KGFC: Keyword-Guided Fusion Correction for Classroom ASR with the SC-ACD Benchmark

Anonymous ACL submission

Abstract

Automatic speech recognition (ASR) systems often degrade sharply in real classrooms because of background noise, overlapping speech, accent variation, and dense subject-specific terminology. These errors are especially harmful in educational settings, where missing or distorted keywords can change the meaning of instructional content. We present **K**eyword-**G**uided **F**usion **C**orrection (KGFC), a lightweight correction model that combines a global representation of the ASR hypothesis with local keyword cues. KGFC is built on a compact 0.5B language model with LoRA adaptation and introduces a fusion gate to balance fluency and keyword fidelity. We also introduce the **S**hanghai **C**lassroom **A**SR **C**orrection **D**ataset (SC-ACD), a classroom benchmark with manually verified transcripts and keyword annotations. Experiments on SC-ACD, noisy AISHELL data, and Chinese spelling-correction benchmarks show that KGFC consistently improves word error rate, character error rate, and keyword coverage over sequence-to-sequence, pretrained correction, and prompt-based baselines. Additional ablations indicate that the keyword encoder, fusion gate, and keyword-aware training objective all make complementary contributions.

1 Introduction

Automatic speech recognition has become increasingly important in education, where it supports applications such as tutoring analytics, classroom observation, and post-class content review (Booth et al., 2024; Southwell et al., 2022). However, real classroom speech is much harder to transcribe than clean read speech. Noise, speaker overlap, pronunciation variation, and rapid topic shifts all increase recognition difficulty (Kothalkar, 2023). In addition, classroom discourse contains dense subject-specific terminology, and standard ASR systems often fail exactly on these high-value words.

This failure mode matters in practice. A transcript can remain broadly fluent while still losing key scientific, mathematical, or historical terms, which makes it less useful for downstream educational analysis. Existing correction methods only partially address this problem. Sequence-to-sequence post-editing models improve local consistency (Sundararaman et al., 2019), reranking methods exploit multiple hypotheses (Chen et al., 2021), and recent large language model approaches improve fluency (Pu et al., 2023). But these methods often lack an explicit mechanism for preserving domain keywords, or they are too expensive for practical classroom deployment.

A classroom transcript is therefore not judged only by overall readability. In many educational scenarios, the most important requirement is that the transcript preserve a small set of pedagogically critical terms: theorem names, scientific concepts, proper nouns, or domain formulas. If these terms are deleted, substituted, or normalized into vague paraphrases, the transcript can become much less useful for retrieval, summarization, or classroom analytics even when the surrounding sentence remains grammatical. This motivates a correction objective that explicitly values content fidelity in addition to general fluency.

We therefore study a more targeted objective: classroom ASR correction that improves transcript quality while explicitly protecting educationally important keywords. Our approach is based on a simple intuition. The raw ASR hypothesis already provides useful global sentence semantics, but class-

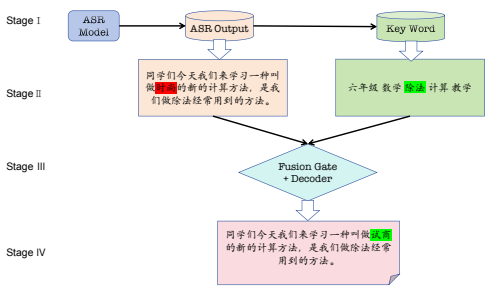

Figure 1: Illustrative classroom correction case. The corrected transcript must preserve key instructional terms rather than only improving local fluency.

room correction also needs a local bias toward terminology that may be rare, domain-specific, or acoustically confusable. KGFC combines these two signals through a keyword-guided fusion mechanism.

KGFC is intentionally lightweight. We build it on Qwen-0.5B and fine-tune it with LoRA (Hu et al., 2022), which keeps adaptation practical while still allowing the model to learn keyword-sensitive correction behavior. The model uses a dedicated keyword encoder, a fusion gate that balances global and local information, and an objective that combines transcript accuracy, keyword coverage, and phoneme consistency. To support evaluation, we construct the Shanghai Classroom ASR Correction Dataset (SC-ACD), derived from large-scale classroom recordings with manually verified transcripts and keyword annotations.

