Selective Context Utilization in Speech-LLM ASR: Lexical Verification and Semantic Disambiguation with GRPO

Anonymous authors
Paper under double-blind review

Abstract

Context can improve entity recognition in speech-enabled large language models (Speech-LLMs), but misleading context can also bias transcription toward unsupported words. Context additionally provides semantic evidence for written distinctions that pronunciation alone cannot resolve. We study these complementary demands through contextual entity recognition and Mandarin third-person pronoun disambiguation. We develop a three-stage framework combining general ASR fine-tuning, robust context fine-tuning with mismatched contexts, and task-specific group relative policy optimization (GRPO). The lexical objective uses phonetic hard negatives and quality-gated distractor rejection, while the semantic objective rewards correct ordered pronoun sequences alongside full-transcription accuracy. We construct a diagnostic benchmark suite with 1,116 lexical utterances and 617 pronoun-disambiguation utterances, supplemented by paired clean/misleading contexts and controlled semantic-context ablations. Our model achieves 97.76% entity recall and 89.70% pronoun macro-F1, exceeding the strongest evaluated pronoun baseline by 7.12 percentage points. Under misleading lexical context, the full pipeline improves entity recall while matching the backbone's distractor false-positive rate. Semantic ablations show that the disambiguation gains depend on matched discourse, although general-ASR performance degrades unevenly across domains. We will release the benchmark upon acceptance. The code is available at `https://anonymous.4open.science/r/SCU-ASR-8E2D`.

1 Introduction

Automatic speech recognition (ASR) remains vulnerable on long-tail names, organizations, locations, products, and technical terms. Early end-to-end contextualization methods encoded supplied phrase lists to help recover rare lexical items from speech (Pundak et al., 2018; Bruguier et al., 2019; Chen et al., 2019; Jain et al., 2020).

Speech-enabled large language models (Speech-LLMs) make the conditioning interface more flexible by accepting context as text (Lakomkin et al., 2024; Gong et al., 2024). Natural-language prompts can carry hotwords, dialogue history, retrieved knowledge, or application metadata without a task-specific bias encoder. In conversational voice assistants, earlier turns can provide information about entities and events that helps transcribe the current utterance. Such information may be particularly useful when the current speech alone leaves the intended written form ambiguous.

Useful context must nevertheless be applied selectively. A prompt may be irrelevant, incomplete, or misleading, and excessive bias can introduce unsupported entities. Most contextual ASR work addresses this as a *lexical verification* problem: context proposes candidate expressions, while acoustic evidence determines whether they occur. Robust models must therefore recognize supported entities without copying phonetically plausible distractors.

We additionally consider *semantic transcription ambiguity*, where the pronunciation of the current utterance cannot uniquely determine the written output. English *he*, *she*, and *it* have distinct pronunciations, whereas Mandarin third-person pronouns share a pronunciation despite differing in reference. For example, a person's name may not indicate which written pronoun is intended, while earlier dialogue may establish the relevant referent attributes. In storytelling, pronoun choice for an

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Examples of semantic transcription ambiguity. Within each subtable, the written forms share a pronunciation but differ in reference (Mandarin) or number (French). Pronunciations are given in the International Phonetic Alphabet (IPA). French examples are illustrative; our experiments focus on Mandarin.

(a) Mandarin pronouns

Form	IPA	Meaning
他	[tʰaɻ]	he
她	[tʰaɻ]	she
它	[tʰaɻ]	it

(b) French utterances

Form	IPA	Meaning
Elle parle.	[ɛl paʁl]	She is speaking.
Elles parlent.	[ɛl paʁl]	They are speaking.

animal character may also depend on personification or an established writing convention. These cases motivate testing whether an ASR model can use available discourse evidence, supplemented where necessary by explicit semantic cues, to resolve ambiguities that pronunciation alone cannot settle. Table 1 shows the Mandarin forms alongside French singular–plural utterances that also share a pronunciation. We use Mandarin pronoun disambiguation as a controlled test of whether the model can use discourse context or auxiliary semantic information to recover such written distinctions. We refer to the three Mandarin forms by their English glosses *He*, *She*, and *It* throughout. This setting complements contextual entity recognition: lexical context is verified against acoustic evidence, whereas semantic context resolves distinctions not encoded in pronunciation. Existing approaches largely strengthen phonetic discrimination (Tang et al., 2024; He et al., 2025; Fu et al., 2026), which cannot by itself resolve exactly homophonic written forms. While semantic pronoun disambiguation has been studied in post-ASR correction (Zhang et al., 2024), we bring it into contextual Speech-LLM recognition itself.

Building on context-aware supervised fine-tuning (SFT) and reinforcement learning (Bai et al., 2024; Kong et al., 2025), we develop a three-stage framework. General ASR SFT first adapts the backbone to the target distribution. Robust Context-SFT then mixes relevant contexts with contexts sampled from unrelated examples, reducing the shortcut of copying contextual tokens. Finally, context-aware GRPO uses two task-specific objectives: a lexical reward for recognizing target entities and rejecting phonetic distractors, and a semantic reward for producing the correct ordered pronoun sequence while preserving the full transcription.

