Context Projector: Complementary Keyword and Dialogue Context Embeddings for LLM-based ASR

Esaú Villatoro-Tello [1,**], *Sergio Burdisso* [1,**], *Shashi Kumar* [1,2], *Hasindri Watawana* [1,2], *Srikanth Madikeri*[3], *Manjunath K E*[4], *Jeena Prakash*[4], *Thibault Bañeras-Roux* [1], *Kadri Hacioglu*[4], *Petr Motlicek* [1,5], *Andreas Stolcke*[4]

[1]Idiap Research Institute, Switzerland
[2]EPFL, Switzerland
[3]University of Zurich, Switzerland
[4]Uniphore, USA & India
[5]Brno University of Technology, Czech Republic
{esau.villatoro, sergio.burdisso}@idiap.ch

Abstract

Contact-center dialogue systems require both accurate transcripts and reliable recognition of business-critical entities. We introduce a hybrid keyword and *context projector* (CP) approach for LLM-based ASR. Rather than appending raw dialogue history, CP encodes previous turns into compact contextual tokens, complemented by automatically extracted keywords. The module is trained while keeping the backbone frozen, enabling parameter-efficient adaptation. Experiments on realistic, multi-domain contact-center data show that naive raw-context prompting can degrade performance, whereas the proposed approach improves the tradeoff between overall and bias-word error, achieving average relative reductions of up to 2.5% overall and 7.2% on bias words. These results show our approach to be an effective and practical enhancement of context-aware LLM-based ASR.

Index Terms: LLM-based ASR, spoken dialogue systems, contact-center speech, contextual ASR, entity-aware evaluation, bias word error rate

1. Introduction

Spoken dialogue systems for contact centers require reliable transcription not only at the word level but also for business-critical entities. Although word error rate (WER) is the standard ASR metric, it does not fully reflect performance in task-oriented applications where downstream components such as dialogue state tracking, routing, retrieval, compliance, and agent assist depend on accurate entity recognition [1, 2, 3]. This makes entity-sensitive metrics such as bias-word error rate (BWER) particularly important. Recent LLM-based ASR architectures offer a promising alternative to conventional end-to-end ASR systems by connecting a pretrained speech encoder to an instruction-tuned LLM through a lightweight speech projector [4, 5, 6]. However, most approaches rely on fixed prompts without contextual information [6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17], limiting their effectiveness in multi-turn dialogue where recognition often depends on conversational history.

A natural solution is to include dialogue history directly in the prompt, leveraging the long-context capabilities of modern LLMs. However, this becomes computationally expensive and increasingly unreliable as context lengths increase [18]. Prior work has therefore explored using only the previous turn [11] or compressed and structured contextual representations [19, 20]. While these methods demonstrate the value of contextual information, a key challenge remains: how to exploit long-range dialogue context without the degradation observed in long-context LLMs [18]. To address this issue, we propose a drop-in context projector that converts previous dialogue utterances into compact contextual tokens, complemented by keyword cues automatically extracted from dialogue history. Figure 1 illustrates the proposed architecture. Although keywords and bias lists are well-established in traditional ASR systems [21, 22, 23], to the best of our knowledge this is the first work to integrate them into an LLM-based ASR framework together with an efficient context projection module.

Our contributions are: *(i)* a parameter-efficient context projection module that combines with a frozen LLM-based ASR backbone,[1] *(ii)* a systematic study of context-injection strategies, showing that raw context can degrade WER, whereas combining keywords with projected context improves the WER–BWER tradeoff, and *(iii)* extensive evaluation on a multi-domain contact-center dataset, demonstrating consistent BWER improvements while maintaining competitive WER.

2. Method

2.1. Base SpeechLLM Model

We use SLAM-ASR as our base LLM-based ASR model [6], whose simple yet effective speech projection module provides a strong foundation for our experiments. Specifically, the speech projector, $sp(\cdot)$, is defined as:

$$\mathbf{e}_i = sp(\mathbf{z}_i) = \mathrm{ReLU}(\mathbf{z}_i W_1 + b_1)W_2 + b_2, \qquad (1)$$

where $\mathbf{z}_i$ denotes the i-th downsampled audio feature, obtained by concatenating k consecutive encoded frames along the temporal dimension. The resulting projection $\mathbf{e}_i$ matches the dimensionality of the LLM input embeddings.

