EHRA: An Effective Hotword Retrieval Approach for Contextual ASR

Chiun-Yu Hua[1], Tien-Hong Lo[1], Yung-Chang Hsu[2], and Berlin Chen[1]
[1]National Taiwan Normal University, Taiwan
[2]EZ-AI, Taiwan
{614k0009c, teinhonglo, berlin}@ntnu.edu.tw
mic@ez-ai.com.tw

Abstract

While contextual ASR enhances rare entity recognition, irrelevant prompts frequently induce hallucinations. To mitigate this challenge, we propose an effective hotword retrieval approach for Mandarin ASR that integrates semantic embedding similarity, Smith-Waterman pinyin alignment, and multi-threshold filtering. Evaluations on the AISHELL-1 dataset, utilizing a frozen Qwen3-ASR-0.6B model, demonstrate that our adaptive filtering mechanism reduces the Biased Character Error Rate (B-CER) to 16.96% while maintaining the Overall CER (O-CER) at 1.88%, effectively improving upon the fixed Top-5 baseline. Notably, the proposed mechanism yields consistent ASR performance improvements across various text encoders.

Keywords: Contextual ASR, hotword retrieval, Qwen3-ASR, AISHELL-1

1 Introduction

Automatic speech recognition (ASR) systems have reached strong overall accuracy on benchmark corpora, yet they remain brittle on rare named entities, product names, and domain-specific terms—collectively called *hotwords*. Such terms are frequent in practical deployments but under-represented in training data, so errors concentrate on entities that the overall character error rate (CER) understates. A growing line of work therefore injects utterance-level context into ASR through *contextual biasing*: shallow fusion with bias language models (Williams et al., 2018), end-to-end attention over bias phrases as in Contextual Listen, Attend and Spell (CLAS) (Pundak et al., 2018), and prompt- or retrieval-based biasing for speech large language models (SpeechLLMs) (Shi et al., 2026). These approaches establish that *correct* hotwords can markedly improve entity recognition.

The converse is equally important: *ineffective retrieval*—wrong, excessive, or indiscriminate prompts—competes for decoder attention, induces hallucinations, and can erase the benefit of a strong ASR backend. (Hou et al., 2025) Empirically, prompting the entire 400-word lexicon or always emitting Fixed Top-K candidates yields little or no ASR gain over a no-hotword baseline, despite high retrieval recall. In short, ASR improvement can be correlated with retrieval effectiveness, not SpeechLLM alone. We therefore present an effective hotword retrieval approach for contextual ASR, based on the premise that ASR improvements cannot rely solely on retrieval recall; retrieval precision plays an equally critical role, and mere inclusion proves ineffective.

Concretely, we contribute a modular pipeline: a text encoder embeds baseline transcripts and lexicon entries; scores are fused with Smith–Waterman alignment on toneless pinyin; subsequently, the collaborative filtering mechanism selects candidates to emit sparse, high-confidence prompts (or outputs an empty set when criteria are unmet); finally, a frozen SpeechLLM consumes this selected list. Crucially, our pipeline is strictly encoder-agnostic. The text encoder serves merely as a pluggable representation module, allowing our fusion and gating strategies to generalize across any compatible text embedder.

Because gains hold for multiple text-encoder instantiations under identical gates, recognition tracks the pipeline rather than a single encoder.

Our contributions are:

Proceedings of the 38th Conference on Computational Linguistics and Speech Processing (ROCLING 2026), pages 207–213
November 6-7, 2026 ©2026 Association for Computational Linguistics

- We empirically demonstrate that while high recall is essential, high precision to mitigate hallucinations is equally indispensable for contextual ASR gains, showing that naive, high-recall policies (full lexicon, Fixed Top-K) fail to improve recognition.

- We propose an effective hotword retrieval approach (EHRA) that combines text-encoder—transcript similarity, pinyin fusion, and multi-metric gating. This pipeline transforms a merely high-recall ranker into a high-quality prompt source for a frozen SpeechLLM, simultaneously maintaining high recall while ensuring high precision.

- We instantiate the same pipeline with multiple peer text encoders and show that whenever gated retrieval remains effective, ASR gains persist in a narrow band, confirming the effectiveness of our encoder-agnostic pipeline.