Our contributions are threefold:

- We study selective context utilization in Speech-LLM ASR through two complementary demands: recognizing context-supported entities while resisting phonetic distractors, and using semantic context to resolve written forms that pronunciation alone cannot distinguish. Mandarin third-person pronouns provide a controlled testbed for the latter.
- We develop a three-stage training framework combining general ASR adaptation, mismatched-context SFT, and task-specific GRPO. The lexical objective combines entity recognition with quality-gated distractor rejection, while the semantic objective targets ordered pronoun disambiguation alongside full-transcription accuracy.
- We construct a diagnostic benchmark suite covering contextual entity recognition and Mandarin pronoun disambiguation, with paired clean/misleading contexts and controlled semantic-context ablations. The suite measures context benefits, susceptibility to distractors, and dependence on semantic evidence. We will release the benchmark upon acceptance.

2 Related Work

2.1 Context Modeling in Speech-LLM-based ASR

Classical contextual ASR differs primarily in how candidate phrases enter the recognizer. CLAS adds a bias encoder and a second attention mechanism to a LAS decoder (Pundak et al., 2018). Phoebe replaces purely graphemic context with pronunciation-aware representations (Bruguier et al., 2019);

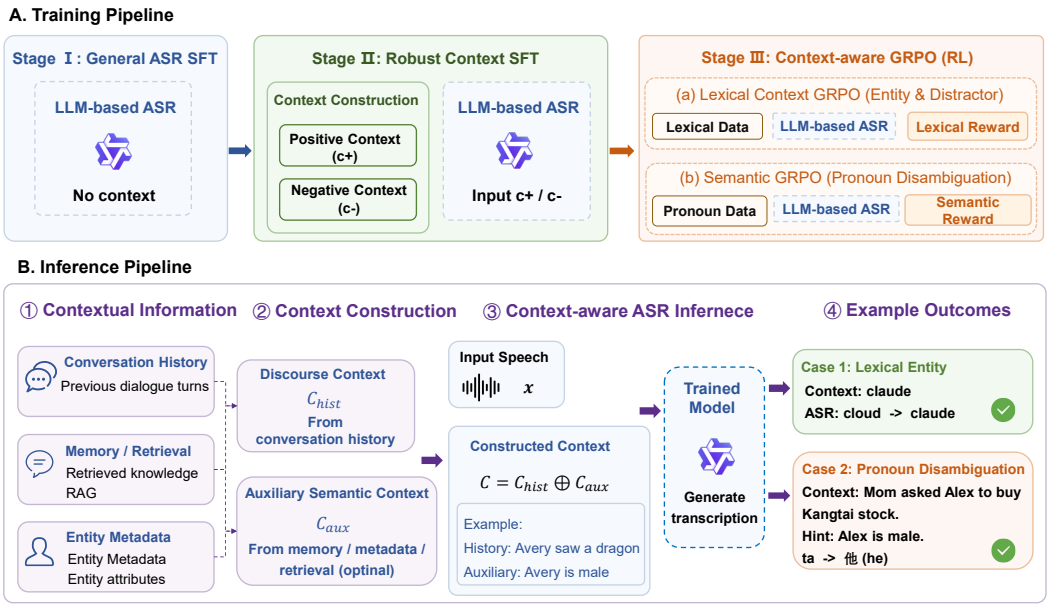


Figure 1: The three-stage framework. Context-aware GRPO combines lexical training with phonetic distractors and semantic pronoun disambiguation.

joint grapheme–phoneme encoders further improve discrimination and generalization to unseen entities (Chen et al., 2019). Contextual RNN-T transfers attention over context embeddings to a transducer and studies unstructured video metadata (Jain et al., 2020). These architectures inject the full candidate set, so their robustness depends on the quality and size of that set.

A second line of work reduces distractor interference before or during decoding. TCPGen restricts copying to paths in a prefix tree, which preserves valid bias-phrase continuations (Sun et al., 2022). CPPN predicts whether each contextual phrase is present and uses that probability to control biasing (Huang et al., 2023). In Speech-LLM systems, CTC-assisted filtering uses a coarse acoustic hypothesis to retain only plausible hotwords (Yang et al., 2024b). BR-ASR instead retrieves a compact bias set from a large inventory (Gong et al., 2025), and explicit ranking provides another way to select among competing bias words (Hou et al., 2025). Unlike these selection pipelines, our phonetic negatives remain in the prompt and train the recognizer itself to reject them.

The source and granularity of context have also expanded. Sun et al. (2024) give an LLM a bias list and few-shot examples during second-pass rescoring, while Xiao et al. (2025) augment contextual ASR with retrieved external evidence. Conversational systems retrieve and select relevant acoustic and textual history (Mu et al., 2026) or jointly prompt with dialogue history and hotwords (Zheng et al., 2026). ContextASR-Bench evaluates such capabilities across domains and entity types (Wang et al., 2025).

Speech-LLMs avoid a dedicated phrase encoder by accepting context as text. Mixed-modal language modeling directly conditions generation on speech and optional text (Lakomkin et al., 2024); prompt-based biasing adapts a speech-enhanced LLM to contextual entities (Gong et al., 2024). Seed-ASR uses context-aware SFT and reinforcement learning in successive stages (Bai et al., 2024). Most closely related to our lexical training, Kong et al. (2025) retrieve hotwords and apply GRPO with keyword and transcription rewards. We differ by training on deliberately confusable phonetic negatives, gating rejection credit by transcription quality, and introducing a separate semantic objective for acoustically identical forms.