Thus, input audio consisting of n downsampled features, $\mathbf{z}_1, \ldots, \mathbf{z}_n$, is fed to the LLM using the following prompt:

```
<s>USER: Transcribe speech to text.
Context:{keywords},{context}. Speech:{speech}
ASSISTANT: {transcript}</s>
```

**indicates the corresponding author.

[1]https://github.com/idiap/llm-asr-context-projector

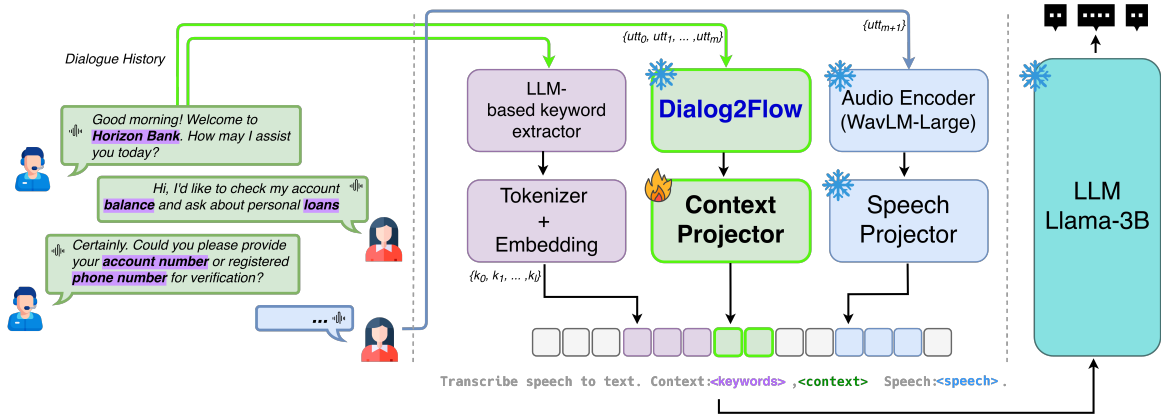

Figure 1: *Proposed dialogue-aware SpeechLLM architecture. A frozen ASR pipeline maps speech to LLM embeddings via $sp(\cdot)$, augmented with a context projector $cp(\cdot)$ encoding dialogue-history embeddings into contextual tokens. Optional keyword cues from previous turns are included in the prompt, and the LLM jointly conditions on speech, context, and keywords to generate the transcript.*

Here, `{speech}` is replaced by the sequence of n projected speech embeddings, $sp(\mathbf{z}_1), \ldots, sp(\mathbf{z}_n)$ before being fed to the LLM. The `{keywords}` field is a comma-separated list of salient keywords automatically extracted from the dialogue history. The `{context}` field is replaced by the sequence of m projected contextual tokens (Section 2.2).[2] During training, `{transcript}` denotes the ground-truth transcription; at inference time, it is generated autoregressively.

All experiments use WavLM-Large [24] as the speech encoder and Llama 3.2 3B Instruct [25] as the decoder LLM. We refer to this configuration as the *base* model. Following the original SLAM-ASR setup, only the speech projector parameters (Eq. 1) are updated during training, while both the speech encoder and the LLM remain frozen.[3]

2.2. The Context Projector: $cp(\cdot)$

To mitigate the degradation caused by directly injecting long-range raw contextual information into LLM-based ASR, we introduce $cp(\cdot)$, a *context projector* that maps dialogue history into a compact set of contextual tokens (see Figure 1).

The context projector is a simple drop-in extension to an existing LLM-based ASR system. After training (or adopting) a base model, we freeze all its components and train only the new projector module. This ensures that optimization focuses solely on learning an effective contextual representation, while avoiding the instability that arises when jointly fine-tuning all the LLM-based ASR components.[4]

Formally, let $\mathbf{x}_1, \ldots, \mathbf{x}_m$ denote the sequence of turn-level contextual embeddings extracted from the dialogue history. These are projected via $cp(\cdot)$ into a compact sequence of m contextual tokens. The resulting projected tokens are replaced by the `{context}` field before being feed to the LLM. The context projector shares the same architecture as the speech projector $sp(\cdot)$ (Equation 1), differing only in input dimensionality, as it operates on sentence-level embedding space.

