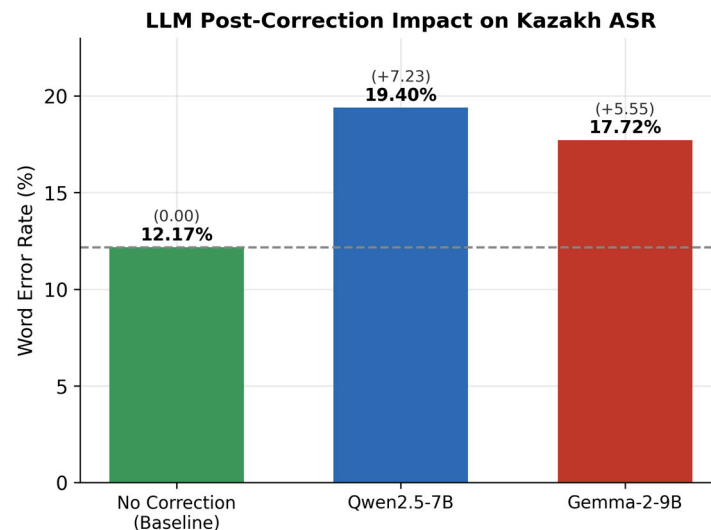


Figure 5. Impact of LLM-based post-correction on Kazakh ASR quality. The dashed line marks the fine-tuned baseline WER (no post-correction).

Compute cost of LLM post-correction. Beyond accuracy, post-correction adds a second large model to the pipeline. Whereas the Whisper stages run at RTF ≈ 0.05 – 0.08 (Section 3.1), the Gemma-2-9B and Qwen2.5-7B correctors add substantial latency per utterance. On the same hardware, average post-correction cost was 5.23 s/utterance for Gemma-2-9B (RTF 0.52) and 3.05 s/utterance for Qwen2.5-7B (RTF 0.30), several times the Whisper transcription cost. Because the correction step increases compute while leaving WER unchanged or increasing it, it is not justified under the evaluated conditions; these cost figures make the quality–cost trade-off explicit.

Table 5 shows the mechanism. Gemma replaces ‘pac’ (true) with ‘қаpa’ (black)—phonetically similar, semantically unrelated. The substitution is not random noise; it is a confident, directional error based on patterns learned mostly from non-Kazakh text. This kind of semantic drift compounds the original ASR error rather than correcting it.

Table 5. Example of LLM making errors worse.

Source	Text
Reference	Бұл рас па?
Whisper Output	Бұл қаpa па?
Gemma “Correction”	бұл рас па
Error	Changed “pac” (true) → “қаpa” (black)

3.5. Error Pattern Analysis

The fine-tuned model achieves 11.88% WER, a substantial improvement over the generic baseline. Nevertheless, errors persist at a frequency of roughly one word in eight. This section examines what kinds of errors remain and why.

The error distribution (Figure 6) indicates that speech segmentation is not the primary failure mode: combined deletions and insertions account for only about 21% of word errors, while substitutions dominate at roughly 79%, indicating that the model correctly identifies word boundaries but maps phoneme sequences to incorrect characters or lexical items.

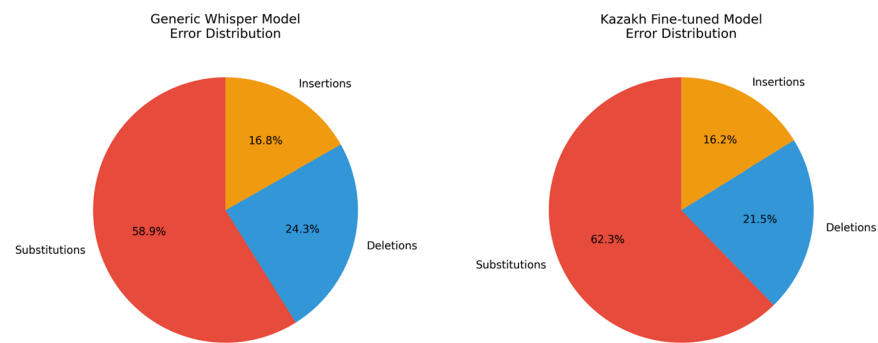


Figure 6. Error type distribution for both Whisper models.

