RCbench: Benchmarking Retrospective Clarification in ASR

Wei-Ting Huang [1] **Chin-Yuan Yeh** [2] **De-Nian Yang** [3] **Ming-Syan Chen** [2 4]

Abstract

Human speakers often clarify newly introduced or ambiguous words later in an utterance, for example by providing an explanation, an acronym expansion, or a spelling cue. However, current ASR systems are generally unable to incorporate such later clarifications to revise earlier parts of the transcript. We define this problem as *retrospective clarification*, and present *RCBench*, a benchmark consisting of three scenarios: *acronym disambiguation*, *disambiguation by semantic explanation*, and *disambiguation by spelling*. On RCBench, we find a substantial gap between current ASR systems and human participants. In particular, within the same utterance, ASR baselines achieve 60–90% accuracy on control words that do not require retrospective clarification, suggesting that the main challenge lies in revising specific target words based on later clarification. Source code: `https://github.com/wweiting/RCbench`

1. Introduction

As speech becomes a primary interface for AI systems, ASR errors increasingly affect not only transcripts but also downstream AI applications that rely on them. A key challenge arises when speakers clarify an ambiguous word only after it has already been spoken.[1] For example, a participant may say, "Please apply the writing principles advocated by Larry muh-KEN-er-nee, spelled M-c-E-n-e-r-n-e-y." A human listener can use the later spelling cue to recover *McEnerney*. In contrast, an ASR system may initially transcribe the name as *McNerney* and fail to revise the earlier token, leaving the downstream system with an incorrect entity.

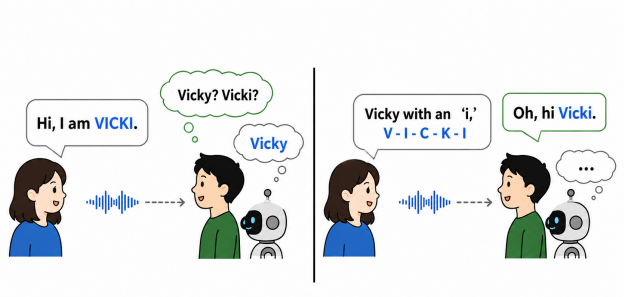

Figure 1. Example of Retrospective Clarification in Conversation

Despite the prevalence of such cases in real-world speech, ASR systems still struggle to revise earlier tokens based on later clarifications. We refer to this problem as *retrospective clarification*. Lattice-based and streaming ASR preserve alternatives for limited revisions during decoding, but they rely on decoding-time evidence and cannot revise earlier outputs using later clarification (Mangu et al., 2000; Sainath et al., 2023). Recent LLM-based methods improve ASR through N-best rescoring (Tur et al., 2024), integrated decoding (Chen et al., 2024), or post-hoc correction (Fang et al., 2025). However, these methods typically rely on candidates generated before the later clarification is interpreted, and thus may fail when the correct token is absent from the ASR candidate space.

Retrospective clarification poses two challenges: identifying the token that may require revision and locating the later span that clarifies it. Clarification can be semantic or spelling-based, without lexical overlap with the ASR error. The task therefore requires linking a potentially incorrect earlier token to delayed clarification first. Existing benchmarks on acronym disambiguation (Jimeno-Yepes et al., 2011; Chen et al., 2023) primarily targets established acronyms with known expansions, whereas retrospective clarification requires resolving novel or context-specific acronyms from later cues. Other benchmarks cover specialized vocabulary or contextual biasing (Huang et al., 2024; Piskala, 2026), but not clarification-driven revision.

[1]Department of Data Science, Soochow University, Taiwan [2]Graduate Institute of Communication Engineering, National Taiwan University, Taiwan [3]Institute of Information Science, Academia Sinica, Taiwan [4]Department of Electrical Engineering, National Taiwan University, Taiwan. Correspondence to: Wei-Ting Huang <bindy.huang@gmail.com>, Chin-Yuan Yeh <cyyeh@arbor.ee.ntu.edu.tw>.