Our narrative is deliberately modest. We do not claim to solve general ASR robustness or broad educational speech understanding. Instead, we focus on a concrete and practically important failure mode: correction systems that improve sentence fluency but still lose the terms that matter most for instructional meaning. This framing keeps the contribution specific and testable, and it also aligns with the empirical design of SC-ACD and the keyword-coverage analysis used throughout the paper.

Our contributions are threefold:

- We propose KGFC, a lightweight classroom ASR correction model that explicitly fuses ASR-semantic context with keyword guidance.
- We introduce SC-ACD, a benchmark for classroom ASR correction with manually verified references and keyword annotations across diverse school subjects.
- We show that KGFC improves WER, CER, and keyword coverage over strong correction baselines on SC-ACD and transfers well to noisy speech and related text-correction settings.

2 Related Work

ASR error correction. Post-ASR correction has been studied with sequence-to-sequence architectures, reranking approaches, and pretrained correction models (Sundararaman et al., 2019; Chen et al., 2021; Cui et al., 2020). These methods improve transcript quality when the error distribution is relatively stable, but they do not directly model the classroom need to preserve subject terminology. Keyword-aware adaptation has also been explored for ASR systems themselves (Shamsian et al., 2024), which supports the broader motivation that domain terms deserve explicit modeling.

Robust speech recognition in noisy or educational settings. Classroom speech remains challenging because it combines adverse acoustics with dynamic pedagogical interaction (Southwell et al., 2022; Kothalkar, 2023). Large-scale weakly supervised ASR models improve robustness (Radford et al., 2023), but downstream correction is still valuable when applications require high transcript fidelity on domain terms.

Lightweight adaptation and deployment. Large language models can be effective post-editors, but cost and latency often prevent classroom deployment. Parameter-efficient adaptation methods such as LoRA (Hu et al., 2022) provide a practical alternative, allowing smaller models to specialize without full fine-tuning. KGFC follows this line by pairing a compact base model with a task-specific fusion design instead of relying on expensive general-purpose LLM prompting.

Educational transcript fidelity. Our motivation is also related to a broader distinction between *fluency-oriented correction* and *content-faithful correction*. In educational transcripts, preserving a small number of highly informative terms can be more important than optimizing average sentence naturalness. KGFC is designed around this distinction: the ASR pathway maintains global sentence structure, while the keyword pathway acts as an