2 Related Work

Retrieval-Augmented Contextual Biasing for SpeechLLMs With the rise of speech large language models (SpeechLLMs), contextual ASR increasingly relies on retrieval-augmented generation (RAG) and prompt-based biasing (Pundak et al., 2018). Instead of decoding-time shallow fusion, recent approaches use a first-pass ASR hypothesis to query a hotword catalog, injecting selected candidates as text prompts into the LLM.

For instance, Xiao et al. (Xiao et al., 2025) explore retrieval-augmented LLM adaptation, while Asano et al. (Asano et al., 2025) combine lexical semantics with pronunciation features. However, by prioritizing broad recall and heavy reliance on semantic spaces, these methods remain vulnerable to homophonic noise and hallucinations.

Retrieval for SpeechLLM biasing. Recent work also explores multi-metric or joint retrieval strategies to filter contextual dictionaries. For instance, Zhou and Li (Zhou and Li, 2026) propose a dynamic dictionary filtering method via joint semantic-pinyin-glyph retrieval for Chinese contextual ASR, supporting the necessity of multi-signal gating and precision control.

As ASR backends move toward promptable SpeechLLMs, retrieving a compact bias subset from a large catalog becomes necessary for scalability. BR-ASR (Gong et al., 2025) retrieves acoustic/textual bias items for SpeechLLM prompting and highlights homophone confusion as a dominant error type—consistent with our use of pinyin alignment as a complementary signal to embedding similarity.

Qwen3-ASR (Shi et al., 2026) supports context tokens in the system prompt as background knowledge for customized recognition without weight updates. Our work is complementary: we keep the SpeechLLM frozen and isolate the effect of *prompt precision* under a realistic closed hotword lexicon.

Text Embeddings for Contextual Retrieval In modern cascading and retrieval-augmented contextual ASR systems, mapping hotword lexicons and ASR transcripts into a shared continuous vector space has become a fundamental approach. High-performance general-purpose text embedding models—such as the BGE series introduced by Xiao (Xiao et al., 2024) through the C-Pack project—provide robust dense and sparse representations that enable efficient semantic similarity search. By leveraging these advanced text encoders, systems can retrieve relevant contextual candidates from large-scale catalogs without exhaustive literal matching.

Beyond monolingual encoders, universal spaces like SONAR (Duquenne et al., 2023) map text and speech into a shared semantic space, enabling zero-shot cross-modal retrieval for hotword matching. Although bridging the modality gap, these universal embeddings struggle with fine-grained phonetic variations—reinforcing the necessity for complementary phonetic alignment and precision gating.

3 Methodology

We formulate hotword-aware recognition as an effective retrieve-then-bias framework for contextual ASR. (Fig. 1) Let $\mathcal{V} = \{w_1, \ldots, w_M\}$ be a closed hotword lexicon ($M = 400$) and $\mathcal{U} = \{u_1, \ldots, u_N\}$ the AISHELL-1 test utter-