2.2 Homophone and Context-Dependent Ambiguity Resolution

Chinese homophone methods intervene at different points in the recognition pipeline. Yang et al. (2024a) use confidence to detect homophone errors before activating contextual bias. Tang et al.

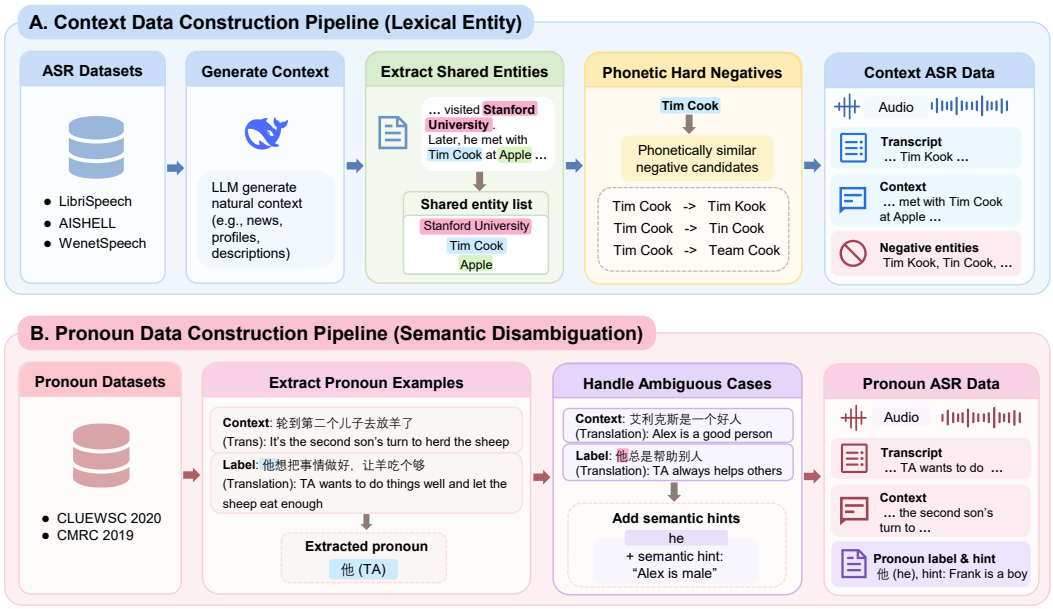


Figure 2: Construction of the two reinforcement-learning data sources. The lexical branch mines phonetic entity distractors and rollout-based hard cases, whereas the semantic branch curates context-dependent pronoun examples and supplements selected contexts with auxiliary semantic cues.

(2024) add a Pinyin regularizer to LLM-based ASR error correction, whereas Zhengjie & Cheng (2025) guide recognition jointly with Pinyin and text. PARCO augments contextual representations with phonemes and learns contrastive entity discrimination (He et al., 2025). PAC goes further by combining grapheme–phoneme context learning with a pronunciation-discriminative RL stage (Fu et al., 2026). Although their mechanisms differ, all exploit phonetic structure to separate candidates. Exact homophones remain underdetermined when the acoustic and phonetic representations are identical.

Mandarin third-person pronouns expose this limitation directly. Zhang et al. (2024) used language models, semantic features, and rules for downstream pronoun-homophone correction. We instead train a contextual Speech-LLM to resolve the written form during recognition, using preceding discourse or optional semantic metadata. This isolates semantic context use from conventional lexical biasing and phonetic discrimination.

3 Method

3.1 Overview and Backbone

Given speech $\mathbf{x}$, context $\mathbf{c}$, and transcription $\mathbf{y}$, the model estimates $P_\theta(\mathbf{y} \mid \mathbf{x}, \mathbf{c})$. We instantiate the framework with Qwen3-ASR: an audio encoder and adapter map the waveform to language-model representations, which are concatenated with the instruction and optional context before autoregressive decoding. The framework itself is applicable to Speech-LLM ASR backbones with the same conditioning interface.

3.2 Three-stage Training

Stage I: General ASR SFT. We fine-tune the pretrained backbone on context-independent speech–transcription pairs using the standard autoregressive cross-entropy objective. This adapts the model to the target distribution and provides an acoustically reliable initialization for contextual training.

Stage II: Robust Context-SFT. Training only with relevant context can teach the model to copy contextual tokens. We therefore replace the relevant context $\mathbf{c}_i^+$ with an unrelated context $\mathbf{c}_i^-$ with

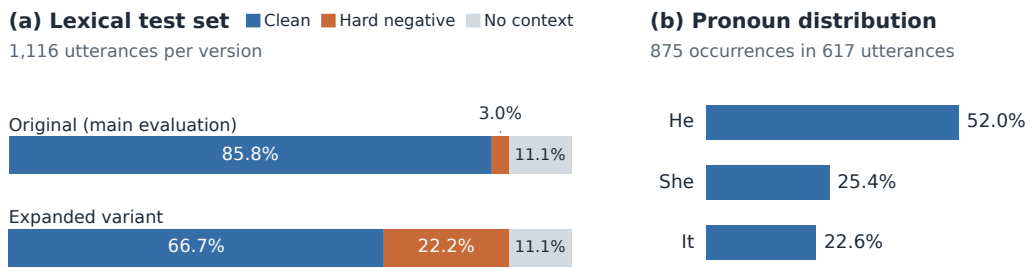


Figure 3: Test-set composition. (a) Original and expanded lexical sets (1,116 utterances each); expanded hard negatives account for 22.2% overall and 25% among contextual samples. (b) Pronoun distribution: 875 occurrences in 617 utterances.