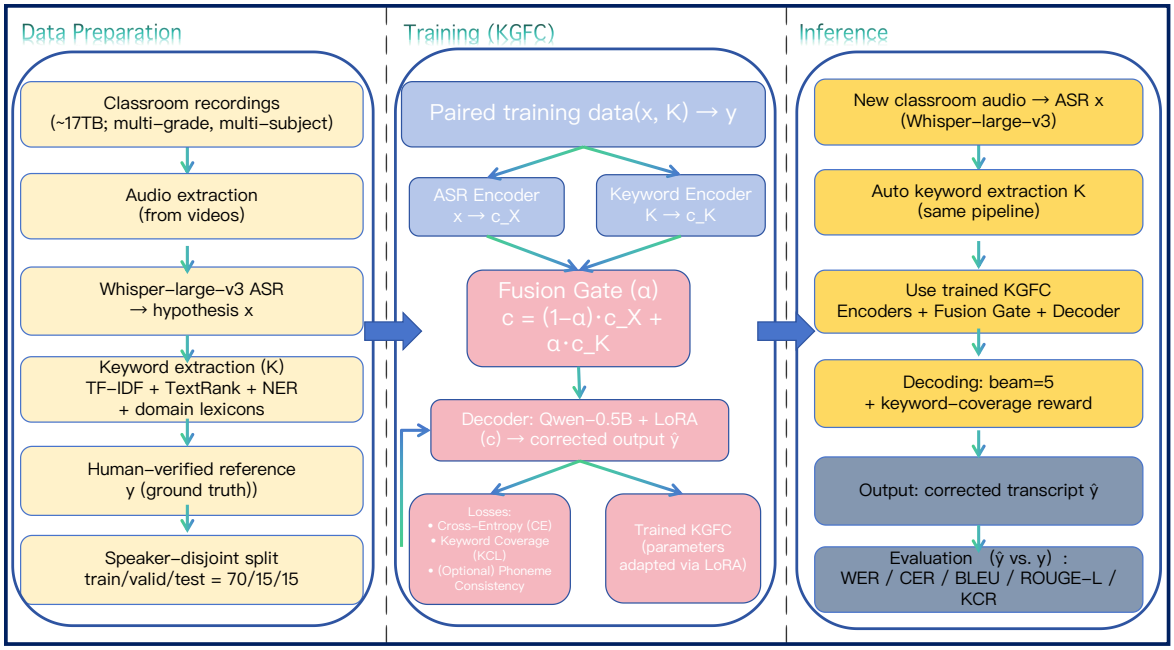

Figure 2: Overview of KGFC and the SC-ACD data pipeline. The system constructs paired ASR hypotheses, verified references, and keyword annotations; encodes the hypothesis and keywords separately; and fuses them with a learnable gate before decoding.

explicit reminder that some local content items are disproportionately valuable.

3 Method

KGFC takes an ASR hypothesis x and a keyword set K as input, and predicts a corrected transcript y. Figure 2 shows the overall design. The method separates the correction problem into four stages: input modeling, dual-path encoding, gated fusion, and keyword-aware decoding. Formally, the model can be viewed as learning

$$y = f_\theta(x, K), \tag{1}$$

where x contributes sentence-level semantic evidence and K provides an explicit local constraint over important terms.

3.1 Input Modeling

The ASR hypothesis provides a *global semantic prior*: even noisy output usually preserves coarse sentence meaning and discourse flow. The keyword set provides a *local correction bias*: it highlights terms that are likely to be educationally important and more costly to lose. During training, keywords are manually verified; during inference, they can be obtained from automatic extraction pipelines based on TF-IDF, TextRank, named-entity recognition, and domain lexicons.

This two-input design reflects a practical observation about classroom errors. Many utterances are not globally unintelligible; instead, they contain a few high-impact local failures. A purely sentence-level corrector may smooth these errors into more common phrases rather than restoring the intended terminology. By injecting a keyword set directly into the model, KGFC turns this weak point into an explicit modeling target.

3.2 Dual-Path Encoding

KGFC encodes the two inputs separately. The ASR hypothesis is mapped to a contextual sentence representation c_X, while the keyword set is embedded and passed through a lightweight transformer encoder to produce a keyword representation c_K. This separation is useful because the two signals play different roles. The hypothesis captures broader fluency and discourse constraints, while the keyword pathway preserves local content cues that standard correction models may underweight.

The separation is also useful from an optimization perspective. If both signals are mixed too early, the model may collapse toward the more frequent or easier-to-learn source, which is usually the sentence context. A dedicated keyword encoder preserves a distinct representational channel for do-

main terms and makes it easier for downstream layers to decide when these terms should override the broader sentence prior.

3.3 Fusion Gate

We combine the two representations with a learnable fusion gate:

$$c = (1 - \alpha)c_X + \alpha c_K, \tag{2}$$

where $\alpha \in [0, 1]$ controls the relative importance of sentence-level semantics and keyword cues. Intuitively, the model can increase keyword reliance when the hypothesis appears noisy or underspecified, while relying more on the hypothesis when the sentence context is already clear.

This gated formulation is intentionally simple. More complex cross-attention or iterative fusion mechanisms are possible, but the present design already captures the core trade-off between fluency and terminology fidelity. It also keeps the model lightweight enough for the deployment scenarios that motivate this work.

3.4 Training Objective

The decoder is initialized from Qwen-0.5B and adapted with LoRA. Training uses three complementary objectives. First, cross-entropy aligns the generated transcript with the reference. Second, a keyword coverage loss penalizes omissions or distortions of important terms. Third, a phoneme-consistency regularizer discourages unnecessary edits that drift too far from the original pronunciation pattern (Mu et al., 2024). Together, these objectives encourage a balance between fluency, faithfulness, and robustness:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_1 \mathcal{L}_{\text{kw}} + \lambda_2 \mathcal{L}_{\text{ph}}, \tag{3}$$

where λ_1 and λ_2 control the influence of keyword fidelity and phonetic regularization.

