Selective Gating for LLM-Based Chinese ASR Error Correction

Yazhou Zhao
Ningxia University, Yinchuan, Ningxia, China

Abstract: This study examines over-correction in Chinese automatic speech recognition (ASR) post-processing with a general-purpose large language model (LLM). Whisper-small generates ASR hypotheses from Common Voice zh-CN, and Qwen2.5-3B-Instruct performs text correction. Three selective strategies-edit-ratio gating, ASR-confidence gating, and a conservative joint gate-are developed on a 100-utterance development set and evaluated with frozen settings on an independent 1,000-utterance test set. Direct LLM correction increases the macro-averaged character error rate (Macro CER) from 34.75% to 44.75%. The joint gate reduces the number of worsened utterances from 469 to 5, but yields a Macro CER of 34.79%, slightly above the Whisper baseline. These results show that, under the present setting, a small general-purpose instruction-tuned LLM is not a reliable unconditional ASR post-processor. Conservative gating substantially limits error propagation, but does not produce a net accuracy gain, highlighting the need for more reliable error detection and named-entity protection.

Keywords: Automatic Speech Recognition; Large Language Model; Error Correction; Whisper; Selective Gating; Character Error Rate

1. Introduction

Automatic speech recognition (ASR) underpins applications such as voice interaction, meeting transcription, and intelligent customer service. Whisper has demonstrated strong zero-shot transfer and robust multilingual recognition through large-scale weakly supervised training [1], while open resources such as Common Voice provide a reproducible foundation for multilingual ASR research [2]. Despite these advances, Chinese ASR remains susceptible to homophones, named entities, colloquial expressions, and acoustic ambiguity, resulting in substitutions, deletions, and entity-related recognition errors.

Large language models (LLMs), with their strong contextual modeling and generation capabilities, have increasingly been investigated for ASR error correction. Prior work has shown that text-based correction models can exploit phonological information to correct Chinese ASR errors [4], while LLM-based methods can incorporate N-best hypotheses and constrained decoding during post-processing [5]. A central difficulty, however, is that a generative model's linguistic prior is not direct evidence of the underlying speech content. Udagawa et al. showed that noisy correction data can induce over-correction in out-of-domain settings and proposed conservative data filtering to improve robustness [6]. A recent benchmark for Chinese ASR error correction likewise found that prompt-only LLM correction is not consistently effective, while fine-tuning and multimodal augmentation offer greater potential [7].

Against this background, we ask a practical question: without retraining either the ASR model or the LLM, and using only the 1-best ASR hypothesis and ASR confidence statistics, can a small general-purpose LLM reliably improve Chinese ASR output, and can conservative selection reduce the risk of harmful edits? We first evaluate direct correction with Qwen2.5-3B-Instruct [3] and then analyze failures in terms of manually categorized error types, LLM edit magnitude, and Whisper average log probability. We subsequently introduce edit-ratio gating, confidence gating, and a joint gating strategy. All thresholds are selected on a 100-utterance development set and frozen before a single evaluation on an independent 1,000-utterance test set, preventing test-set-driven threshold tuning.

2. Method and Experimental Setup

2.1 ASR and LLM Correction Pipeline

The pipeline comprises Whisper-based recognition, Qwen-based correction, text normalization, and selective output acceptance. Whisper-small is used as the ASR model with the language fixed to Chinese, temperature set to

 Copyright @ STEMM Institute Press

0, and condition_on_previous_text disabled. In addition to the 1-best hypothesis, segment-level statistics including avg_logprob, no_speech_prob, and compression_ratio are recorded. The simple average of segment-level avg_logprob is used as a proxy for recognition uncertainty. It is not a calibrated probability of correctness and is therefore used only as an empirical gating signal.

The correction model is a locally deployed Qwen2.5-3B-Instruct model with 4-bit NF4 quantization and deterministic greedy decoding. The prompt enforces a minimum-edit principle: only clear recognition errors should be corrected; sentence style should not be polished; no new information should be added; existing information should not be deleted; and uncertain content should remain unchanged. To reduce abnormal foreign-language generation, if the Chinese input contains no Latin letters but the LLM output introduces new Latin letters, the system falls back to the original ASR hypothesis. This is a manually designed heuristic and does not use the reference transcript.