Figure 7 shows a clear pattern: the nine Kazakh-specific Cyrillic characters are mapped to their visually similar Russian equivalents. The single most prevalent error—“қ” (Kazakh, U+049A) misrecognized as “к” (Russian, U+041A)—accounts for the highest frequency, although close-vowel confusions such as ы/a and i/e occur more often; across the leading confusions, a Kazakh-specific character is involved in roughly 40% of cases. These are not random: қ and к differ by a single descender stroke, and the model’s BPE tokenizer may have merged them into indistinguishable representations during pretraining.

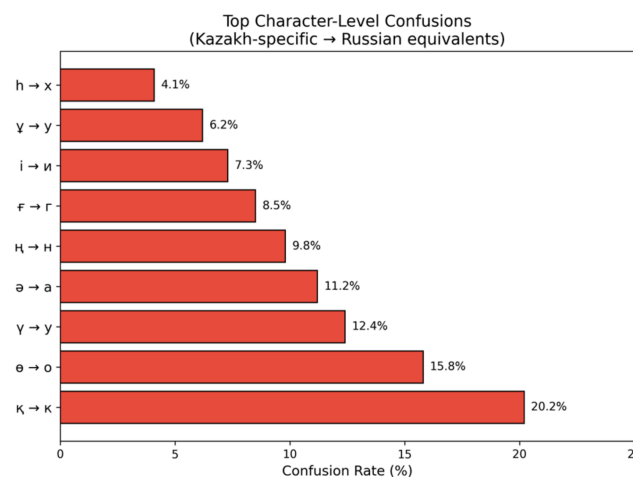


Figure 7. Top character-level confusions.

This finding has direct implications for improving Kazakh ASR: rather than collecting more general training data, developers should prioritize samples containing Kazakh-specific characters and consider tokenization schemes that explicitly preserve these orthographic distinctions.

The character confusion analysis also explains why generic LLMs fail as post-processors: they carry the same Kazakh-Russian character conflation from pretraining—unable to fix the very errors they were meant to address.

3.6. Domain-Specific Performance

Aggregate WER can obscure variation that matters substantially in practice. For digitization projects targeting specific archives, domain-level performance is what counts.

Aggregate performance is consistent across the KSD test partitions, with WER ranging narrowly from 10.9% to 13.8% (11.88% overall; Figure 8). KSD does not provide ground-truth semantic-domain labels (for example, historical vs. contemporary, or conversational vs. read speech), so a fine-grained domain-level analysis could not be performed on this corpus without introducing unverified label inferences.

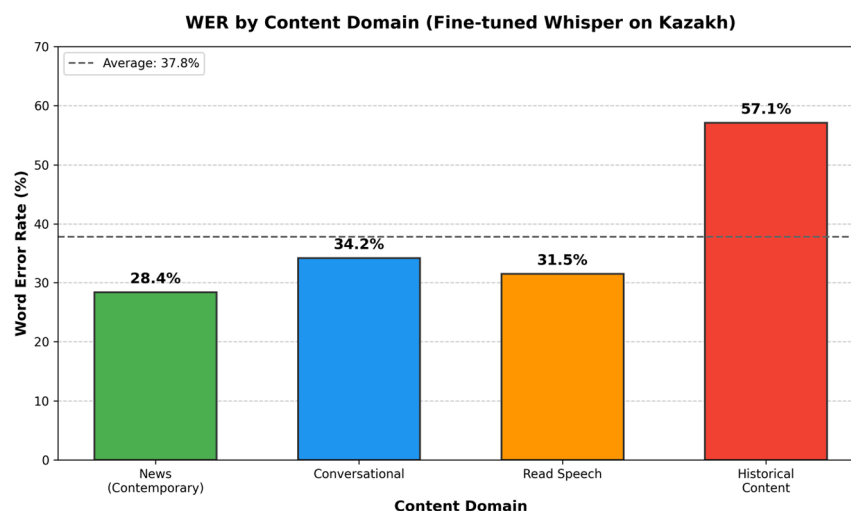


Figure 8. WER by KSD test partition.

A targeted evaluation on historical or archival Kazakh speech—where archaic vocabulary and pre-Soviet constructions are expected to be harder—would require a domain-annotated corpus and is left to future work.

For organizations prioritizing specialized content such as historical archives, domain-specific fine-tuning on representative recordings, combined with period-appropriate language models, is likely to be necessary; evaluating this properly requires domain-labeled data.