3.5 Decoding and Inference

At inference time, KGFC receives a new ASR hypothesis and a keyword set produced by the automatic extraction pipeline. The decoder then generates the corrected transcript using the same fused representation used during training. The original setup further applies beam search with a keyword-aware reward so that hypotheses preserving more validated terms are preferred. This choice is important because it aligns decoding with the same fidelity objective already encouraged by training.

4 SC-ACD Dataset

SC-ACD is designed to reflect practical classroom correction conditions rather than idealized clean speech. The corpus is derived from about 17TB of classroom recordings spanning primary, middle, and high school settings. It covers multiple disciplines, including language, mathematics, English, physics, chemistry, biology, history, geography, politics, arts, and information technology.

Each sample contains an ASR hypothesis generated by Whisper-style transcription, a manually verified reference transcript, and a small set of validated keywords. The dataset is split into train, validation, and test portions with disjoint speakers. Raw ASR outputs remain challenging, with high baseline WER and CER, which makes the benchmark suitable for studying both robustness and keyword preservation.

SC-ACD differs from standard speech or spelling-correction corpora in two ways. First, it explicitly targets educational discourse with subject terminology. Second, it pairs correction targets with keyword annotations, which makes it possible to measure whether a model preserves semantically critical content rather than only improving surface fluency.

4.1 Collection and Annotation

The dataset contains more than 30,000 manually verified transcripts, and each sample is paired with keyword annotations derived from a combination of lexical extraction and human validation. During construction, the annotation process emphasizes educationally meaningful terms rather than arbitrary high-frequency tokens. This is important because the benchmark is meant to approximate downstream classroom use cases, where preserving the right concept names matters more than preserving generic function words.

4.2 Benchmark Properties

SC-ACD is difficult for at least three reasons. First, classroom audio contains environmental noise and frequent speaker overlap. Second, the discourse is heterogeneous across subjects and grade levels. Third, the benchmark rewards systems that preserve local terminology, which is a stricter goal than optimizing transcript fluency alone. These properties make SC-ACD complementary to both standard noisy-ASR evaluation and text-only correction benchmarks.

Table 1: Structured summary of SC-ACD.

Aspect	Summary
Source	About 17TB of classroom recordings from primary, middle, and high schools
Annotation unit	ASR hypothesis, verified reference transcript, and keyword annotations
Domain coverage	Language, mathematics, English, science, humanities, arts, and IT
Split	70% train / 15% validation / 15% test with disjoint speakers
Raw difficulty	Baseline ASR remains noisy, motivating post-correction
Keyword protocol	3–5 keywords from TF-IDF, TextRank, and NER, then manually checked

Table 1 highlights that SC-ACD combines both scale and heterogeneity. The benchmark is not only large, but also structurally aligned with the classroom correction problem because it provides paired references and keyword annotations across multiple school subjects.

5 Experiments

5.1 Setup

Datasets. We evaluate on SC-ACD as the main benchmark, on a noisy subset of AISHELL-1 for cross-domain speech robustness (Bu et al., 2017), and on Wang271K and LEMON to compare against Chinese correction-oriented baselines (Wang et al., 2018; Wu et al., 2023).

Baselines. We compare against the uncorrected ASR output (**Original**), a transformer sequence-to-sequence corrector (**Seq2Seq**) (Sundararaman et al., 2019), **MacBERT** (Cui et al., 2020), and a prompt-based **GPT-4o** correction setting motivated by recent LLM post-editing work (Pu et al., 2023).

Metrics. We report word error rate (WER), character error rate (CER), and keyword coverage rate (KCR). Lower WER/CER and higher KCR are better. Operationally, KCR measures the proportion of reference keywords retained by the corrected transcript:

$$\mathrm{KCR} = \frac{\#\ \text{preserved reference keywords}}{\#\ \text{reference keywords}}. \quad (4)$$

This metric is central because the motivating educational failure mode is often a local terminology loss rather than a globally incoherent sentence.