probability p_{neg}, while leaving speech and transcription unchanged:

$$\tilde{\mathbf{c}}_i = \begin{cases} \mathbf{c}_i^+, & \text{with probability } 1 - p_{\text{neg}}, \\ \mathbf{c}_i^-, & \text{with probability } p_{\text{neg}}. \end{cases} \tag{1}$$

Stage III: Context-aware GRPO. Token-level SFT does not directly optimize the desired sequence-level behaviors. We therefore apply GRPO (Shao et al., 2024) to both lexical verification and semantic disambiguation. Each input produces a group of candidate transcriptions whose relative advantages are computed with the reward defined for its task. We use the standard clipped GRPO objective with KL regularization to the Context-SFT policy.

3.3 RL Data Construction

Figure 2 summarizes Stage III data construction. Lexical hard cases combine phonetic distractors (Section 3.4.1) with base-model rollout selection. We construct context-dependent pronoun disambiguation examples from CMRC 2019 (Cui et al., 2020) and CLUEWSC 2020(Xu et al., 2020) texts. We extract occurrences of the three target pronoun characters from the reference texts in their original order and manually retain examples whose intended written forms require contextual reasoning. For selected examples in which relevant referent attributes, such as gender, are underspecified in the original context, we supplement the context with auxiliary semantic cues guided by the reference pronoun forms.

3.4 Lexical Context GRPO

3.4.1 Phonetic Hard-Negative Construction

Random unrelated contexts teach coarse robustness but provide weak supervision for acoustically confusable entities. For each Chinese character z, we obtain tone-3 Pinyin $\pi(z) = (o(z), r(z), t(z))$, where o, r, and t are the initial, final, and tone. A fuzzy mapping ϕ merges commonly confused initials and finals, such as retroflex and non-retroflex consonants. We assign a character-level distance $d(z, z') \in \{0, 1, 2, 3\}$ according to exact syllable, tone, and fuzzy-syllable agreement; the complete rule is given in Appendix A.1. Lexically different exact homophones may have zero distance and are retained as the hardest negatives.

For equal-length entities $e = (z_1, \ldots, z_L)$ and $e' = (z'_1, \ldots, z'_L)$, we rank candidates using

$$\bar{D}(e, e') = \frac{1}{3L} \sum_{i=1}^{L} d(z_i, z'_i), \qquad \mathcal{N}_K(e) = \text{TopK}\{e' \neq e : \bar{D}(e, e') \leq \tau\}. \tag{2}$$

Candidates occurring in the reference are removed, and the remaining K nearest entities are inserted into the prompt as distractors.

Table 2: Character error rate (CER, %) on four Mandarin test sets. Lower is better. WS denotes WenetSpeech.

Model	AISHELL-1	AISHELL-2	WS-Net	WS-Meeting
Qwen2-Audio	1.52	3.08	8.40	7.64
Qwen3-ASR	1.48	2.71	4.97	5.88
FunASR-Nano	1.88	2.61	5.30	5.31
Step-Audio 2	0.63	**2.10**	4.75	4.67
Kimi-Audio	**0.60**	2.56	6.28	5.37
FireRedASR2-LLM	0.64	2.15	**4.44**	**4.32**
Ours (best)	3.07	2.99	5.55	7.53

3.4.2 Entity and Distractor Reward

Let $\mathcal{E} = \{e_m\}_{m=1}^{M}$ and $\mathcal{N} = \{n_k\}_{k=1}^{K}$ denote the target `must_emit` entities and `must_reject` distractors. We remove the model-specific language prefix, apply Unicode normalization, and discard punctuation and whitespace. Let $\nu(\cdot)$ denote this transformation and write $\bar{\mathbf{y}} = \nu(\hat{\mathbf{y}})$. We combine exact occurrence with the edit similarity of the closest hypothesis substring, where $[x]_+ = \max(x, 0)$:

$$R_{\text{ent}} = \frac{1}{M} \sum_{m=1}^{M} \left(0.75\, \mathbf{1}[\nu(e_m) \in \bar{\mathbf{y}}] + 0.25 \left[1 - \frac{\min_{s \preceq \bar{\mathbf{y}}} \text{ED}(\nu(e_m), s)}{|\nu(e_m)|} \right]_+ \right). \tag{3}$$

The 0.75:0.25 weighting keeps exact recognition dominant while retaining a dense signal for near misses.