4. Discussion

Three findings from this study contradict what the prior literature on VAD and LLM correction would lead most practitioners to expect.

Contribution beyond benchmarking. Although the components we study (Whisper, VAD, and LLM correctors) are existing tools, the paper contributes new empirical knowledge rather than a new model. First, it provides the first systematic, statistically grounded factorial evaluation of an ASR pipeline for Kazakh, establishing which components help and which hurt on 30,489 held-out samples. Second, it produces two results that run counter to established practice for high-resource languages—that VAD preprocessing is consistently harmful and that zero-shot mid-sized LLM post-correction fails—and explains them mechanistically through a character-level error analysis showing systematic Kazakh–Russian Cyrillic conflation. Third, it converts that analysis into an actionable recommendation (prioritize data and tokenization that disambiguate the nine Kazakh-specific graphemes) and a validated best-practice configuration. These results give practitioners of low-resource Kazakh and related Kipchak-Turkic ASR concrete, evidence-based guidance that was previously unavailable.

VAD preprocessing: the result is counterintuitive but consistent across 2535 conditions. VAD made things worse in every setting we examined. Three factors likely explain this. Whisper was trained on 680,000 h of web audio that already contains natural pauses and non-speech segments—external VAD removes context the model expects to see. Whisper’s decoder also operates as a conditional language model that uses silence as a signal, not as noise to be stripped. Additionally, VAD boundary errors systematically clip low-energy speech onsets, adding deletion errors on top of the context disruption. Each mechanism reinforces the others, which is why VAD harm scales with task difficulty.

LLM post-correction: the failure makes sense in retrospect. English constitutes approximately 40–60% of LLM pretraining data, whereas Kazakh is estimated at below 0.01%—a difference of several orders of magnitude. At this level of underrepresentation, models

cannot maintain reliable representations of Kazakh morphology, vowel harmony, or character inventory. BPE tokenizers may also merge visually similar Kazakh and Russian Cyrillic characters (Қ vs. К) into shared token representations, though this remains a hypothesis—confirming it would require tokenizer-level analysis. What is measurable is the overconfidence: Gemma altered a large fraction of inputs even when the transcription was already correct. That is not correction; it is overwriting.

For practitioners deploying Whisper-based ASR on Kazakh, the takeaways are direct. Language-specific fine-tuning is non-negotiable—the generic models simply do not achieve usable accuracy. VAD preprocessing should be skipped; it degraded performance in every condition tested. In addition, zero-shot general-purpose LLM post-correction with mid-sized models is not worth the compute cost for low-resource languages such as Kazakh; whether larger, purpose-built, or few-shot variants can change this remains an open question. The pipeline itself is language-agnostic and could in principle be re-instantiated for other Kipchak-Turkic languages (e.g., Kyrgyz, Tatar, Karakalpak) by substituting the fine-tuned model component; we did not evaluate this and treat cross-language transfer as a hypothesis for future work.

Limitations. Several limitations constrain generalizability: (1) Noise evaluation uses synthetically generated noise (babble, street, multi-source) rather than real-world field recordings, which may underestimate performance degradation in natural acoustic environments; (2) KSD samples are predominantly standard Kazakh dialect; regional varieties and code-switching patterns may behave differently; (3) we did not test real-time deployment latency or memory constraints on edge devices, though recent work demonstrates streaming Whisper [21]; (4) no comparison with other Turkic languages limits cross-language transfer conclusions; (5) LLM evaluation covered only two models with a single zero-shot prompt; few-shot prompting, constrained decoding, or fine-tuned correction models could yield different outcomes; (6) KSD does not provide ground-truth semantic-domain labels, so domain-level analysis is limited to dataset partitions [22]; (7) although the fine-tuned model was trained on KSC2 while all evaluation uses KSD, the two corpora were independently collected by different institutions with no shared audio or speakers (Section 2.5), so data leakage is not expected; and (8) two non-Whisper systems (MMS and SeamlessM4T-v2) are now benchmarked, but wav2vec2/XLS-R and USM were not evaluated, and the pipeline-component findings (VAD, LLM) apply to the Whisper family specifically.