Implementation. KGFC is built on Qwen-0.5B with LoRA adaptation. The original study uses rank 8, $\alpha = 16$, dropout 0.05, AdamW, and beam search with a keyword-aware reward at inference time.

Table 2: Implementation summary for KGFC.

Component	Configuration
Base model	Qwen-0.5B
Adaptation	LoRA fine-tuning
Decoding	Beam search with keyword-aware reward
Objectives	Cross-entropy, keyword coverage loss, phoneme-consistency regularization
Inference keywords	Automatic extraction with lexical and entity cues
Main metrics	WER, CER, KCR

Evaluation design. The benchmark suite intentionally mixes one in-domain classroom dataset with out-of-domain noisy speech and text-correction datasets. This allows us to test two different hypotheses: whether KGFC is best aligned with the classroom setting it was designed for, and whether keyword-guided correction still helps when the data distribution shifts.

Table 2 shows that the proposed system remains compact. This compactness is relevant to the overall claim of the paper because the method is intended as a realistic classroom correction pipeline rather than a purely scaling-based solution.

5.2 Main Results

Table 3 reports the main results. KGFC achieves the strongest overall performance on all reported datasets. On SC-ACD, it reduces WER from 22.4 to 8.71 and raises KCR to 93.4, substantially ahead of both classical correction models and the prompt-based LLM baseline. The same pattern appears on noisy AISHELL, where KGFC yields the lowest WER and CER while maintaining strong keyword coverage.

These gains suggest that keyword guidance is doing more than simply making the output more fluent. If the model were only improving general language quality, we would expect smaller improvements on KCR. Instead, KGFC produces the largest margins exactly on the metric that reflects preservation of educationally important content.

Table 4 makes the improvement pattern more explicit. The gains on CER are moderate but consistent, while the gains on KCR are large on every benchmark. This is exactly the behavior we would expect from a keyword-aware model: once sentence-level fluency is already competitive, the main remaining benefit comes from preserving the

Table 3: Main results. WER/CER are percentages; higher KCR is better.

Dataset	Model	WER	CER	KCR
SC-ACD	Original	22.40	13.70	68.2
	Seq2Seq	12.63	6.74	72.4
	MacBERT	10.38	5.91	72.6
	GPT-4o	10.14	5.82	75.1
	KGFC	**8.71**	**4.97**	**93.4**
AISHELL	Original	13.20	7.42	60.2
	Seq2Seq	8.54	4.61	72.1
	MacBERT	7.91	4.35	73.2
	GPT-4o	7.43	4.10	76.1
	KGFC	**6.85**	**3.87**	**91.7**
LEMON	MacBERT	–	7.20	72.3
	GPT-4o	–	6.85	74.7
	KGFC	–	**6.50**	**88.6**
Wang271K	MacBERT	–	5.10	68.3
	GPT-4o	–	4.95	72.8
	KGFC	–	**4.70**	**90.2**

Table 4: KGFC gains over the strongest non-KGFC baseline on each benchmark. Negative WER/CER changes are improvements.

Dataset	ΔWER	ΔCER	ΔKCR
SC-ACD	-1.43	-0.85	+18.3
AISHELL	-0.58	-0.23	+15.6
LEMON	–	-0.35	+13.9
Wang271K	–	-0.25	+17.4

content items that general correction systems still overlook.

5.3 Data Efficiency

An additional question is whether keyword guidance remains useful when the amount of supervised classroom correction data is limited. Figure 3 summarizes the error-rate trend under different training sample sizes using the supplementary project results. KGFC remains the strongest method across the full range and degrades more gracefully than the comparison systems when the number of labeled samples is small.

This behavior is important for educational deployment. Collecting and manually validating classroom transcripts is expensive, and many institutions will not have access to large correction corpora. A method that only works after substantial in-domain annotation would therefore be less attractive in practice. The observed trend suggests that keyword guidance provides a useful inductive bias even before the model sees a large amount of classroom-specific supervision.