Global quality is measured by $R_{\text{ASR}} = \max\{1 - \text{CER}(\bar{\mathbf{y}}, \nu(\mathbf{y})), 0\}$. For distractors, we first compute their absence rate and then gate it by transcription quality:

$$R_{\text{rej}} = \left(1 - \frac{1}{K} \sum_{k=1}^{K} \mathbf{1}[\nu(n_k) \in \bar{\mathbf{y}}] \right) \text{clip}\left(\frac{R_{\text{ASR}} - 0.5}{0.3}, 0, 1 \right). \tag{4}$$

Thus, an empty or poor transcript cannot receive a high score merely by omitting every distractor: rejection credit is zero up to ASR accuracy 0.5 and becomes fully active at 0.8. We combine the three complementary signals as

$$R_{\text{lex}} = 0.40 R_{\text{ASR}} + 0.35 R_{\text{ent}} + 0.25 R_{\text{rej}}. \tag{5}$$

If either annotation set is empty, its component is omitted and the remaining weights are renormalized.

3.5 Semantic Ambiguity GRPO

Each semantic example contains speech, discourse context, and an ordered pronoun annotation $\mathbf{p} = (p_1, \ldots, p_M)$, where $p_i \in \{他, 她, 它\}$. Optional metadata can additionally specify entity type, gender, identity, or a local writing convention. These sources are concatenated under a fixed prompt template.

From prediction $\hat{\mathbf{y}}$, we extract every occurrence of the three target pronoun characters in order to obtain $\hat{\mathbf{p}}$. We compare the two sequences by normalized edit similarity and add an exact-match bonus:

$$R_{\text{pron}} = 0.5\, \mathbf{1}[\hat{\mathbf{p}} = \mathbf{p}] + 0.5 \left[1 - \frac{\text{ED}(\hat{\mathbf{p}}, \mathbf{p})}{\max\{|\hat{\mathbf{p}}|, |\mathbf{p}|, 1\}} \right]_+. \tag{6}$$

The soft term provides graded feedback for a near-correct sequence, whereas the exact term is activated only when identity, count, and order all match. The semantic reward balances this targeted signal with the full-transcription score:

$$R_{\text{sem}} = 0.60 R_{\text{ASR}} + 0.40 R_{\text{pron}}. \tag{7}$$

Using full rather than pronoun-masked CER prevents the model from maximizing the task reward by emitting only pronouns. Samples without pronoun annotations fall back to R_{ASR}. Lexical and semantic samples use R_{lex} and R_{sem}, respectively.

Table 3: Contextual ASR results (%) on the lexical and semantic evaluation sets. BCER and UCER measure character error rates on context-biased and remaining unbiased characters, respectively; B-Recall denotes context-biased-word recall. He-R, She-R, and It-R denote class-wise pronoun recall. Best results are shown in **bold**.

Model	Lexical Context				Pronoun Disambiguation			
	CER ↓	BCER ↓	UCER ↓	B-Recall ↑	He-R ↑	She-R ↑	It-R ↑	Macro-F1 ↑
Qwen2-Audio	9.11	6.59	9.46	89.42	83.96	20.27	26.26	46.40
Qwen3-ASR	2.30	1.21	2.46	97.16	96.04	28.83	48.48	61.14
FunASR-Nano	2.73	3.39	2.64	92.15	**96.70**	5.41	18.18	36.69
Step-Audio 2	3.30	3.31	3.30	93.08	93.85	72.97	70.71	82.58
Kimi-Audio	4.52	3.04	4.74	94.10	89.89	63.51	52.02	73.42
FireRedASR2-LLM	3.19	3.91	3.08	91.63	88.79	18.02	59.60	56.40
Ours (Context-SFT)	2.36	1.18	2.53	97.31	94.29	24.32	36.36	54.03
Ours (GRPO-RL)	**2.14**	**1.18**	**2.28**	**97.76**	92.09	**81.08**	**94.44**	**89.70**

4 Experiments

4.1 Experimental Setup

Training. We use Qwen3-ASR-1.7B (Shi et al., 2026) throughout the three stages. Stage I uses 50k hours of general speech–transcription data.

Stage II uses 50k hours in total: 40k hours of general ASR data and 10k hours of contextual ASR data, with $p_{\text{neg}} = 0.5$ in Eq. 1. Both SFT stages use a learning rate of 5×10^{-6} and a per-GPU batch size of 32.

For Stage III, model rollouts retain 10,000 difficult samples per task from the data sources in Section 3.3. The resulting 20,000 samples are used in a single GRPO run implemented with verl (Sheng et al., 2024). Each training batch contains only lexical examples or only semantic examples, and a data field identifies the task for reward selection. The joint run uses a global prompt batch size of 128, a PPO mini-batch size of 32, and an actor learning rate of 5×10^{-6}. We sample six responses per prompt at temperature 1.0. Across all stages, we use AdamW, a 5% warmup ratio, bfloat16 precision, and NVIDIA A100 GPUs.

Evaluation sets. The main lexical benchmark contains 1,116 utterances: 958 with clean context, 34 with phonetic hard negatives, and 124 without context or target-entity annotations. Its 992 contextual utterances contain 1,258 target-entity annotations and 57 distractor annotations. An expanded variant retains the same audio and increases hard-negative utterances to 248, with 271 distractor annotations, leaving 744 clean-context and 124 context-free utterances. Hard negatives thus account for 25% of contextual utterances (22.2% overall). Robustness evaluation pairs clean and misleading contexts for these 248 utterances, yielding 496 evaluation records.