The Machine Learning for Audio Workshop at the 43^{rd} International Conference on Machine Learning 2026, Seoul, South Korea.

[1]Conversation analysis has observed that speakers clarify their own words during utterance, a phenomenon known as *self-repair* (Schegloff et al., 1977). Such repairs are central to *grounding*, the collaborative process through which speakers and listeners establish mutual understanding (Clark & Brennan, 1991). Human listeners revise provisional interpretations once clarification arrives. ASR systems, in contrast, commit to a token and move on.

Table 1. Coverage of clarification phenomena in existing ASR benchmarks. Spell. = spelling-based; Acro. = acronym; Rare. = rare or out-of-vocabulary lexical items such as names, locations, and domain-specific terminology; Retro-Revis. = whether later evidence is required to revise earlier recognition outputs

Benchmark	Spell.	Acro.	Rare.	Retro-Revis.
Czech 2014 (Czech, 2014)	✓			
Jimeno 2011 (Jimeno-Yepes et al., 2011)		✓	✓	
GLADIS (Chen et al., 2023)		✓		
CB-Conformer (Xu et al., 2023)			✓	
Earnings21 (Rio et al., 2021)			✓	
ConEC (Huang et al., 2024)			✓	
ContextASR (Wang et al., 2025)			✓	
ProfASR (Piskala, 2026)			✓	
RCBench (ours)	✓	✓	✓	✓

We therefore introduce *RCBench*, a benchmark for evaluating retrospective clarification in ASR. RCBench includes 180 sentences across three categories: *acronym disambiguation*, where the speaker provides the full form of the intended abbreviation; *semantic explanation*, where the intended word is clarified by its meaning or definition; and *spelling-based clarification*, where novel words and foreign names are resolved through explicit spelling cues. Each sentence is recorded by five speakers.

Across all retrospective clarification categories, ASR baselines fall substantially behind human performance. Human achieves the highest accuracy in every category, while no baseline exceeds 2% accuracy in acronym disambiguation and most remain below 20% in the other categories. In contrast, control-group accuracy reaches 60–90%, where the control group consists of non-target words that do not require retrospective clarification. This contrast suggests that the main failure is not general transcription quality, but the ability to link later clarifications to earlier target words. Retrospective clarification thus exposes a distinct gap in current ASR systems: recognizing the surrounding utterance does not imply the ability to revise the clarified word.

2. Related Work

2.1. ASR Hypothesis Revision and Correction

Automatic speech recognition (ASR) has long studied mechanisms for maintaining and revising competing hypotheses as additional evidence becomes available. However, these mechanisms primarily address system-side uncertainty or instability rather than clarification-driven cases, where later discourse explicitly supplies explanatory evidence that changes the interpretation of an earlier utterance.

Lattice-based methods preserve alternative hypotheses in compact weighted graphs over possible word sequences (Mangu et al., 2000; Mohri et al., 2002; Byrne, 2006; Kombrink et al., 2011), allowing decoders to defer commitment among competing paths. While such methods support revision only when the correct alternative is preserved in the lattice, the revision space remains constrained by recognition-time hypotheses and their associated acoustic-language model scores (Mangu et al., 2000; Mohri et al., 2002). Streaming and incremental ASR systems permit limited revision of partial outputs through backtracking or hypothesis replacement to improve stability and reduce flickering (Selfridge et al., 2011; McGraw & Gruenstein, 2012; Sainath et al., 2023; Zhang et al., 2024). These revisions are primarily designed to manage the latency–stability tradeoff for unstable outputs during streaming recognition, and are not designed to identify a later clarification and use it to revise a specific earlier token (Baumann et al., 2009).