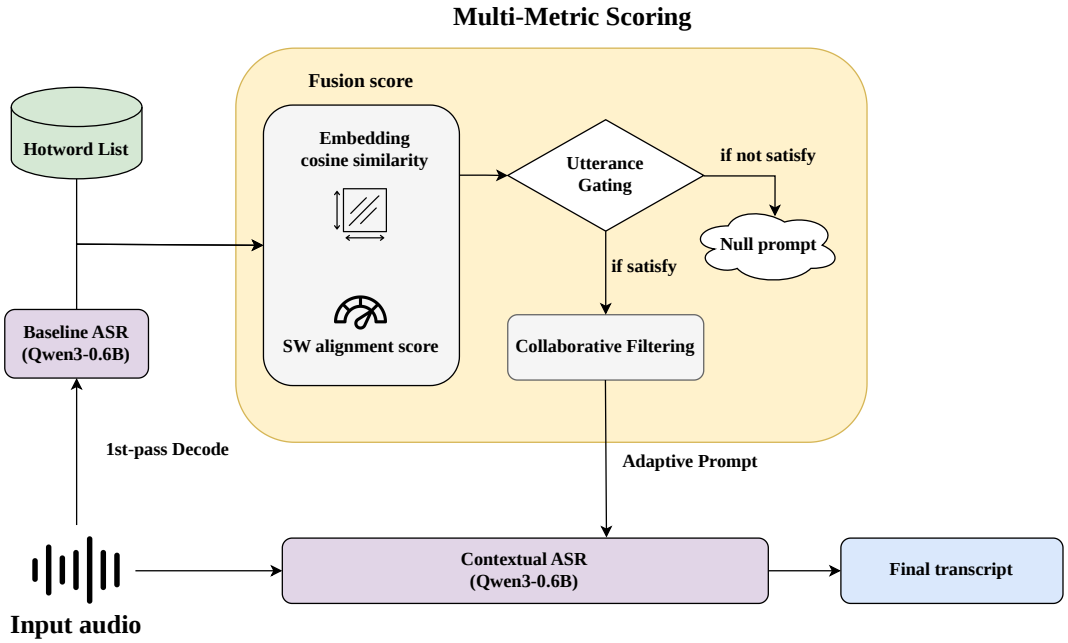

Figure 1: The effective hotword retrieval pipeline for contextual ASR, integrating multi-metric semantic-phonetic scoring and adaptive gating to mitigate hallucinations.

ances ($N = 7{,}176$) (Bu et al., 2017). For each utterance u, the retriever returns a (possibly empty) set $\hat{\mathcal{H}}(u) \subseteq \mathcal{V}$. A frozen ASR model then produces a hypothesis $y = \mathrm{ASR}(u, \hat{\mathcal{H}}(u))$. The design goal is not merely to maximize retrieval recall, but to maximize the *downstream* benefit of $\hat{\mathcal{H}}(u)$—which, we argue, requires high precision and the option to emit $\hat{\mathcal{H}}(u) = \emptyset$ when evidence is weak.

3.1 Feature Extraction

Representations. We encode each hotword string using a pre-trained text encoder (SONAR text encoder) to obtain the hotword embeddings, E_{hw}. For each test utterance, we extract an additional representation: a transcript embedding, E_{tr}, by processing the baseline ASR hypothesis without hotword context through the text encoder. All extracted embeddings are L2-normalized, and the respective cosine similarity matrices are computed as $S_{\mathrm{tr}} = E_{\mathrm{tr}} E_{\mathrm{hw}}^{\top}$.

String Similarities. In parallel to embedding extraction, we compute the lexical and phonetic similarities between the baseline hypothesis and each hotword. The metrics encompass Levenshtein-based local matching, character-bigram Jaccard similarity, and Smith-Waterman (SW) local alignment applied to toneless phonetic representations (pinyin). The SW alignment scores are normalized to a $[0, 1]$ range based on the maximum attainable match score. This phonetic-based SW alignment is particularly crucial for Mandarin hotwords, as it captures the phonetic neighborhoods frequently preserved in ASR character errors, which pure semantic embeddings often fail to resolve.

3.2 Adaptive Prompt Construction

To overcome the limitations of relying solely on pure transcript embedding similarity (S_{tr}) and naive static Top-K retrieval, we integrate the scoring and selection phases into a cohesive pipeline. Our method first fuses phonetic features to enhance ranking robustness, followed by a multi-metric margin-based gating mechanism to adaptively construct a precise candidate list.

Phonetic Fusion. We bridge the gap between semantic relevance and acoustic confusion by fusing the pure transcript embedding similarity (S_{tr}) with the phonetic SW alignment score (S_{sw}). The fusion mechanism ap-

plies row-wise min-max normalization followed by equal averaging:

$$\text{fuse}(A, B) = \frac{1}{2}\text{minmax}_{\text{row}}(A) + \frac{1}{2}\text{minmax}_{\text{row}}(B) \quad (1)$$

Empirical evaluation demonstrates that this fused score (S_{fus}) provides a significantly more robust ranking in noisy ASR conditions than pure embeddings alone. Consequently, we adopt it as the primary metric for our precision-oriented selection stage.