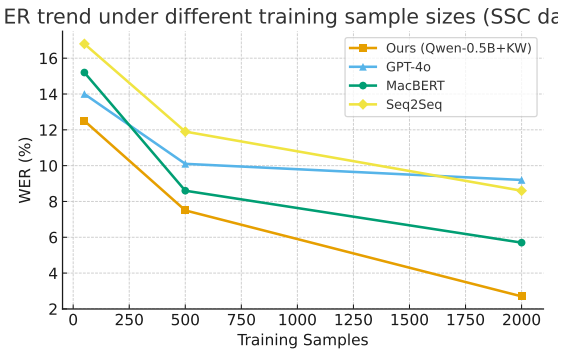

Figure 3: WER trend under different training sample sizes. KGFC remains consistently better than the comparison methods, and the relative advantage is especially visible in lower-resource settings.

Table 5: Ablation study on SC-ACD.

Variant	WER	CER	KCR
KGFC	**8.71**	**4.97**	**93.4**
w/o keyword encoder	10.61	6.01	–
w/o fusion gate	12.11	6.89	–
w/o keyword coverage loss	9.91	5.63	88.2
Prompt-only correction	11.94	6.77	90.3

5.4 Ablation Study

We next isolate the contribution of each component on SC-ACD. Table 5 shows that removing the keyword encoder or the fusion gate substantially hurts WER and CER, and removing the keyword coverage loss lowers KCR by more than five points. The full model therefore benefits from both architectural and objective-level keyword guidance.

5.5 Analysis

The main empirical takeaway is that classroom correction benefits from explicitly separating global fluency and local terminology preservation. Prompt-based LLM correction is already competitive on WER and CER, but it trails KGFC on KCR, which suggests that generic generation quality does not guarantee pedagogically useful correction. The gain is also consistent with the motivating case in Figure 1: educational transcripts are often judged by whether key concepts survive the pipeline.

KGFC also appears to transfer beyond the exact classroom setting. The gains on AISHELL and the Chinese correction benchmarks suggest that keyword guidance is not overfitting to one corpus. At the same time, the strongest improvements occur on SC-ACD, which supports the claim that dataset design and model design are aligned around the classroom use case.

5.6 Cross-domain Generalization

The benchmark suite reveals an encouraging pattern: KGFC helps both in-domain and out-of-domain settings, but the type of gain differs across datasets. On SC-ACD and AISHELL, the method improves both error rates and keyword coverage. On the text-oriented LEMON and Wang271K benchmarks, the clearest effect appears on CER and KCR. This suggests that keyword guidance is not tied to one specific acoustic domain. Instead, it acts as a generic bias toward preserving informative local content, which remains useful even when the surface task shifts from speech correction to related character-level correction.

5.7 Why KCR Improvements Matter

The KCR margins in Table 4 are particularly important because they are much larger than the corresponding CER improvements. If KGFC were only a stronger generic decoder, we would expect similar relative gains across all metrics. Instead, the method improves most on the metric that operationalizes its design objective. This makes the evidence more persuasive than a small average error-rate gain alone.

5.8 Interpreting the Ablations

The ablation results also separate the roles of the three main design choices. Removing the keyword encoder weakens the model's ability to represent local terminology. Removing the fusion gate harms performance even more, indicating that simply exposing the model to keywords is not enough; the model also needs a mechanism to balance them against noisy sentence context. Finally, removing keyword coverage loss reduces KCR substantially, which shows that the training objective matters in addition to the architecture. These trends support the broader claim that KGFC is not just a smaller LLM baseline with auxiliary text pasted into the prompt.

5.9 Deployment Considerations

KGFC is designed for a practical operating regime. Relative to large prompt-based correction systems, its main advantage is not only accuracy but predictability: the keyword set offers an explicit, inspectable signal about what the model should protect. This can be valuable in education, where practitioners often want to understand *why* a transcript changed. A lightweight LoRA-adapted model also offers a more realistic deployment path for repeated classroom processing than repeatedly querying a very large general-purpose model.