Recent LLM-based approaches improve ASR through N-best rescoring, integrated decoding, and post-hoc transcript correction. N-best rescoring methods (Udagawa et al., 2022; CHEN et al., 2023; Ogawa et al., 2024; Chen, 2025) use LLMs to rescore or rewrite candidate transcripts produced by an ASR system, but their revision is limited to hypotheses already present in the candidate list. If the intended form is absent, later clarification cannot introduce it. Integrated decoding methods (Seide et al., 2024; Chen et al., 2024; Wan et al., 2026; Xia et al., 2026) incorporate LLMs into the recognition pipeline, sometimes with access to acoustic features (Seide et al., 2024), but they still operate during recognition rather than explicitly identifying a later clarification and linking it back to an earlier target. Post-hoc correction methods (Singla et al., 2022; Fang et al., 2025) edit completed transcripts, but typically perform general transcript refinement rather than clarification-grounded revision of a specific earlier token. Thus, existing LLM-based ASR methods do not directly evaluate or target the core capability required by retrospective clarification: using delayed clarification cues to revise an earlier recognition error.

2.2. Related Benchmarks

Existing ASR benchmarks seldom test retrospective clarification, as summarized in Table 1. Some evaluate spelled-out words (Czech, 2014) or acronym disambiguation (Jimeno-Yepes et al., 2011; Chen et al., 2023; Haffoudhi et al., 2026), but these settings typically assume that the ambiguous mention is already correctly observed, or that the intended expansion belongs to a known inventory. Other benchmarks focus on rare words, domain terminology, named entities, or contextual biasing (Xu et al., 2023; Rio et al., 2021; Huang et al., 2024; Wang et al., 2025; Piskala, 2026). These benchmarks evaluate recognition under difficult lexical conditions, but not whether a system can revise an earlier ASR error using a clarification that appears later in the same utterance. RCBench addresses this gap by evaluating clarification-driven revision across acronym disambiguation, semantic explanation, and spelling-based clarification.

Category	Subcategory	Example Target	Example Sentence
Acronym Disambiguation	First Character Only	BEAM / BEEM	We utilized BEAM, the Binary Extraction and Analysis Module, and BEEM, the Binary Extraction and Evaluation Module.
		MYTH / MITH	Users can access MYTH, the Multimodal Yielding Task Heuristic, or MITH, the Meaning Informed Transcript Harmonization.
	First-few Characters	CoVer / CovEr	In this setup, CoVer means Contextual Verification, while CovEr refers to Coverage Error Reporting.
		SpArC / SpARC	Our framework uses SpArC for Sparse Argumentation Code, while SpARC refers to Spatial Analysis and Reasoning Component.
Disambiguation by Semantic Explanation	Compound Word Confusion	ArcAgent / ArkAgent	The tool we're deploying is ArcAgent—'arc' like a curve, not ArkAgent, 'ark' like Noah's Ark.
		ByteBridge / BiteBridge	What I actually said was ByteBridge. I mean 'byte' like digital data, instead of BiteBridge, 'bite' like taking a bite of food.
	Explain by Character Hint	Katherine / Catherine	The name is Katherine with a K, not Catherine with a C.
		Kris / Chris	I'm referring to Kris with a K, not Chris with Ch.
Disambiguation by Spelling	Novel Word	stasult	The experiment finally produced a stasult. The letters are S-T-A-S-U-L-T. I mean a result that remains stable across repeated tests.
		viviage	The poster has a memorable viviage. The letters are V-I-V-I-A-G-E. This means a vivid image that stays clearly in your mind after you see it.
	Asian Name Translation	CHUNG, YU-CHIAO	The entry for Chung Yu-Chiao is ready—it is formatted as C-H-U-N-G-Y-U-C-H-I-A-O.
		LING, MEI-CHIEH	You can identify Ling Mei-Chieh by her ID tag, which reads L-I-N-G-M-E-I-C-H-I-E-H.

Table 2. Challenging phenomena that existing methods fail to handle.