Utterance Gating. We define the maximum phonetic match for an utterance u as $g(u) = \max_j S_{\text{phon}}(u, w_j)$. If $g(u) < \tau_{\text{utt}}$ (a pre-defined utterance-level threshold), we set the candidate set $\hat{\mathcal{H}}(u) = \emptyset$ and bypass the contextual biasing process entirely. This initial gate proactively suppresses false-positive prompts for utterances lacking target hotwords.

Collaborative Margin-based Filtering. For utterances that pass the initial gate, a candidate hotword w_j is retained only if it simultaneously satisfies absolute thresholding and relative margin constraints across our complementary metrics:

$$\begin{aligned} S_*(u, w_j) &\geq \tau_*, \quad \forall * \in \{\text{fus}, \text{phon}, \text{lex}\} \\ S_{\text{fus}}(u, w_j) &\geq \max_k S_{\text{fus}}(u, w_k) - \delta, \end{aligned} \quad (2)$$

where $*$ denotes the scoring metric and τ_* is its corresponding absolute threshold. The within-utterance relative margin δ discards candidates that score significantly lower than the top-ranked hotword. Surviving candidates are sorted by S_{fus} and truncated to a predefined maximum length (e.g., Top-K) for ASR prompting.

Operating Point Optimization. The threshold hyperparameters (τ_{utt}, τ_{fus}, τ_{phon}, τ_{lex}) and the margin δ are determined via grid search on a validation set to optimally balance precision and recall. At this calibrated operating point, our proposed pipeline—integrating phonetic fusion and rigorous multi-metric pruning—achieves a strictly bounded false-positive rate. This holistic approach represents a significant and robust improvement over baseline strategies that rely on pure embedding similarities with static Top-K retrieval.

3.3 Contextual ASR

We transcribe all test utterances using a pretrained contextual ASR model, which is kept frozen during inference.

Prompting Mode. If the candidate set $\hat{\mathcal{H}}(u) = \emptyset$, we perform standard decoding without any contextual bias. Otherwise, the model is directly conditioned on a concatenated textual representation of the retrieved candidate hotwords.

Design Rationale. The proposed pipeline intentionally decouples *ranking* from *admission*. The transcript-based similarity effectively captures semantic and orthographic proximity within the embedding space, whereas the phonetic alignment recovers confusable hotwords that ASR character substitutions frequently leave partially intact. The strict admission thresholds convert a high-recall ranker into a sparse, high-precision context set. This design reflects a conservative biasing philosophy: when the system is uncertain, withholding contextual bias is strictly preferable to forcing noisy Top-K candidates that inevitably induce ASR hallucination.

4 Experimental Setup

4.1 Dataset and Lexicon

We evaluate on the AISHELL-1 test set (N = 7,176 utterances) using its official 400-term hotword lexicon and associated keyword audio clips for text–audio prompting. Ground-truth labels denote an utterance as *positive* if it contains at least one lexicon term (808 positives / 6,368 negatives). All ASR metrics evaluate entity matching against the entire 400-word lexicon after standard text normalization (punctuation stripping and space removal).

The hotword lexicon comprises 400 distinct Mandarin entries with a mean length of 3.49 characters (range 2–8; 109 bigrams, 163 trigrams, 45 four-, 27 five-, 29 six-, 15 seven- and 12 eight-character entries). Semantically the lexicon is dominated by personal names, followed by organisations and media outlets, place names and a small number of multi-word

Method	O-CER↓	B-CER↓	Rec.↑	Prec.↑	F1↑
Baseline	2.09	39.16	63.48	99.50	0.775
GT	1.74	7.04	92.89	99.89	0.963
Full lexicon	2.13	39.69	63.06	99.50	0.772
Fixed Top-5	2.28	25.23	78.24	94.85	0.857
EHRA (Ours)	**1.88**	**16.96**	**83.01**	98.87	**0.902**

Table 1: Contextual ASR on AISHELL-1 test with a fixed Qwen3-ASR-0.6B backend.

	Hotword Retrieval				Contextual ASR				
Text Encoder	Prec.↑	Rec.↑	F1↑	Neg. FP↓	O-CER↓	B-CER↓	Rec.↑	Prec.↑	F1↑
SONAR	0.802	0.852	0.827	0.021	1.88	16.96	83.01	98.74	0.902
BGE-M3	**0.809**	0.852	**0.830**	0.021	1.88	17.24	83.01	98.74	0.902
Qwen3-Embedding-0.6B	0.799	**0.855**	0.826	0.022	1.88	**16.82**	**83.23**	98.74	**0.903**

Fixed Top-5 is ineffective for all encoders (Neg. FP = 1.0).