5.10 Boundary Cases

KGFC is not expected to help equally in all settings. If keyword extraction is poor, the model may overemphasize irrelevant terms. Likewise, if the input utterance is extremely noisy or the reference keywords are too sparse, the keyword pathway may have limited room to help. Another boundary case arises when the correct output genuinely requires paraphrasing rather than preserving the extracted term verbatim. These cases suggest that future work should study confidence-aware keyword selection and more flexible notions of content preservation.

6 Conclusion

This paper addressed a specific but practically important weakness of classroom ASR: many correction systems can improve surface fluency while still failing to preserve the subject terms that carry most of the instructional meaning. To study this problem, we introduced SC-ACD, a classroom ASR correction benchmark built from large-scale educational recordings with manually verified references and keyword annotations. We also proposed KGFC, a keyword-guided fusion model that keeps the correction pipeline lightweight while making terminology preservation an explicit modeling objective.

Across SC-ACD, noisy AISHELL speech, and related Chinese correction benchmarks, the empirical picture is consistent. KGFC improves standard transcript quality metrics such as WER and CER, but its most distinctive advantage appears on keyword coverage. This matters because educational usefulness is often determined less by whether every sentence sounds maximally fluent than by whether critical concepts, formulas, names, and domain expressions survive the recognition-and-correction pipeline. The ablation results reinforce the same story: the keyword encoder, fusion gate, and keyword-aware objective each contribute to the observed gains, which suggests that the improvement comes from the overall design rather than from scale alone.

More broadly, we view this work as an argument for task-aware ASR correction. In classroom settings, evaluation criteria should reflect the fact that some local content items are disproportionately valuable. A correction model that is explicitly

biased toward preserving these items can therefore be more useful than a stronger generic generator that optimizes average fluency alone. We hope SC-ACD provides a useful benchmark for this perspective and that KGFC offers a simple, extensible baseline for future work on educational speech systems.

The benchmark contribution is also important in its own right. Many existing correction datasets emphasize either clean transcription recovery or generic text normalization, which makes it difficult to isolate the educational failure cases that motivate this work. By pairing noisy classroom hypotheses with manually verified references and validated keyword annotations, SC-ACD makes it possible to evaluate not only whether a model reduces average error, but also whether it preserves the instructional content that downstream users actually care about. We expect this benchmark structure to be useful beyond KGFC itself, including for retrieval-augmented correction, constrained decoding, terminology-aware reranking, and future multimodal systems that combine speech with slides, blackboard content, or lesson plans.

From a systems perspective, the results also suggest a practical compromise between specialized correction quality and deployment cost. Educational institutions often need a pipeline that can be run repeatedly, audited by instructors, and adapted to new subjects without the expense of relying on a very large general-purpose model for every transcript. KGFC points to one feasible design pattern for that setting: a compact adapted decoder supplies the main correction capacity, while an explicit keyword channel provides a transparent mechanism for protecting high-value terms. This combination is useful not only because it improves accuracy, but also because it offers a clearer interface for human oversight. Teachers or curriculum designers can inspect the keywords, understand what the model is being encouraged to preserve, and refine the constraint source when subject terminology changes over time.

Several promising directions follow from this study. Future research could extend keyword guidance to multilingual classrooms, multimodal lecture settings, and longer discourse segments where context spans many turns. It would also be useful to replace fixed keyword lists with confidence-aware or dynamically generated content constraints, and to evaluate whether better correction translates into downstream gains for tutoring analytics, retrieval, summarization, or classroom search. In this sense, we see KGFC not as a complete solution to robust educational ASR, but as a practical step toward correction systems that are judged by the information they preserve, not only by the fluency they produce.

7 Limitations

This work has several limitations. First, SC-ACD is large and diverse, but it remains focused on Chinese classroom speech and may not capture the full range of multilingual or multimodal educational settings. Second, the keyword annotations are central to the method, so automatic keyword extraction quality can affect downstream correction. Third, KGFC is lightweight relative to large LLM baselines, but it still depends on curated correction data and a task-specific training pipeline. Finally, keyword coverage is only one proxy for educational usefulness; future work should study whether the corrected transcripts improve downstream tutoring, analytics, or search tasks.