3. RCbench

3.1. Design Principles

Retrospective clarification refers to cases where speakers provide additional explanations after an utterance to clarify what they intended to say. Existing ASR methods remain limited in handling such cases because they cannot reliably backtrack to identify incorrect words in earlier utterances or extract useful later information to support revision. In addition, existing benchmarks often focus on limited subproblems, such as acronym disambiguation, rather than systematically covering the full scope of retrospective clarification. As a result, we present a new benchmark that highlights these challenges and evaluates whether ASR systems can revise earlier recognition outputs based on later explanatory context.

3.2. Phenomenon Taxonomy

To highlight retrospective clarification problems in real-world speech, we construct a dataset covering several representative categories. Examples are shown in Table 2.

The first category, ***Acronym Disambiguation***, includes *First Character Only* and *First-few Characters*, where the target terms are acronyms or abbreviations derived from multiple words. The second category, ***Disambiguation by Semantic Explanation***, includes *Compound Word Confusion* and *Explain by Character Hint*, where speakers clarify a word through semantic or character-level explanations. The final category, ***Disambiguation by Spelling***, includes *Novel Word* and *Foreign Name Translation*, where the target terms are likely to be out-of-vocabulary, and speakers therefore tend to spell out the target word. We also define a ***Control*** set over non-target words, which do not require retrospective clarification.

3.3. Evaluation & Metrics

RCBench does not evaluate general-purpose ASR transcription quality. Instead, we focus on accuracy and average targeted word error rate (WER), which measure the correctness of labeled target words that have corresponding later clarifications. For instance, given the sentence *"Our loyal users are moving toward a subscriever model, spelled S-U-B-S-C-R-I-E-V-E-R. This refers to a lifetime commitment where you subscribe once and enjoy the service forever without recurring fees."*, we only evaluate the correctness of the target word *subscriever*. We report accuracy to measure the proportion of sentences in which the target regions are correctly recognized, and average targeted WER to quantify the degree of recognition errors within the target regions. We also observe that ASR systems rarely make errors outside the targeted regions.

Benchmark	Whisper-large-v3		ProGRes		FHMV		Typeless AI		Human	
	ACC	WER	ACC	WER	ACC	WER	ACC	WER	ACC	WER
First-char Abbr.	0	65.67	0	71.33	0	67.33	1.33	59.00	20.00	48.00
Few-char Abbr.	0	39.67	0	61.00	0	51.33	0	42.33	16.00	38.00
Compound Word	0	69	0	77.33	2.00	75.00	3.33	62.67	8.00	64.00
Explain by Hint	0	64	4.00	75.34	8.00	62.33	67.34	21.33	84.00	12.00
Novel Word	16	83.33	10.67	89.33	10.00	89.33	68.67	31.33	84.00	16.00
Translated Name	30.67	24.14	2.00	91.11	14.67	65.68	51.33	35.06	84.00	10.78
Control	**86.98**	**14.27**	**63.24**	**37.94**	**76.13**	**27.74**	**83.73**	**17.80**	84.73	17.99

Table 3. Baseline Comparison.

4. Experiments

4.1. Dataset

To evaluate existing baselines, we construct a dataset containing 180 sentences. The dataset covers six subcategories: First Character Only and First-few Characters for acronym disambiguation; Compound Word Confusion and Explain by Character Hint for semantic explanation; and Novel Word and Asian Name Translation for spelling-based clarification. Each subcategory includes 30 distinct sentences. We ask five human speakers to record each sentence, yielding 900 speech recordings in total. This design ensures balanced coverage across clarification types while capturing variation across speakers and accents.

4.2. Baselines

We evaluate several baselines that use LLMs to improve ASR outputs. First, we include ProGRes (Tur et al., 2024), which improves ASR performance by integrating confidence scores and LLM sequence scoring with a prompt-based generative mechanism to refine and expand N-best hypotheses. We also include Fewer Hallucinations, More Verification (Fang et al., 2025) as a post-hoc correction baseline. This work proposes a three-step framework for reducing hallucinations in ASR correction, consisting of error pre-detection, CoT-based iterative correction, and reasoning process verification.