[2]Depending on the experimental setting, `{context}` may alternatively be replaced by m raw contextual utterances (Section 3.1)..

[3]A *base* model is trained with empty `{keywords}`,`{context}` fields.

[4]In complementary experiments, unfreezing the backbone models consistently led to unstable training and degraded performance.

2.3. Dialog2Flow Sentence Embeddings

A key component of the context projector is a sentence-level embedding step that transforms raw dialogue utterances into structured contextual representations. In all experiments, we employ Dialog2Flow (D2F) embeddings [26].

D2F provides sentence representations optimized for task-oriented dialogue modeling. Unlike conventional sentence embeddings that primarily capture semantic similarity, D2F maps utterances into a latent space organized according to their communicative function or underlying dialogue action. This structure is learned using a soft contrastive objective during pretraining on 3.4 million utterances spanning 52 domains. As a result, conversational turns that serve similar functional roles are mapped to nearby regions of the embedding space.

This action-aware organization makes D2F particularly suitable for contextual projection in dialogue ASR, as it provides compact representations that preserve workflow-relevant information while reducing raw textual variability. In our experiments, we use the `dialog2flow-joint-bert-base` model.[5]

2.4. The Keywords Extractor

For keyword extraction, we employ Gemma3 (27B) [27]. Given each utterance in the dialogue history, we prompt the model to extract salient single- and multi-word keywords intended to populate a lightweight memory mechanism.

We select Gemma based on a zero-shot comparison of entity recall. Across all evaluated domains, Gemma achieves the highest recall (90.6%), outperforming Phi 4-14B (86.3%) [28], Llama 3.2-8B (80.7%) [25], DeepSeek-R1-32B (78.7%) [29], and Qwen3-30B (75%) [30]. Because our hybrid context pipeline relies on preserving domain entities within the keyword memory, we prioritize recall and therefore adopt Gemma as the keyword extractor.

2.5. Training and Inference Implementation Details

All experimental settings follow the original SLAM-ASR work [6]. The speech encoder produces representations at

[5]`https://huggingface.co/sergioburdisso/dialog2flow-joint-bert-base`

Table 1: *DefinedAI dataset statistics (per-split view).*

Split	Domain	#Utts	#Entities	Total (h)
Train	Banking	26,496	9,852	54.5
	Healthcare	4,708	1,480	11.5
	Insurance	30,758	12,253	65.2
	Retail	10,981	2,938	27.8
	Teleco	18,889	6,392	50.9
Dev	Banking	1,447	527	2.9
	Healthcare	236	62	0.5
	Insurance	1,794	681	3.9
	Retail	623	111	1.5
	Teleco	1,130	382	3.1
Test	Banking	3,166	1,215	6.5
	Healthcare	488	154	1.1
	Insurance	3,670	1,432	7.7
	Retail	1,312	324	3.3
	Teleco	2,243	711	6.0
Total:		*107,941*	*38,514*	*246.4*

50 Hz. With a downsampling factor of $k = 5$, this yields downsampled features $\mathbf{z}_i$ at 100 ms resolution, resulting in 10 projected speech embeddings $sp(\mathbf{z}_i)$ per second of audio.

Both the speech and context projectors use a single hidden layer of dimension 2048. Decoding is performed using beam search with beam size 4. Models are trained with AdamW [31], using a learning rate of 10^{-4}, batch size 10, and early stopping based on development-set cross-entropy loss.

All computations are performed in *bfloat16* precision. Training is conducted on a single NVIDIA H100 GPU (80 GB VRAM), while inference is run on an NVIDIA RTX 3090 GPU. Each model is trained for 7 epochs: 5 for the speech projector, followed by 2 additional epochs for the context projector.

2.6. Datasets

Our experiments use scripted call-center conversations between agents and customers from Defined.ai.[6] We adopt the data split shown in Table 1, comprising 107,941 utterances, 38,514 entity mentions, and 246.4 total hours across train, development, and test sets. The corpus spans five domains—Banking, Healthcare, Insurance, Retail, and Telecommunications—enabling evaluation on heterogeneous vocabulary and entity distributions.