2.2 Evaluation Metrics and Gating Strategies

Before evaluation, Traditional Chinese characters are converted to Simplified Chinese; text is lowercased; and spaces and common Chinese and English punctuation marks are removed. The primary metric is character error rate (CER). Given a reference R and a system hypothesis H, let S, D, and I denote the numbers of character substitutions, deletions, and insertions, respectively. CER is defined as (S + D + I) / N, where N is the number of characters in the reference. Both Macro CER and Corpus CER are reported. Macro CER is the arithmetic mean of sentence-level CER values, whereas Corpus CER is computed after accumulating edit counts over the entire corpus. Because the dataset contains many short utterances, Macro CER can be sensitive to a small number of short samples; the two metrics are therefore reported together.

To constrain over-editing, we define the edit ratio as edit_ratio = ED(H_ASR, H_LLM) / |H_ASR|, where ED denotes the character-level edit distance between the normalized ASR hypothesis and the LLM output. The edit gate accepts only candidates whose modifications are sufficiently small, whereas the confidence gate accepts an LLM output only when Whisper avg_logprob falls below a predefined threshold. The two confidence thresholds used in this study were selected separately because the standalone confidence gate and the joint gate optimize different decision rules. For the standalone confidence gate, we swept avg_logprob thresholds {-1.00, -0.80, -0.70, -0.60, -0.50, -0.45, -0.40, -0.35, -0.30, -0.25, -0.20} on the 100-utterance development set. A threshold of -0.60 yielded the lowest development-set Macro CER (38.30%) and was therefore frozen for the standalone gate. For the joint gate, we performed a two-dimensional sweep over confidence thresholds {-0.80, -0.70, -0.60, -0.50, -0.40} and edit-ratio thresholds {0.05, 0.10, 0.15, 0.20, 0.25, 0.30}. Although the combination (-0.60, 0.30) achieved the lowest development-set Macro CER, we selected the more conservative combination avg_logprob <= -0.70 and edit_ratio <= 0.10 because it produced 1 improved and 0 worsened utterances, with a Macro CER of 38.60% and a Corpus CER of 22.09%. Thus, -0.60 is the empirically best standalone confidence threshold under the development-set Macro-CER criterion, whereas -0.70 is part of a deliberately more conservative joint decision rule aimed at suppressing harmful LLM edits. Both settings were frozen before evaluation on the independent test set.

2.3 Dataset and Experimental Configuration

The experiments use the Chinese (China) (zh-CN) subset of Mozilla Common Voice Scripted Speech 26.0. The official data page reports 852,010 speech clips in this release, including 239.59 hours of validated speech [8]. The locally downloaded validated.tsv contains 190,017 valid records. We first randomly sample 100 utterances as a development set for prompt checking, error diagnosis, and gating-threshold selection. After excluding these audio files, 1,000 utterances are sampled from the remaining records with random_state = 2026 to form the independent test set. No overlap is observed between the two sets at the audio_path level. Speaker-level and transcript-level disjointness, however, were not explicitly verified; accordingly, the test set is not claimed to be speaker-disjoint.

3. Results and Analysis

3.1 Effect of Direct LLM Correction

On the development set, Whisper obtains a Macro CER of 38.67% and a Corpus CER of

Copyright @ STEMM Institute Press

22.18%. After directly accepting all Qwen outputs, the two metrics increase to 42.94% and 27.45%, respectively. Among the 100 utterances, only 8 improve, 51 remain unchanged, and 41 become worse. These results indicate that Qwen can correct some obvious recognition errors; however, under the present zero-shot correction setting with only the 1-best hypothesis as input, newly introduced errors outweigh successful corrections.

The same tendency is more evident on the independent test set. The Whisper baseline achieves 34.75% Macro CER and 24.50% Corpus CER, whereas Always LLM reaches 44.75% and 30.56%, respectively. Among the 1,000 test utterances, direct LLM correction improves only 72 utterances, worsens 469, and leaves 459 unchanged. Thus, general language modeling capability alone is insufficient to justify unconditional LLM-based ASR post-processing.

3.2 Error Types in LLM Over-Correction

To characterize the sources of degradation, we manually categorize the 41 development-set utterances whose CER increases after LLM correction. As shown in Table 1, semantic replacement accounts for 34.15%, content addition for 21.95%, and named-entity corruption for 17.07%. Semantic replacement and named-entity corruption together account for 21 of the 41 worsened utterances (51.22%), indicating that the major risk is not merely character-level variation but semantic rewriting driven by the model's language prior.