In addition to these academic baselines, we include Whisper-large-v3 as a conventional ASR baseline. We also evaluate Typeless AI, a commercial AI speech transcription tool that converts users' spoken input into polished written messages.

4.3. Main Results

We evaluate accuracy to measure how many target words are correctly recognized, and average targeted WER to quantify the degree of recognition errors within the target words. Accuracy measures whether the target word is fully corrected, while targeted WER captures the degree of remaining errors. The results indicate that all ASR-based baselines perform substantially worse than human prediction on retrospective clarification, showing that they do not reliably revise earlier target words even when later clarification is available. Whisper-large-v3 shows the weakest performance, achieving 0% accuracy in four subcategories and WERs ranging from 45.33% to 80.67%. The two research baselines also fail on acronym disambiguation, with 0% accuracy on both abbreviation-related tests and WERs above 50% across all target categories. Typeless AI performs best among the system baselines, reaching 67.34% ACC on Explain by Hint and 68.67% on Novel Word, but it still remains below human performance in all target-category accuracy scores. In contrast, all baselines achieve much higher accuracy on the control group, ranging from 63.24% to 87.39%, where no retrospective revision is required. This contrast suggests that the main difficulty lies not in transcription quality, but in using later clarifying evidence to revise earlier target words. This gap is especially notable because the same systems often recognize the surrounding context correctly, but fail to use that context as evidence for revising the target word. These results indicate that retrospective clarification is not simply a rare-word recognition problem, but a targeted revision problem. Detailed results are reported in Table 3.

5. Conclusion

This limitation is increasingly consequential as speech and dictation become primary interfaces for AI systems. Our results show that current methods can transcribe ordinary control cases reasonably well, yet fail when correct recognition requires using later clarification to revise an earlier target word. The large gap between control accuracy and retrospective-clarification accuracy indicates that current systems lack a dedicated capability for clarification-driven revision. RCbench therefore highlights a concrete limitation in current ASR evaluation and points to the need for systems that can treat later spoken evidence as grounds for revising earlier hypotheses. These findings suggest that future ASR evaluation should explicitly measure whether systems can revise earlier outputs using later clarification.

References

Baumann, T., Atterer, M., and Schlangen, D. Assessing and improving the performance of speech recognition for incremental systems. In Ostendorf, M., Collins, M., Narayanan, S., Oard, D. W., and Vanderwende, L. (eds.), *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 380–388, Boulder, Colorado, June 2009. Association for Computational Linguistics. URL https://aclanthology.org/N09-1043/.

Byrne, W. Minimum bayes risk estimation and decoding in large vocabulary continuous speech recognition. *IEICE Transactions on Information and Systems*, E89-D(3): 900–907, March 2006. doi: 10.1093/ietisy/e89-d.3.900. URL https://doi.org/10.1093/ietisy/e89-d.3.900.

Chen, C. Advancing speech-to-text adaptation for large speech models. 2025. doi: 10.32657/10356/201854. URL https://doi.org/10.32657/10356/201854.

CHEN, C., Hu, Y., Yang, C.-H. H., Siniscalchi, S. M., Chen, P.-Y., and Chng, E.-S. Hyporadise: An open baseline for generative speech recognition with large language models. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 31665–31688. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6492267465a7ac507be1f9fd1174e78d-Paper-Datasets_and_Benchmarks.pdf.

Chen, L., Varoquaux, G., and Suchanek, F. M. GLADIS: A general and large acronym disambiguation benchmark. In Vlachos, A. and Augenstein, I. (eds.), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2073–2088. Association for Computational Linguistics, May 2023. doi: 10.18653/v1/2023.eacl-main.152. URL https://aclanthology.org/2023.eacl-main.152/.