The data distribution reflects realistic production settings. Insurance and Banking are the largest domains (30,758 and 26,496 training utterances, respectively), whereas Healthcare is comparatively low-resource (4,708 training utterances). We preserve the original split to assess the robustness of both WER and BWER under domain and entity-frequency shifts, particularly in scenarios where contextual modeling is expected to benefit rare or sparsely observed entities.

2.7. Evaluation Metrics

We evaluate system performance using three complementary metrics: WER, BWER, and entity-level F1 (harmonic mean of precision and recall) [32]. Standard **WER** measures overall transcription quality, computed as the normalized edit distance between the hypothesis (hyp) and reference (ref), including substitutions, insertions, and deletions. **BWER** assesses robustness on domain-specific terminology. Entity spans are identified in the reference, and substitutions or deletions within these spans are counted. BWER is the ratio of entity-specific errors to the total number of entity tokens. **Entity-level F1** evaluates recovery of complete entity mentions. True positives (TPs) occur when a reference entity is exactly matched in the hypothesis, false negatives (FNs) when it is omitted or misrecognized. False positives (FPs) are spurious entities appearing in the hypothesis. Precision, recall, and F1 are computed from these counts.

[6] https://www.defined.ai

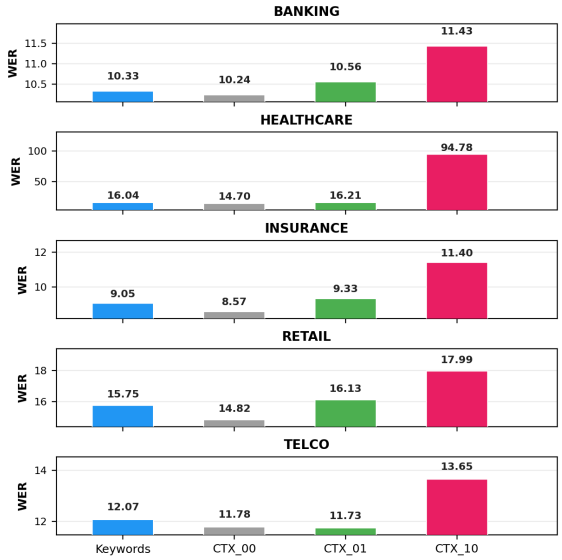

Figure 2: *ASR performance (% WER) across domains when different types of contextual information (keywords and previous turns) are appended directly to the prompt. CTX_00 denotes the* base *LLM-based ASR model without any context.*

3. Experimentation

3.1. Assessing the Impact of Adding Raw Context

Figure 2 shows that injecting contextual information in raw textual form is generally detrimental to recognition quality. Across domains, most prompt variants that append either keywords or previous dialogue turns (1 or 10) yield higher WER than the no-context model (CTX_00, i.e., the *base* model), indicating that naive context addition does not consistently improve transcription accuracy.

As discussed in Section 1, this degradation likely arises from prompt clutter and the increased difficulty for the LLM to disentangle acoustically grounded evidence from auxiliary textual input, particularly when the contextual content grows.[7]

3.2. Impact of Keywords + Context Projection

The results in Figure 3 support our complementary design, combining two context signals: (i) raw keywords tracked by the LLM-based keyword extractor, and (ii) projected representations of previous utterances via $cp(\cdot)$. We evaluate history windows ranging from a single previous turn up to the full available dialogue history (CTX=[1,5,10,ALL]), observing that context size is a key control variable. Increasing context generally improves performance until a domain-specific saturation point, beyond which additional turns yield limited gains or introduce variability. This suggests that useful dialogue context is often

[7] We do not report results for appending the entire dialogue history as raw context, as this often caused out-of-memory (OOM) errors due to batch size in our experiments.