Table 1. Manual Categorization of LLM-Induced Errors on the Development Set

Error type	Count	Percentage
Semantic replacement	14	34.15%
Content addition	9	21.95%
Named-entity corruption	7	17.07%
Information deletion	5	12.20%
Hallucination	3	7.32%
Unnecessary rewriting	3	7.32%

For example, the LLM changes an already correct sentence meaning "The government began to mobilize the police" into one meaning "The government began to mobilize the troops," which is a semantic replacement rather than an ASR correction. It also tends to replace rare names or low-frequency expressions with more common alternatives. In a successful case, an ASR hypothesis containing two phonetically plausible but incorrect expressions was corrected to "Huangmei County" and the "Beijing–Kowloon Railway" when the surrounding context provided sufficiently strong lexical constraints. These examples suggest that the usefulness of LLM correction is highly context dependent: linguistic priors can help when the intended expression is strongly constrained by context, but may override transcription fidelity for named entities or semantically ambiguous inputs. The manual labels were produced by a single annotator, and no inter-annotator agreement was computed; therefore, this analysis is intended to characterize error mechanisms rather than support independent statistical inference.

3.3 Development-Set Performance of Selective Gating

Given the poor performance of unconditional correction, we next examine whether selectively accepting LLM modifications can reduce the associated risk. On the development set, the mean edit_ratio values are 0.1421 for improved utterances, 0.0881 for unchanged utterances, and 0.1806 for worsened utterances, but the three distributions overlap substantially. Therefore, a small edit does not necessarily imply a correct edit. Similarly, avg_logprob values overlap across improved, unchanged, and worsened groups. Low ASR confidence indicates that the ASR hypothesis may be unreliable, but does not imply that the LLM has enough information to reconstruct the correct transcript. The development-set results of the four correction strategies are summarized in Table 2.

Table 2. Development-Set Results of Different Correction Strategies

Method	Macro CER	Corpus CER	LLM accepted	Improved	Worsened
Whisper	38.67%	22.18%	0	-	-
Always LLM	42.94%	27.45%	100	8	41
Edit Gate	38.71%	22.26%	3	1	2
Confidence Gate	38.30%	22.34%	31	3	5
Joint Gate	38.60%	22.09%	15	1	0

The edit gate accepts only three candidates and greatly limits the modification range, but two of the accepted candidates still become worse. The confidence gate accepts 31 candidates and

 Copyright @ STEMM Institute Press

reduces Macro CER to 38.30%, while its Corpus CER is 22.34% and therefore does not outperform Whisper on that metric. The final joint gate uses a more conservative threshold combination, accepts 15 candidates, and produces no worsened utterances on the development set, with Macro CER and Corpus CER of 38.60% and 22.09%, respectively. This marginal development-set gain is used only for strategy selection and is not interpreted as evidence of a stable performance improvement.

3.4 Independent Test and Generalization Analysis

After freezing the prompt, model configuration, and gating thresholds, we evaluate the system once on the independent 1,000-utterance test set without further adjustment. The main results are summarized in Table 3.

Table 3. Main Results on the Independent Test Set

Method	Macro CER	Corpus CER	LLM accepted	Improved	Unchanged	Worsened
Whisper	34.75%	24.50%	0	-	-	-
Always LLM	44.75%	30.56%	1000	72	459	469
Confidence Gate	39.26%	25.28%	243	16	907	77
Final Joint Gate	34.79%	24.53%	91	0	995	5

Confidence Gate accepts 243 LLM candidates on the test set, of which 16 improve and 77 worsen. Its Macro CER and Corpus CER are 39.26% and 25.28%, respectively, both worse than the Whisper baseline. Final Joint Gate further reduces the accepted set to 91 utterances, corresponding to 9.10% of the test set. Only five utterances are worsened, far fewer than the 469 worsened by Always LLM. This result shows that the joint constraints substantially restrict the scope of harmful edits introduced by generative correction.

However, this reduction in harmful edits does not translate into a net gain in recognition accuracy. Final Joint Gate obtains 34.79% Macro CER and 24.53% Corpus CER, both slightly higher than the Whisper baseline values of 34.75% and 24.50%. None of the 91 accepted candidates achieves a sentence-level CER improvement. Consequently, the reduction from 469 worsened utterances to 5 primarily reflects the rejection of most LLM modifications and should not be described as an improvement in correction accuracy. No statistical significance test is conducted for the small difference between 34.75% and 34.79%, and we therefore do not interpret this difference as statistically meaningful.