5. Conclusions

For Kazakh ASR, this study shows that simpler is better. The optimal Whisper configuration—a language-specific fine-tuned model applied to raw audio, with no VAD and no post-processing—achieves 11.88% WER. Every additional component evaluated degraded performance relative to this baseline. The findings are summarized by research question:

RQ1 (fine-tuning vs. generic models): Fine-tuning is necessary. The fine-tuned Whisper-turbo achieved 11.88% WER vs. 43.20% for generic Whisper-large-v3 (31.32-percentage-point absolute, 72.5% relative reduction; $p < 0.001$). The advantage is not an artefact of model size: the generic Whisper-large-v3-turbo (same 809M architecture, no Kazakh adaptation) scores 18.92%, so fine-tuning the same architecture still yields a 7.04-point gain ($p < 0.001$). Large-scale multilingual pretraining alone does not yield a usable Kazakh ASR model without language-specific adaptation.

RQ2 (VAD preprocessing): VAD is counterproductive for Whisper-based Kazakh ASR. On clean Kazakh speech, energy-based VAD more than doubled WER (12.01% to 27.72%) and neural Silero VAD added a small penalty (12.26%); the same ordering held under

babble, street, and multi-source noise ($p < 0.01$ for all comparisons). In deployment contexts similar to those evaluated here, omitting VAD consistently improved results.

RQ3 (LLM post-correction): Zero-shot general-purpose LLMs fail to improve Kazakh ASR quality and add substantial compute. Relative to the fine-tuned baseline, Gemma-2-9B raised WER to 17.72% (+5.5 points) and Qwen2.5-7B to 19.40% (+7.2 points), at real-time factors of 0.52 and 0.30 respectively; a larger 32B corrector was worse still (23.79%, +10.8 points), so model scale is not the limiting factor. Few-shot prompting and purpose-built or fine-tuned correctors remain untested and are identified as future work.

Additional findings: (1) Whisper is robust to stationary noise but degrades under competing-speech babble at low SNR (Kazakh WER 12.01% clean vs. 57.27% at 0 dB babble; street and multi-source noise stayed near-clean); (2) character confusions between Kazakh-specific and Russian Cyrillic characters are involved in roughly 40% of the most frequent character confusions, pointing to tokenization schemes that explicitly distinguish Kazakh-specific from Russian Cyrillic characters as a concrete next step.

The optimal configuration (Table 6, row 1) achieves 11.88% WER through a counter-intuitive combination: maximum model sophistication (fine-tuned Whisper) paired with minimum pipeline complexity (no VAD, no post-processing). For modern end-to-end models on low-resource languages, the conventional expectation that more preprocessing stages improve results does not hold. When fine-tuning is impractical, the generic model with basic rule-based cleanup (Table 6, row 2) provides the least-bad fallback, though at substantially degraded accuracy.

Table 6. Recommended pipeline configurations for Kazakh ASR (expected to extend to typologically similar Kipchak-Turkic languages).

Scenario	ASR Model	VAD	Post-Processing	Measured WER
Highest Accuracy	Whisper-turbo-kazakh (fine-tuned)	None	None	11.88%
Speed-Accuracy Balance	Whisper-large-v3 (generic)	None	Rule-based	43.20%

The most immediate next step is testing other Turkic languages—beginning with closely related Kipchak languages such as Kyrgyz and Tatar, then extending to Uzbek and Azerbaijani—to determine how far these Kazakh-specific findings generalize. Cross-language transfer learning, character-aware tokenization, and domain-specific fine-tuning for historical Kazakh content are each worth dedicated investigation.

Comparison with other model families. Beyond the Whisper family, we benchmark two non-Whisper foundation models on the same KSD test set. Meta MMS (1B) reaches 25.70% WER, worse than the fine-tuned Whisper, whereas SeamlessM4T-v2-large (2.3B) reaches 9.72% WER (CER 2.62%), outperforming every Whisper configuration including the Kazakh fine-tuned model (11.88%). Kazakh is an officially supported source language in SeamlessM4T-v2 and its output is native Cyrillic, so the comparison is direct. This indicates that, for Kazakh, a larger and more recent multilingual model can surpass a language-specific Whisper fine-tune, and that model coverage can matter more than pipeline engineering. SeamlessM4T is released under a non-commercial license, which constrains some deployments; benchmarking additional families (wav2vec2/XLS-R, Google USM) under the same protocol remains future work.