We also emphasize that KGFC targets a narrow scientific claim. The method is designed to reduce a specific correction failure mode, namely the loss of high-value local content in otherwise usable transcripts. It does not directly address speaker diarization, multimodal grounding, or long-context discourse consistency. These remain open problems for future classroom speech systems.

References

Brandon M. Booth, Jennifer Jacobs, Jeffrey B. Bush, Brent Milne, Tom Fischaber, and Sidney K. DMello. 2024. Human-tutor coaching technology (htct): Automated discourse analytics in a coached tutoring model. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*, LAK '24, page 725–735, New York, NY, USA. Association for Computing Machinery.

Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. 2017. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In *2017 20th conference of the oriental chapter of the international coordinating committee on speech databases and speech I/O systems and assessment (O-COCOSDA)*, pages 1–5. IEEE.

Ziqiang Chen, Bo Li, and Bo Xu. 2021. A comparative study of n-best list re-ranking strategies for asr error correction. In *ICASSP 2021 - IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7418–7422. IEEE.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 657–668, Online. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Prasanna V Kothalkar. 2023. Teacher-student learning based acoustic models for robust speech recognition in naturalistic childhood classroom settings. *Speech communication*.

Bingshen Mu, Yangze Li, Qijie Shao, Kun Wei, Xucheng Wan, Naijun Zheng, Huan Zhou, and Lei Xie. 2024. Mmger: Multi-modal and multi-granularity generative error correction with llm for joint accent and speech recognition. *Preprint*, arXiv:2405.03152.

Jie Pu, Thai-Son Nguyen, and Sebastian Stüker. 2023. Multi-stage large language model correction for speech recognition. *arXiv preprint arXiv:2310.11532*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Aviv Shamsian, Aviv Navon, Neta Glazer, Gill Hetz, and Joseph Keshet. 2024. Keyword-guided adaptation of automatic speech recognition. *Preprint*, arXiv:2406.02649.

Rosy Southwell, Samuel Pugh, M Perkoff, Charis Clevenger, Jeffrey Bush, Rachel Lieber, Wayne Ward, Peter Foltz, and Sidney D'Mello. 2022. Challenges and feasibility of automatic speech recognition for modeling student collaborative discourse in classrooms. *International Educational Data Mining Society*.

Anurag Sundararaman, Sandeep Kumar, and Sanjeev Khudanpur. 2019. Sequence-to-sequence models for asr error correction. In *ICASSP 2019 - IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5651–5655. IEEE.

Dingmin Wang, Yan Song, Jing Li, Jialong Han, and Haisong Zhang. 2018. A hybrid approach to automatic corpus generation for chinese spelling check. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2517–2527.

Hongqiu Wu, Shaohua Zhang, Yuchen Zhang, and Hai Zhao. 2023. Rethinking masked language modeling for chinese spelling correction. *arXiv preprint arXiv:2305.17721*.

A Appendix: Dataset and Training Details

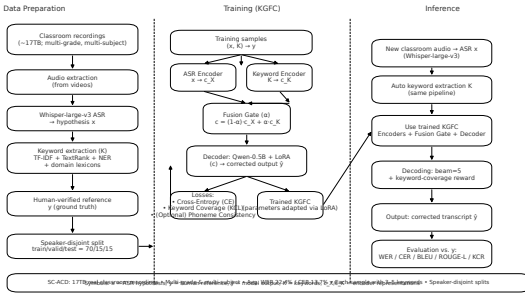

Figure 4: Supplementary flowchart from the original project materials. We keep it in the appendix to support the end-to-end data and training description without overcrowding the main paper.

B Appendix: Additional Discussion

- **Why keyword coverage matters.** In classroom speech, rare domain terms often carry most of the instructional meaning, so preserving them is more important than optimizing general fluency alone.

- **Where KGFC is likely to help most.** The method is especially useful when the raw transcript remains roughly understandable but misses or distorts a small set of critical terms.

- **Potential failure cases.** If keyword extraction is noisy, the fusion gate may overemphasize irrelevant terms. This suggests a useful future direction in confidence-aware keyword selection.