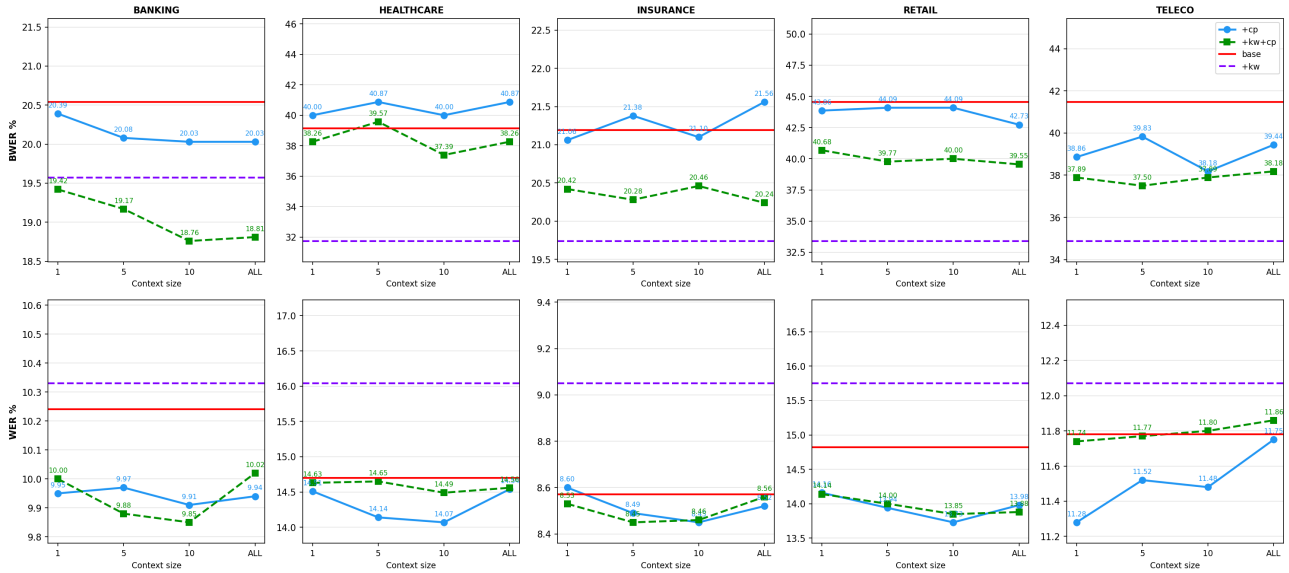

Figure 3: *ASR results across domains as a function of context size. WER and BWER are shown for context projection (+cp, blue solid) and hybrid keyword+context projection (+kw+cp, green dashed). The base model (red solid) and raw-keyword condition (+kw, purple dashed) are shown for reference. Results show that the complementary setting generally achieves a better WER–BWER tradeoff, especially at moderate context sizes.*

concentrated in a relatively short, recent window, typically the 10 previous turns.

A consistent pattern is that using only raw keywords (*base+kw*) tends to improve entity-focused recognition (BWER), but can degrade overall transcription quality (WER). This trend is evident in Table 2, where the *base+kw* setup achieves better BWER than *base*, but at the expense of higher WER. Intuitively, raw keywords bias decoding toward entity mentions, but without structured contextual grounding they may distract the model from acoustically supported tokens.

When keyword cues are combined with projected dialogue history (*base+kw+cp*), the WER–BWER tradeoff improves substantially. Across all domains, our proposed approach decreases BWER over *base* while keeping WER below base model performance in all cases. Compared with *base+kw* alone, the addition of $cp(\cdot)$ stabilizes decoding and restores general ASR quality, indicating that the context projector provides the structure necessary to make raw keyword information reliably useful. Overall, our complementary approach offers the most practical operating point for dialogue-aware ASR: stronger entity accuracy without incurring the WER penalty associated with raw-context prompting (Figure 2).

4. Conclusions

We introduced a hybrid keyword and context projector approach for LLM-based ASR that enhances contextual grounding without modifying the frozen backbone model. Experiments show that naively appending raw dialogue history degrades transcriptions, whereas projecting past turns into compact contextual tokens complemented with automatically extracted keywords provides a more robust WER–BWER tradeoff—achieving up to 2.5% relative improvement in WER, up to 7.2% in BWER, and up to 3.7% in F1-score on average. Future work will explore compressing past spoken utterances to further reduce the BWER gap between the *+kw* and *+kw+*$cp(\cdot)$ approaches.