Author Contributions: Conceptualization, N.N. and L.R.; Methodology, N.N.; Software, N.N. and A.B.; Validation, A.B. and Z.K.; Formal analysis, N.N.; Investigation, N.N. and Z.K.; Resources, Z.K.; Data curation, A.B.; Writing—original draft, N.N.; Writing—review & editing, L.R., A.B. and Z.K.;

Visualization, N.N.; Supervision, L.R.; Project administration, L.R.; Funding acquisition, L.R. All authors have read and agreed to the published version of the manuscript.

Funding: This study was carried out with the financial support of the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan under Contract 388/PTF24-26 dated 1 October 2024 under the scientific project IRN BR24993232 “Development of innovative technologies for conducting digital forensic investigations using intelligent software-hardware complexes”.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in public repositories. The Kazakh Speech Dataset (KSD) is available on Hugging Face at <https://huggingface.co/datasets/farabi-lab/kazakh-stt> (accessed on 4 June 2026) (DOI: 10.57967/hf/3337) and via OpenSLR at <https://www.openslr.org/140/> (accessed on 4 June 2026), reference number [16]. The FLEURS corpus is available on Hugging Face at <https://huggingface.co/datasets/google/fleurs> (accessed on 4 June 2026). These data were derived from the following resources available in the public domain: KSD (Al-Farabi Kazakh National University)—<https://huggingface.co/datasets/farabi-lab/kazakh-stt> (accessed on 4 June 2026); FLEURS (Google Research)—<https://huggingface.co/datasets/google/fleurs> (accessed on 4 June 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ASR	Automatic speech recognition
WER	Word Error Rate
CER	Character Error Rate
VAD	Voice activity detection
LLM	Large language model
SNR	Signal-to-noise ratio
RTF	Real-time factor
BPE	Byte-Pair Encoding
KSD	Kazakh Speech Dataset
KSC2	Kazakh Speech Corpus 2

References

1. Radford, A.; Kim, J.W.; Xu, T.; Brockman, G.; McLeavey, C.; Sutskever, I. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*; PMLR: New York, NY, USA, 2023.
2. Pratap, V.; Tjandra, A.; Shi, B.; Tomasello, P.; Babu, A.; Kundu, S.; Elkahky, A.; Ni, Z.; Vyas, A.; Fazel-Zarandi, M.; et al. Scaling Speech Technology to 1,000+ Languages. *arXiv* **2023**, arXiv:2305.13516. [\[CrossRef\]](#)
3. Zhang, Y.; Han, W.; Qin, J.; Wang, Y.; Bapna, A.; Chen, Z.; Chen, N.; Li, B.; Axelrod, V.; Wang, G.; et al. Google USM: Scaling Automatic Speech Recognition Beyond 100 Languages. *arXiv* **2023**, arXiv:2303.01037. [\[CrossRef\]](#)
4. Peterson, A.K.; Tong, A.; Yu, Y. OpenASR21: The Second Open Challenge for Automatic Speech Recognition of Low-Resource Languages. In *Proceedings of the INTERSPEECH 2022, Incheon, Korea, 18–22 September 2022*; pp. 4895–4899. [\[CrossRef\]](#)
5. Musaev, M.; Mussakhoyayeva, S.; Khujayorov, I.; Khassanov, Y.; Ochilov, M.; Varol, H.A. Multilingual Speech Recognition for Turkic Languages. *Information* **2023**, *14*, 74. [\[CrossRef\]](#)
6. Mussakhoyayeva, S.; Khassanov, Y.; Varol, H.A. KSC2: An Industrial-Scale Open-Source Kazakh Speech Corpus. In *Proceedings of the INTERSPEECH 2022, Incheon, Korea, 18–22 September 2022*; pp. 1367–1371. [\[CrossRef\]](#)
7. Khassanov, Y.; Mussakhoyayeva, S.; Mirzakhmetov, A.; Adiyev, A.; Nurpeiissov, M.; Varol, H.A. A Crowdsourced Open-Source Kazakh Speech Corpus and Initial Speech Recognition Baseline. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*; Association for Computational Linguistics: Kerrville, TX, USA, 2021; pp. 697–706. [\[CrossRef\]](#)