Table 2: Performance of various text encoders under the proposed adaptive pipeline.

phrases. The entries draw on 633 distinct characters, of which 152 occur in three or more entries; this heavy character-level overlap is what makes the retrieval task challenging, since embedding similarity alone cannot separate entries that differ by a single homophonous character.

4.2 Pipeline Instantiations

All configurations share the same interpolation weight ($\alpha = 0.5$), string similarity metrics, collaborative filtering thresholds, and frozen Qwen3-ASR-0.6B backend.We adopt Qwen3-ASR-0.6B as the shared backend because the retrieval stage consumes baseline transcripts rather than the final hypothesis, so recognition quality affects retrieval only through the phonetic evidence it exposes; a compact backbone is sufficient to produce that evidence while keeping the full 7,176-utterance pipeline cheap enough to sweep thresholds exhaustively. To isolate the impact of the text encoder f_{text}, we benchmark three off-the-shelf models as equal peers: the SONAR text encoder, BGE-M3, and Qwen3-Embedding-0.6B. Embeddings are cached upfront, and a shared speech encoder generates E_{sp} across all peers when applicable. Unbiased baseline hypotheses $h(u)$ and downstream contextual decoding both utilize the same ASR checkpoint. Retrieved prompts are strictly capped at five hotwords per utterance.

4.3 Compared Conditions

We evaluate our approach against a Ground Truth (GT) upper bound (gold hotwords) and several weak retrieval baselines, including a no-hotword baseline, full 400-word lexicon injection, and static Top-5 retrieval. For downstream ASR, the adaptively retrieved candidates are directly provided as textual prompts. Utterances bypassed by the utterance gate default to plain ASR decoding without prompts.

4.4 Metrics

Retrieval. Retrieval effectiveness is measured via instance-level precision, recall, and F1. We additionally report utterance-level recall and the negative false-positive rate (the fraction of true-negative utterances incorrectly assigned any prompt).

ASR. We report overall CER (O-CER) and biased CER (B-CER) on hotword character mass, with entity-level Precision / Recall / F1. Definitions follow standard contextual ASR practice: for each hotword, TP/FN/FP are counted by string containment in reference vs. hypothesis; B-CER accumulates hotword lengths into reference/error character pools accordingly. We break results down on (i) the full test set, (ii) prompted utterances, and (iii) unprompted utterances.

5 Results

5.1 Ineffective Retrieval Does Not Improve ASR

Table 1 details the contextual ASR performance on the AISHELL-1 test set utilizing a fixed Qwen3-ASR-0.6B backend. The baseline model, which excludes any hotword information, yields an O-CER of 2.09%, a B-CER of 39.16%, and an entity-level F1 score of 0.775. This indicates that while the ASR system accurately transcribes the vast majority of utterances, substantial room for improvement remains for hotword recognition. When the full lexicon is directly provided, the model experiences hallucinations, resulting in a marginal degradation in both O-CER (2.13%) and B-CER (39.69%). Although the conventional Fixed Top-5 retrieval method improves the F1 score to 0.857, it compromises overall recognition accuracy, deteriorating the O-CER to 2.28%.

In contrast, our proposed approach effectively mitigates the hallucination problem. It substantially reduces the B-CER from the baseline of 39.16% to 16.96% and elevates the F1 score to 0.902. Furthermore, unlike the Fixed Top-5 strategy, our approach significantly suppresses the false-positive rate on negative samples. Consequently, it enhances the recognition of biasing words without degrading the transcription quality of general speech; rather, it improves the O-CER to 1.88%. These results demonstrate that the proposed pipeline achieves an optimal balance between rare word biasing and general speech recognition, drastically closing the gap with the Ground Truth (GT) upper bound.