Chen, P., Sun, S., Shan, C., Yang, Q., and Xie, L. Streaming decoder-only automatic speech recognition with discrete speech units: A pilot study. In *Interspeech 2024*, pp. 4468–4472. ISCA, September 2024. doi: 10.21437/Interspeech.2024-1853. URL https://doi.org/10.21437/interspeech.2024-1853.

Clark, H. H. and Brennan, S. E. *Grounding in communication.*, pp. 127–149. American Psychological Association, 1991. doi: 10.1037/10096-006. URL https://doi.org/10.1037/10096-006.

Czech, L. *A System for Recognizing Natural Spelling of English Words*. PhD thesis, Diploma Thesis, Karlsruhe Institute of Technology, 2014.

Fang, Y., Chen, B., Peng, J., Li, X., Xi, Y., Zhang, C., and Zhong, G. Fewer hallucinations, more verification: A three-stage llm-based framework for asr error correction. *arXiv preprint*, arXiv:2505.24347, 2025. URL https://arxiv.org/abs/2505.24347.

Haffoudhi, S., Suchanek, F. M., and Holzenberger, N. LELA: an LLM-based Entity Linking Approach with Zero-Shot Domain Adaptation. working paper or preprint, January 2026. URL https://hal.science/hal-05445830.

Huang, R., Yarmohammadi, M., Trmal, J., Liu, J., Raj, D., Garcia, L. P., Ivanov, A., Ehlen, P., Yu, M., Rastrow, A., Povey, D., and Khudanpur, S. ConEC: Earnings call dataset with real-world contexts for benchmarking contextual speech recognition. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N. (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 3700–3706. ELRA and ICCL, May 2024. URL https://aclanthology.org/2024.lrec-main.328/.

Jimeno-Yepes, A. J., McInnes, B. T., and Aronson, A. R. Exploiting mesh indexing in medline to generate a data set for word sense disambiguation. *BMC Bioinformatics*, 12(1), June 2011. doi: 10.1186/1471-2105-12-223. URL https://doi.org/10.1186/1471-2105-12-223.

Kombrink, S., Mikolov, T., Karafiát, M., and Burget, L. Recurrent neural network based language modeling in meeting recognition. In *Interspeech 2011*, pp. 2877–2880. ISCA, August 2011. doi: 10.21437/interspeech.2011-720. URL https://doi.org/10.21437/interspeech.2011-720.

Mangu, L., Brill, E., and Stolcke, A. Finding consensus in speech recognition: word error minimization and other applications of confusion networks. *Computer Speech & Language*, 14(4):373–400, October 2000. doi: 10.1006/csla.2000.0152. URL https://doi.org/10.1006/csla.2000.0152.

McGraw, I. and Gruenstein, A. Estimating word-stability during incremental speech recognition. In *Interspeech 2012*, pp. 1019–1022. ISCA, September 2012. doi: 10.21437/interspeech.2012-301. URL https://doi.org/10.21437/interspeech.2012-301.

Mohri, M., Pereira, F., and Riley, M. Weighted finite-state transducers in speech recognition. *Computer Speech &*

Language, 16(1):69–88, January 2002. doi: 10.1006/csla.2001.0184. URL https://doi.org/10.1006/csla.2001.0184.

Ogawa, A., Kamo, N., Matsuura, K., Ashihara, T., Moriya, T., Kano, T., Tawara, N., and Delcroix, M. Applying llms for rescoring n-best asr hypotheses of casual conversations: Effects of domain adaptation and context carry-overg. *arXiv preprint*, arXiv:2406.18972, 2024. URL https://arxiv.org/abs/2406.18972.

Piskala, D. ProfASR-bench: A professional-talk ASR dataset for high-stakes applications exposing the context-utilization gap, 2026. URL https://openreview.net/forum?id=o2k64ag4HA.