The semantic benchmark contains 617 utterances with discourse context, including 40 with auxiliary semantic hints. Its 875 annotated pronoun occurrences comprise 455 *He*, 222 *She*, and 198 *It*. Figure 3 summarizes the lexical evaluation conditions in panel (a) and the distribution of annotated pronoun occurrences in panel (b).

Normalization and metrics. We normalize references and predictions identically, removing task-irrelevant punctuation and standardizing whitespace while preserving annotated entities and target pronoun characters. We report character error rate (CER) on both sets. For lexical evaluation, we additionally report BCER and UCER, which measure errors on context-biased characters and the remaining unbiased characters, respectively, as well as B-Recall for context-biased words. On the semantic set, we report pronoun accuracy, class-wise recall (He-R, She-R, and It-R), and macro-F1. Together, these metrics assess overall transcription quality, entity recognition, and semantic disambiguation.

Table 4: Context robustness and training ablations. Misleading contexts replace target entities with phonetically similar distractors. C-SFT denotes contextual supervised fine-tuning; Full model uses all three training stages. B-Recall and D-FPR denote context-biased-word recall and distractor false-positive rate, respectively. Values are percentages, except B-Recall Drop: clean minus misleading B-Recall in percentage points. Best results are shown in **bold**.

Model	Clean Context		Misleading Context			Robustness Gap
	CER ↓	B-Recall ↑	CER ↓	B-Recall ↑	D-FPR ↓	B-Recall Drop ↓
Qwen3-ASR	3.28	91.75	3.68	81.27	8.86	10.48
C-SFT ($p_{\text{neg}} = 0$)	3.40	89.52	4.05	67.84	21.83	21.68
C-SFT ($p_{\text{neg}} = 0.5$)	3.01	92.49	3.98	69.21	21.03	23.28
Direct GRPO	**2.88**	92.38	3.66	80.95	**8.48**	11.43
Full model	2.89	**93.02**	**3.20**	**83.49**	8.86	**9.52**

Baselines. We compare against six publicly available ASR and audio-language models: Qwen2-Audio (Chu et al., 2024), the original Qwen3-ASR-1.7B backbone (Shi et al., 2026), FunASR-Nano (Gao et al., 2023), Step-Audio 2 (Wu et al., 2025), Kimi-Audio (Ding et al., 2025), and FireRedASR2-LLM (Xu et al., 2026). On the contextual benchmarks, Qwen2-Audio, Qwen3-ASR, Step-Audio 2, and Kimi-Audio receive the same textual context through their instruction interfaces. FunASR-Nano and FireRedASR2-LLM do not provide an instruction-following or context-injection interface, so we evaluate them using their recommended context-free configurations. All models receive the same audio inputs.

5 Results

5.1 General ASR Performance

Compared with Qwen3-ASR, our model shows limited degradation on AISHELL-2 and WS-Net but larger degradation on AISHELL-1 and WS-Meeting (Table 2). We hypothesize that these losses reflect data-distribution imbalance and forgetting during contextual optimization.

5.2 Contextual ASR Performance

On the lexical set, our GRPO-trained model achieves the best CER, UCER, and B-Recall while tying Context-SFT for the best BCER (Table 3). These joint gains indicate that GRPO strengthens the recognition of context-supported entities without degrading the surrounding transcription.

The effect of GRPO is more pronounced on semantic ambiguity. Context-SFT alone retains a strong majority-class bias, whereas GRPO raises macro-F1 from 54.03% to 89.70%, surpassing the strongest baseline, Step-Audio 2, by 7.12 percentage points. This gain is primarily driven by substantially better recognition of the *She* and *It* classes, with a modest decrease in recall for the *He* class.

5.3 Context Robustness Analysis

Without mismatched-context augmentation, C-SFT degrades under both clean and misleading contexts (Table 4). Setting $p_{\text{neg}} = 0.5$ substantially improves clean-context recognition and yields modest gains under misleading context. Nevertheless, its high D-FPR and large B-Recall drop show that augmentation alone is insufficient to prevent over-reliance on contextual entities.

The comparison with Direct GRPO further demonstrates the complementarity of the two training stages. Direct GRPO achieves the lowest clean-context CER and D-FPR, but its correct-entity recall under misleading context remains close to the backbone. The full model raises misleading-context B-Recall from 80.95% to 83.49%, reduces CER from 3.66% to 3.20%, and obtains the smallest B-Recall drop (9.52 percentage points), while maintaining the backbone-level D-FPR of 8.86%. These results

Table 5: Semantic context ablation for Ours (GRPO-RL) (%). Mismatched contexts are obtained by shuffling the original contexts across utterances and are treated as a context-sensitivity diagnostic. The hint subset contains 40 utterances with 52 annotated pronoun occurrences. Best results within each subset are shown in **bold**.

Evaluation Subset	Context Condition	CER ↓	Macro-F1 ↑
Full TA-Bench (617)	No context	4.82	54.17
	Mismatched context	5.60	47.35
	Original context	**2.89**	**89.70**
Hint subset (40)	Original context	2.80	63.51
	Original context + semantic hint	**1.23**	**93.93**

indicate that C-SFT develops effective context utilization, whereas GRPO makes this utilization more selective, producing the best overall recall–robustness trade-off.