Table 2: *WER, BWER, and NE-F1 results across five domains. Relative improvements (%) over the base model are shown in parentheses, with green for gains and red for degradations. CTX denotes the number of dialogue turns used as context.* ***Bold*** *and underlined mark the best and second-best results among context-augmented systems.*

Domain	Method	CTX	WER(Δ%)	BWER(Δ%)	F1
Banking	*base*	-	10.2	20.5	0.83
	+kw	-	10.3 (-0.9%)	19.6 (4.4%)	0.84
	+cp(·)	10	9.9 (2.9%)	20.0 (2.4%)	0.83
	+kw+cp(·)	10	**9.9** (2.9%)	**18.8** (8.3%)	**0.84**
Healthcare	*base*	-	14.7	39.1	0.67
	+kw	-	16.0 (-8.8%)	31.7 (18.9%)	0.75
	+cp(·)	10	**14.0** (4.7%)	40.0 (-2.3%)	0.65
	+kw+cp(·)	10	14.5 (1.4%)	**37.4** (4.4%)	**0.71**
Insurance	*base*	-	8.6	21.19	0.82
	+kw	-	9.0 (-4.7%)	19.7 (7.1%)	0.84
	+cp(·)	10	**8.5** (1.2%)	21.1 (0.4%)	0.82
	+kw+cp(·)	5	**8.5** (1.2%)	**20.3** (4.2%)	**0.83**
Retail	*base*	-	14.8	44.6	0.68
	+kw	-	15.8 (-6.8%)	33.4 (25.1%)	0.77
	+cp(·)	10	**13.7** (7.4%)	44.1 (1.1%)	0.70
	+kw+cp(·)	10	13.9 (6.1%)	**40.0** (10.3%)	**0.73**
Teleco	*base*	-	11.8	41.5	0.69
	+kw	-	12.1 (-2.5%)	34.9 (15.9%)	0.74
	+cp(·)	1	**11.3** (4.2%)	38.9 (6.3%)	0.70
	+kw+cp(·)	1	11.7 (0.9%)	**37.9** (8.7%)	**0.71**
Average Δ%	*+kw*	-	-4.7%	**14.3%**	**7.2%**
	+cp(·)	≈10	**4.1%**	1.6%	0.3%
	+kw+cp(·)	≈10	2.5%	7.2%	3.7%

5. Acknowledgments

This work was supported by an Idiap Research Institute and Uniphore collaboration project. Part of this work was also supported by EU Horizon 2020 project ELOQUENCE.

6. Generative AI Use Disclosure

Generative AI was used to proofread the paper, fix orthographic and grammatical errors, and make long sections more concise. Additionally, AI tools were used to improve the look and feel of the main plots of the paper.

7. References

[1] Y.-Y. Wang, L. Deng, and A. Acero, "Spoken language understanding," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 16–31, 2005.

[2] D. Hakkani-Tür, G. Tür, A. Celikyilmaz, Y.-N. Chen, J. Gao, L. Deng, and Y.-Y. Wang, "Multi-domain joint semantic frame parsing using bi-directional RNN-LSTM," in *Proc. Interspeech*, 2016, pp. 715–719.

[3] M. Henderson, B. Thomson, and J. D. Williams, "The second Dialog State Tracking challenge," in *Proceedings of the 15th annual meeting of the special interest group on discourse and dialogue (SIGDIAL)*, 2014, pp. 263–272.

[4] C. Tang, W. Yu, G. Sun, X. Chen, T. Tan, W. Li, L. Lu, M. Zejun, and C. Zhang, "SALMONN: Towards Generic Hearing Abilities for Large Language Models," in *Proc. 12th Intl. Conf. on Learning Representations*, 2024.

[5] J. Wu, Y. Gaur, Z. Chen, L. Zhou, Y. Zhu, T. Wang, J. Li, S. Liu, B. Ren, L. Liu *et al.*, "On decoder-only architecture for speech-to-text and large language model integration," in *Proc. IEEE ASRU*, 2023.

[6] Z. Ma, G. Yang, Y. Yang, Z. Gao, J. Wang, Z. Du, F. Yu, Q. Chen, S. Zheng, S. Zhang *et al.*, "An embarrassingly simple approach for LLM with strong ASR capacity," preprint arXiv:2402.08846, 2024.

[7] W. Yu, C. Tang, G. Sun, X. Chen, T. Tan, W. Li, L. Lu, Z. Ma, and C. Zhang, "Connecting speech encoder and large language model for ASR," in *Proc. IEEE ICASSP*, 2024.