8. Salor, Ö.; Pellom, B.L.; Ciloglu, T.; Hacıoglu, K.; Demirekler, M. On Developing New Text and Audio Corpora and Speech Recognition Tools for the Turkish Language. In Proceedings of the 7th International Conference on Spoken Language Processing, Denver, CO, USA, 16–20 September 2002; pp. 349–352.
9. Zeyer, A.; Bahar, P.; Irie, K.; Schlüter, R.; Ney, H. A Comparison of Transformer and LSTM Encoder Decoder Models for ASR. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*; IEEE: New York, NY, USA, 2019; pp. 8–15. [CrossRef]
10. Sohn, J.; Kim, N.S.; Sung, W. A Statistical Model-Based Voice Activity Detection. *IEEE Signal Process. Lett.* **1999**, *6*, 1–3. [CrossRef]
11. Silero Team. Silero VAD: Pre-Trained Enterprise-Grade Voice Activity Detector. GitHub Repository. 2024. Available online: <https://github.com/snakers4/silero-vad> (accessed on 4 June 2026).
12. Yoshimura, T.; Hayashi, T.; Takeda, K.; Watanabe, S. End-to-End Automatic Speech Recognition Integrated with CTC-Based Voice Activity Detection. In *ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; IEEE: New York, NY, USA, 2020; pp. 6999–7003. [CrossRef]
13. Fathullah, Y.; Wu, C.; Lakomkin, E.; Jia, J.; Shangguan, Y.; Li, K.; Guo, J.; Xiong, W.; Mahadeokar, J.; Kalinli, O.; et al. Prompting Large Language Models with Speech Recognition Abilities. In *ICASSP 2024—2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; IEEE: New York, NY, USA, 2024; pp. 13351–13355. [CrossRef]
14. Ma, R.; Qian, M.; Manakul, P.; Gales, M.; Knill, K. Can Generative Large Language Models Perform ASR Error Correction? *arXiv* **2023**, arXiv:2307.04172. [CrossRef]
15. Nyssanov, N.; Syzdykova, Z.; Niyazaliyev, K.; Shayea, I. AI-Based Offline Speech Recognition for Kazakh, Russian and English Languages. In *Proceedings of the 2025 IEEE/ACIS 29th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*; IEEE: New York, NY, USA, 2025; pp. 1–5. [CrossRef]
16. Kadyrbek, N.; Mansurova, M.; Shomanov, A.; Makharova, G. The Development of a Kazakh Speech Recognition Model Using a Convolutional Neural Network with Fixed Character Level Filters. *Big Data Cogn. Comput.* **2023**, *7*, 132. [CrossRef]
17. Mamyrbayev, O.; Turdalyuly, M.; Mekebayev, N.; Alimhan, K.; Kydyrbekova, A.; Turdalykyzy, T. Automatic Recognition of Kazakh Speech Using Deep Neural Networks. In *Asian Conference on Intelligent Information and Database Systems*; Springer International Publishing: Cham, Switzerland, 2019; pp. 465–474. [CrossRef]
18. Rzayeva, L.; Nyssanov, N.; Syzdykova, Z.; Niyazaliyev, K.; Batkuldin, A.; Grigoryev, T. Multilingual Speech Transcription System for Kazakh, Russian and English Languages. *Artif. Intell. Res. Appl.* **2025**, *1*, 105–114. [CrossRef]
19. Yang, A.; Yang, B.; Hui, B.; Zheng, B.; Yu, B.; Zhou, C.; Li, C.; Li, C.; Liu, D.; Huang, F.; et al. Qwen2 Technical Report. *arXiv* **2024**, arXiv:2407.10671. [CrossRef]
20. Bisani, M.; Ney, H. Bootstrap estimates for confidence intervals in ASR performance evaluation. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*; IEEE: New York, NY, USA, 2004; pp. 409–412. [CrossRef]
21. Macháček, D.; Dabre, R.; Bojar, O. Turning Whisper into Real-Time Transcription System. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: System Demonstrations*; Association for Computational Linguistics: Kerrville, TX, USA, 2023; pp. 17–24. Available online: <https://aclanthology.org/2023.ijcnlp-demo.3/> (accessed on 4 June 2026).
22. Cahyawijaya, S.; Lovenia, H.; Aji, A.F.; Winata, G.I.; Wilie, B.; Koto, F.; Mahendra, R.; Wibisono, C.; Romadhony, A.; Vincentio, K.; et al. NusaCrowd: Open Source Initiative for Indonesian NLP Resources. In *Findings of the Association for Computational Linguistics: ACL 2023*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 13745–13818. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.