The joint gate shows a very small CER advantage on the development set, but this advantage does not reproduce on the independent test set. This suggests that thresholds derived from only 100 development utterances are sample-dependent, particularly because Macro CER can be sensitive to short utterances and a small number of changed examples. More importantly, the result clarifies the limitation of the proposed signals: avg_logprob and edit_ratio can filter risky candidates, but they cannot determine whether an LLM correction is actually correct. Stable net gains would require more direct acoustic evidence or a stronger estimate of error correctability.

3.5 Discussion

From a modeling perspective, ASR error correction requires fidelity to the underlying speech signal, whereas a general-purpose instruction-tuned LLM is primarily optimized to generate linguistically plausible text. These objectives are not fully aligned. When an ASR hypothesis contains a local homophone error and the context strongly constrains the intended expression, the LLM's language prior can be beneficial. When the input contains rare entities, proper names, or missing acoustic information, however, a text-only LLM cannot reliably distinguish between an ASR mistake and an inherently uncommon expression in the original speech, which can lead to semantic replacement and hallucination.

The results therefore favor selective, risk-aware correction over unrestricted LLM rewriting. Future work can proceed in three directions. First, ASR N-best hypotheses or token-level acoustic confidence can provide richer candidate and acoustic information. Second, named-entity dictionaries, entity protection, and domain constraints may reduce active rewriting of low-frequency entities. Third, supervised fine-tuning on ASR error pairs or audio-text multimodal correction can better align the correction model with the ASR task. Recent Chinese ASR error-correction research also suggests that multimodal augmentation is more promising

Copyright @ STEMM Institute Press

than prompt-only correction [7].

4. Conclusion

This study evaluates Qwen2.5-3B-Instruct as a post-processor for Whisper-small Chinese ASR and investigates edit-ratio gating, ASR-confidence gating, and a joint gating strategy. On the independent test set, direct LLM correction increases Macro CER from 34.75% to 44.75% and worsens 469 utterances, showing that a small general-purpose LLM is unsuitable as an unconditional ASR post-processor under the present setting. Manual analysis identifies semantic replacement, content addition, and named-entity corruption as the major risks. The joint gate limits the LLM acceptance rate to 9.10% and reduces the number of worsened utterances to five, but its Macro CER and Corpus CER remain at 34.79% and 24.53%, respectively, providing no net accuracy gain. Thus, the main value of conservative gating is risk control rather than direct recognition improvement. Future work should incorporate finer-grained acoustic evidence, entity constraints, and ASR-specific correction training to better determine when a hypothesis should be corrected and when it should be preserved.

References

[1] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," in Proceedings of the 40th International Conference on Machine Learning, PMLR, vol. 202, pp. 28492–28518, 2023.

[2] R. Ardila, M. Branson, K. Davis, et al., "Common Voice: A Massively-Multilingual Speech Corpus," in Proceedings of the Twelfth Language Resources and Evaluation Conference, Marseille, France: European Language Resources Association, pp. 4218–4222, 2020.

[3] A. Yang, B. Yang, B. Zhang, et al., "Qwen2.5 Technical Report," arXiv:2412.15115, 2024.

[4] Z. Fang, R. Zhang, Z. He, H. Wu, and Y. Cao, "Non-Autoregressive Chinese ASR Error Correction with Phonological Training," in Proceedings of NAACL 2022: Human Language Technologies, Seattle, United States: Association for Computational Linguistics, pp. 5907–5917, 2022, doi: 10.18653/v1/2022.naacl-main.432.

[5] R. Ma, M. Qian, M. Gales, and K. Knill, "ASR Error Correction using Large Language Models," arXiv:2409.09554, 2024.

[6] T. Udagawa, M. Suzuki, M. Muraoka, and G. Kurata, "Robust ASR Error Correction with Conservative Data Filtering," in Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track, Miami, Florida, US: Association for Computational Linguistics, pp. 256–266, 2024.

[7] V. J. Wei, W. Wang, D. Jiang, Y. Song, and L. Wang, "ASR-EC Benchmark: Evaluating Large Language Models on Chinese ASR Error Correction," in Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track, Suzhou, China: Association for Computational Linguistics, pp. 1567–1575, 2025, doi: 10.18653/v1/2025.emnlp-industry.110.

[8] Mozilla Data Collective, "Common Voice Scripted Speech 26.0 - Chinese (China)," dataset, 2026.

 Copyright @ STEMM Institute Press