5.4 Semantic Context Ablation

Removing discourse context reduces the GRPO-trained model's macro-F1 from 89.70% to 54.17%, while mismatched context lowers it to 47.35% (Table 5). CER follows the same ordering, indicating that semantic gains depend on matched context. Because shuffled context may conflict with the original reference, we treat this condition as a context-sensitivity diagnostic rather than a conventional accuracy benchmark.

The 40-utterance hint subset contains cases for which the original discourse provides insufficient semantic evidence. Adding an explicit semantic hint raises macro-F1 from 63.51% to 93.93% while also reducing CER.

6 Conclusions

We presented a three-stage framework for selective context utilization in Speech-LLM ASR, combining general ASR adaptation, robust Context-SFT with mismatched contexts, and task-specific GRPO for lexical verification and Mandarin pronoun disambiguation. On our diagnostic benchmark suite, the model achieves 97.76% B-Recall and 89.70% pronoun macro-F1, exceeding the strongest evaluated pronoun baseline by 7.12 percentage points. Under misleading lexical context, the full pipeline yields the smallest entity-recall drop among evaluated variants while matching the backbone's distractor false-positive rate. Semantic ablations show that the disambiguation gains depend on matched discourse, with auxiliary hints providing further benefits when discourse alone is insufficient.

Together, these findings highlight two complementary requirements for contextual ASR: using acoustic evidence to verify lexical candidates and using semantic context to resolve written distinctions that pronunciation alone cannot determine. Contextual improvements nevertheless come with general-ASR degradation that varies across domains. Future work will investigate more diverse general-ASR replay, improved data mixing, and regularization to mitigate this trade-off, while extending semantic disambiguation beyond Mandarin pronouns.

AI use statement

We used generative AI tools to assist with data construction, method implementation, software development, manuscript preparation, and literature discovery. Specifically, generative AI was used to generate contextual information for a subset of the training data based on the corresponding transcription text, and to assist with the implementation of parts of the proposed methods and related code. We also used generative AI for translation, language polishing, improving manuscript readability, identifying relevant literature, and assisting in the creation of the overview figure.

All AI-assisted outputs included in the final work were reviewed by the authors. AI-assisted code was inspected and tested before use, and AI-assisted literature discovery was verified against the original

sources. The authors take full responsibility for the final content of the paper and all associated artifacts.

Reproducibility statement

To support reproducibility, Section 4.1 describes the model backbone, training setup, and evaluation protocol. Section 3.3 describes the construction of the lexical and semantic reinforcement-learning data, and Section 3.4.1 together with Appendix A.1 explains phonetic hard-negative construction and the character-level distance rule. The training objectives and task-specific reward functions are described in Section 3. The anonymous repository linked in the abstract provides inference instructions and four demonstration examples. The current release does not include the full training code or the complete training and evaluation datasets. We plan to release the training code, the test sets, and a subset of the training data upon acceptance.

References

Ye Bai, Jingping Chen, Jitong Chen, Wei Chen, Zhuo Chen, Chuang Ding, Linhao Dong, Qianqian Dong, Yujiao Du, Kepan Gao, et al. Seed-asr: Understanding diverse speech and contexts with llm-based speech recognition. *arXiv preprint arXiv:2407.04675*, 2024.

Antoine Bruguier, Rohit Prabhavalkar, Golan Pundak, and Tara N Sainath. Phoebe: Pronunciation-aware contextualization for end-to-end speech recognition. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6171–6175. IEEE, 2019.

Zhehuai Chen, Mahaveer Jain, Yongqiang Wang, Michael L Seltzer, and Christian Fuegen. Joint grapheme and phoneme embeddings for contextual end-to-end asr. In *Proc. Interspeech 2019*, pp. 3490–3494, 2019.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

Yiming Cui, Ting Liu, Ziqing Yang, Zhipeng Chen, Wentao Ma, Wanxiang Che, Shijin Wang, and Guoping Hu. A sentence cloze dataset for chinese machine reading comprehension. In *Proceedings of the 28th international conference on computational linguistics*, pp. 6717–6723, 2020.

Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, et al. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*, 2025.

Li Fu, Yu Xin, Sunlu Zeng, Lu Fan, Youzheng Wu, and Xiaodong He. Pac: Pronunciation-aware contextualized large language model-based automatic speech recognition. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 16377–16381. IEEE, 2026.

Zhifu Gao et al. Funasr: A fundamental end-to-end speech recognition toolkit. In *INTERSPEECH*, 2023.

Xun Gong, Anqi Lv, Zhiming Wang, and Yanmin Qian. Contextual biasing speech recognition in speech-enhanced large language model. In *INTERSPEECH*, 2024.

Xun Gong, Anqi Lv, Zhiming Wang, Huijia Zhu, and Yanmin Qian. Br-asr: Efficient and scalable bias retrieval framework for contextual biasing asr in speech llm. *arXiv preprint arXiv:2505.19179*, 2025.

Jiajun He, Naoki Sawada, Koichi Miyazaki, and Tomoki Toda. Parco: Phoneme-augmented robust contextual asr via contrastive entity disambiguation. In *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–7. IEEE, 2025.

Haoxiang Hou, Xun Gong, Wangyou Zhang, Wei Wang, and Yanmin Qian. Ranking and selection of bias words for contextual bias speech recognition. In *Proc. Interspeech 2025*, pp. 5183–5187, 2025.