5.2 Pipeline Effectiveness Holds Across Peer Text Encoders

Since our method functions as a pipeline with a pluggable text encoder, a natural question arises as to whether the ASR performance gains are strictly dependent on a specific encoder, or if these improvements are fundamentally attributable to the proposed fusion and admission mechanism. By keeping the SW-pinyin, Levenshtein cues, precise adaptive thresholds, and the ASR backend fixed, we evaluate SONAR, BGE-M3, and Qwen as equivalent peers (Table 2).

For all evaluated encoders, the Fixed Top-5 strategy consistently yields a high false alarm rate and remains ineffective. In contrast, the application of Multi-Metric Scoring successfully restores high-precision retrieval performance. Correspondingly, the overall entity-level ASR F1 scores stabilize within a narrow yet strong range (0.902–0.903), all of which significantly outperform both the baseline (0.775) and the Fixed Top-K approach ($\approx$ 0.857). Thus, as long as the gating stage of the pipeline operates successfully, effective retrieval reliably translates into ASR improvements. This confirms that the causal link is at the method level, rather than being dependent on the specific encoder brand.

5.3 The Pipeline Transfers Across Backbone Scales

A natural concern is whether thresholds calibrated on one backbone remain valid for another. We therefore replace Qwen3-ASR-0.6B with Qwen3-ASR-1.7B, keeping the retrieval stage, the operating point, and the hotword list *entirely unchanged*. The larger backbone naturally achieves higher standalone performance (35.29% vs. 39.16% B-CER), yet the benefit of EHRA is essentially identical in relative terms: a 56.7% B-CER reduction at 0.6B and a 57.2% reduction at 1.7B. No re-tuning was required.

We additionally re-derived the hotword list from the 1.7B baseline transcripts using the same thresholds. This raises retrieval recall from 85.24% to 87.47%, but *does not* improve ASR (15.38% vs. 15.10% B-CER). The retrieval stage can therefore be instantiated with a small, inexpensive backbone and its output reused by a larger one—a practical decoupling of retrieval cost from recognition quality.

6 Conclusion

This study investigates Mandarin hotword-aware speech recognition and establishes that precision gains driven by an effective scoring pipeline yield substantial improvements in contextual ASR performance. Specifically, a modular pipeline is proposed to jointly integrate semantic and phonetic information during the retrieval phase. By incorporating a collaborative filtering mechanism, this framework ef-

fectively transforms a high-recall, false-alarm-prone ranking stage into a high-precision context generator. By substantially suppressing false-positive context injections on negative utterances, the proposed approach significantly reduced the biased character error rate (B-CER) from the 39.16% baseline to a range of 16.82%–17.24% and elevated the entity-level F1 score to 0.903. Simultaneously, it improved the overall transcription quality, reducing the overall CER (O-CER) to 1.88%. Furthermore, extensive evaluations across peer text encoders confirm that these robust ASR gains are fundamentally attributable to the proposed fusion and admission design, rather than specific encoder architectures. Future research directions include the exploration of end-to-end tunable gating mechanisms, the integration of larger open-domain lexicons, and systematic comparisons with jointly trained CLAS-style bias encoders.

References

Yuya Asano, Sabit Hassan, Paras Sharma, Anthony B Sicilia, Katherine Atwell, Diane Litman, and Malihe Alikhani. 2025. Contextual asr error handling with llms augmentation for goal-oriented conversational ai. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 374–386.

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.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. 2023. Sonar: sentence-level multimodal and language-agnostic representations. *arXiv preprint arXiv:2308.11466*.

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

Haoxiang Hou, Xun Gong, Wangyou Zhang, Wei Wang, and Yanmin Qian. 2025. Ranking and Selection of Bias Words for Contextual Bias Speech Recognition. In *Interspeech 2025*, pages 5183–5187.

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

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

Ian Williams, Anjuli Kannan, Petar Aleksic, David Rybach, and Tara Sainath. 2018. Contextual speech recognition in end-to-end neural network systems using beam search. In *Proc. Interspeech 2018*, pages 2227–2231.

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

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 641–649.

Shilin Zhou and Zhenghua Li. 2026. Jspg: Dynamic dictionary filtering via joint semantic-pinyin-glyph retrieval for chinese contextual asr. *arXiv preprint arXiv:2605.16896*.