Rio, M. D., Delworth, N., Westerman, R., Huang, M., Bhandari, N., Palakapilly, J., McNamara, Q., Dong, J., Zelasko, P., and Jette, M. Earnings-21: A practical benchmark for asr in the wild. *arXiv preprint*, arXiv:2104.11348, 2021. URL https://arxiv.org/abs/2104.11348.

Sainath, T. N., Prabhavalkar, R., Bapna, A., Zhang, Y., Huo, Z., Chen, Z., Li, B., Wang, W., and Strohman, T. Joist: A joint speech and text streaming model for asr. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pp. 52–59. IEEE, January 2023. doi: 10.1109/SLT54892.2023.10022774. URL https://doi.org/10.1109/slt54892.2023.10022774.

Schegloff, E. A., Jefferson, G., and Sacks, H. The preference for self-correction in the organization of repair in conversation. *Language*, 53(2):361–382, June 1977. doi: 10.2307/413107. URL https://doi.org/10.2307/413107.

Seide, F., Shi, Y., Doulaty, M., Gaur, Y., Jia, J., and Wu, C. Speech reallm – real-time speech recognition with multimodal language models by teaching the flow of time. In *Interspeech 2024*, pp. 1900–1904. ISCA, September 2024. doi: 10.21437/Interspeech.2024-571. URL https://doi.org/10.21437/interspeech.2024-571.

Selfridge, E., Arizmendi, I., Heeman, P., and Williams, J. Stability and accuracy in incremental speech recognition. In Chai, J. Y., Moore, J. D., Passonneau, R. J., and Traum, D. R. (eds.), *Proceedings of the SIGDIAL 2011 Conference*, pp. 110–119, Portland, Oregon, June 2011. Association for Computational Linguistics. URL https://aclanthology.org/W11-2014/.

Singla, K., Jalalvand, S., Kim, Y.-J., Price, R., Pressel, D., and Bangalore, S. Seq-2-seq based refinement of asr output for spoken name capture. In *Interspeech 2022*, pp. 3963–3967. ISCA, September 2022. doi: 10.21437/Interspeech.2022-10885. URL https://doi.org/10.21437/interspeech.2022-10885.

Tur, A. D., Moumen, A., and Ravanelli, M. Progres: Prompted generative rescoring on asr n-best. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pp. 600–607. IEEE, December 2024. doi: 10.1109/SLT61566.2024.10832194. URL https://doi.org/10.1109/slt61566.2024.10832194.

Udagawa, T., Suzuki, M., Kurata, G., Itoh, N., and Saon, G. Effect and analysis of large-scale language model rescoring on competitive asr systems. In *Interspeech 2022*, pp. 3919–3923. ISCA, September 2022. doi: 10.21437/Interspeech.2022-11123. URL https://doi.org/10.21437/interspeech.2022-11123.

Wan, G., Zhang, W., Zhang, J.-X., Xiong, S., Gao, J., and Ye, Z. Streaming speech recognition with decoder-only large language models and latency optimization. *arXiv preprint*, arXiv:2601.22779, 2026. URL https://arxiv.org/abs/2601.22779.

Wang, H., Ma, L., Guo, D., Wang, X., Xie, L., Xu, J., and Lin, J. Contextasr-bench: A massive contextual speech recognition benchmark. *arXiv preprint*, arXiv:2507.05727, 2025. URL https://arxiv.org/abs/2507.05727.

Xia, Y., Tang, J., Hou, J., Xu, G., and Yao, H. Uni-asr: Unified llm-based architecture for non-streaming and streaming automatic speech recognition. *arXiv preprint*, arXiv:2603.11123, 2026. URL https://arxiv.org/abs/2603.11123.

Xu, Y., Liu, B., Huang, Q., Song, X., Wu, Z., Kang, S., and Meng, H. CB-conformer: Contextual biasing conformer for biased word recognition. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10095469.

Zhang, Z., Gan, R., Yuan, P., and Jin, H. Correcting pronoun homophones with subtle semantics in Chinese speech recognition. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N. (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 4047–4058, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.360/.