Kaixun Huang, Ao Zhang, Zhanheng Yang, Pengcheng Guo, Bingshen Mu, Tianyi Xu, and Lei Xie. Contextualized end-to-end speech recognition with contextual phrase prediction network. *arXiv preprint arXiv:2305.12493*, 2023.

Mahaveer Jain, Gil Keren, Jay Mahadeokar, Geoffrey Zweig, Florian Metze, and Yatharth Saraf. Contextual rnn-t for open domain asr. *arXiv preprint arXiv:2006.03411*, 2020.

YuXiang Kong, JunFeng Hou, Jian Tang, Bingqing Zhu, Jicheng Zhang, and Shaofei Xue. Contextual biasing for llm-based asr with hotword retrieval and reinforcement learning. *arXiv preprint arXiv:2512.21828*, 2025.

Egor Lakomkin, Chunyang Wu, Yassir Fathullah, Ozlem Kalinli, Michael L Seltzer, and Christian Fuegen. End-to-end speech recognition contextualization with large language models. In *Icassp 2024-2024 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 12406–12410. IEEE, 2024.

Bingshen Mu, Hexin Liu, Hongfei Xue, Kun Wei, and Lei Xie. Hearing more with less: Multimodal retrieval-and-selection augmented conversational llm-based asr. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 32519–32527, 2026.

Golan Pundak, Tara N Sainath, Rohit Prabhavalkar, Anjuli Kannan, and Ding Zhao. Deep context: end-to-end contextual speech recognition. In *2018 IEEE spoken language technology workshop (SLT)*, pp. 418–425. IEEE, 2018.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, et al. Qwen3-asr technical report. *arXiv preprint arXiv:2601.21337*, 2026.

Chuanneng Sun, Zeeshan Ahmed, Yingyi Ma, Zhe Liu, Lucas Kabela, Yutong Pang, and Ozlem Kalinli. Contextual biasing of named-entities with large language models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 10151–10155. IEEE, 2024.

Guangzhi Sun, Chao Zhang, and Philip C Woodland. Minimising biasing word errors for contextual asr with the tree-constrained pointer generator. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:345–354, 2022.

Zhiyuan Tang, Dong Wang, Shen Huang, and Shidong Shang. Pinyin regularization in error correction for chinese speech recognition with large language models. *arXiv preprint arXiv:2407.01909*, 2024.

He Wang, Linhan Ma, Dake Guo, Xiong Wang, Lei Xie, Jin Xu, and Junyang Lin. Contextasr-bench: A massive contextual speech recognition benchmark. *arXiv preprint arXiv:2507.05727*, 2025.

Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, et al. Step-audio 2 technical report. *arXiv preprint arXiv:2507.16632*, 2025.

Cihan Xiao, Zejiang Hou, Daniel Garcia-Romero, and Kyu J Han. Contextual asr with retrieval augmented large language model. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Kaituo Xu, Yan Jia, Kai Huang, Junjie Chen, Wenpeng Li, Kun Liu, Feng-Long Xie, Xu Tang, and Yao Hu. Fireredasr2s: A state-of-the-art industrial-grade all-in-one automatic speech recognition system. *arXiv preprint arXiv:2603.10420*, 2026.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, et al. Clue: A chinese language understanding evaluation benchmark. In *Proceedings of the 28th international conference on computational linguistics*, pp. 4762–4772, 2020.

Chengxu Yang, Lin Zheng, Sanli Tian, Gaofeng Cheng, Sujie Xiao, and Ta Li. Contextual biasing with confidence-based homophone detector for mandarin end-to-end speech recognition. In *Proc. Interspeech 2024*, pp. 747–751, 2024a.

Guanrou Yang, Ziyang Ma, Zhifu Gao, Shiliang Zhang, and Xie Chen. Ctc-assisted llm-based contextual asr. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pp. 126–131. IEEE, 2024b.

Zhaobo Zhang, Rui Gan, Pingpeng Yuan, and Hai Jin. Correcting pronoun homophones with subtle semantics in chinese speech recognition. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 4047–4058, 2024.

Jie Zheng, Gaofeng Cheng, Xuyang Wang, Qingwei Zhao, and Yonghong Yan. Multilevel contextual prompting for conversational asr: unifying conversation history and hotwords with speech llm. *Speech Communication*, pp. 103458, 2026.

Jie Zhengjie and Gaofeng Cheng. Pinyin-guided chinese speech recognition with large language model. In *Proc. Interspeech 2025*, pp. 564–568, 2025.

A Appendix

A.1 Full Character-level Phonetic Distance

Let $b(z) = (o(z), r(z))$ be the base syllable and $\tilde{b}(z) = \phi(o(z), r(z))$ its fuzzy-normalized form. The character-level distance used for hard-negative mining is

$$d(z, z') = \begin{cases} 0, & b(z) = b(z') \wedge t(z) = t(z'), \\ 1, & b(z) = b(z') \wedge t(z) \neq t(z'), \\ 1, & b(z) \neq b(z') \wedge \tilde{b}(z) = \tilde{b}(z') \wedge t(z) = t(z'), \\ 2, & b(z) \neq b(z') \wedge \tilde{b}(z) = \tilde{b}(z') \wedge t(z) \neq t(z'), \\ 3, & \text{otherwise.} \end{cases} \tag{8}$$