[8] S. Kumar, I. Thorbecke, S. Burdisso, E. Villatoro-Tello, M. K E, K. Hacioğlu, P. Rangappa, P. Motlicek, A. Ganapathiraju, and A. Stolcke, "Performance evaluation of SLAM-ASR: The good, the bad, the ugly, and the way forward," in *Proc. IEEE ICASSP Workshop*, 2025.

[9] M. Yang, S.-J. Chen, J. Xie, and H. L. John Hansen, "Bridging the modality gap: Softly discretizing audio representation for LLM-based automatic speech recognition," in *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2025, pp. 1–7.

[10] Y. Fang, J. Peng, X. Li, Y. Xi, C. Zhang, G. Zhong, and K. Yu, "Low-resource domain adaptation for speech LLMs via text-only fine-tuning," preprint arXiv:2506.05671, 2025.

[11] Šimon Sedláček, B. Yusuf, J. Švec, P. Hegde, S. Kesiraju, O. Plchot, and J. Černocký, "Approaching Dialogue State Tracking via Aligning Speech Encoders and LLMs," in *Proc. Interspeech*, 2025, pp. 1748–1752.

[12] M. Wang, W. Han, I. Shafran, Z. Wu, C.-C. Chiu, Y. Cao, N. Chen, Y. Zhang, H. Soltau, P. K. Rubenstein *et al.*, "SLM: Bridge the thin gap between speech and text foundation models," in *Proc. IEEE ASRU*, 2023.

[13] G. Yang, Z. Ma, F. Yu, Z. Gao, S. Zhang, and X. Chen, "MaLa-ASR: Multimedia-assisted LLM-based ASR," in *Proc. Interspeech*, 2024.

[14] A. Carofilis, S. Burdisso, E. Villatoro-Tello, S. Kumar, K. Hacioglu, S. Madikeri, P. Rangappa, M. K. E, P. Motlicek, S. Venkatesan, and A. Stolcke, "Text-only adaptation in LLM-based ASR through text denoising," 2026. [Online]. Available: https://arxiv.org/abs/2601.20900

[15] S. Burdisso, E. Villatoro-Tello, S. Kumar, S. Madikeri, A. Carofilis, P. Rangappa, M. K. E, K. Hacioglu, P. Motlicek, and A. Stolcke, "Reducing prompt sensitivity in LLM-based speech recognition through learnable projection," in *Proc. IEEE ICASSP*, 2026.

[16] S. Burdisso, E. Villatoro-Tello, T. Bañeras-Roux, S. Kumar, S. Madikeri, M. K E, P. Motlicek, and A. Stolcke, "Avoiding catastrophic forgetting in text-only adaptation of LLM-based ASR via multi-view text denoising," in *Proc. Interspeech*, 2026.

[17] T. Bañeras-Roux, S. Burdisso, E. Villatoro-Tello, D. Sánchez-Cortés, S. Liu, S. Baroudi, S. Kumar, H. Watawana, M. K E, K. Hacioglu, P. Motlicek, and A. Stolcke, "Closing the speech-text gap with limited audio for effective domain adaptation in LLM-Based ASR," in *Proc. Interspeech*, 2026.

[18] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang, "Lost in the middle: How language models use long contexts," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 157–173, 2024. [Online]. Available: https://aclanthology.org/2024.tacl-1.9/

[19] L. Druart, V. Vielzeuf, and Y. Estève, "Is one brick enough to break the wall of spoken dialogue state tracking?" *arXiv preprint arXiv:2311.04923*, 2023.

[20] N. E. Ghazal, A. Caubrière, and V. Vielzeuf, "The speech-LLM takes it all: A truly fully end-to-end spoken dialogue state tracking approach," *arXiv preprint arXiv:2510.09424*, 2025.

[21] A. Shamsian, A. Navon, N. Glazer, G. Hetz, and J. Keshet, "Keyword-Guided Adaptation of Automatic Speech Recognition," in *Proc. Interspeech*, 2024, pp. 732–736.

[22] Y. Sudo, Y. Fukumoto, M. Shakeel, Y. Peng, and S. Watanabe, "Contextualized automatic speech recognition with dynamic vocabulary," in *2024 IEEE Spoken Language Technology Workshop (SLT)*, 2024, pp. 78–85.

[23] I. Thorbecke, E. Villatoro-Tello, J. P. Zuluaga, S. Kumar, S. Burdisso, P. Rangappa, A. Carofilis, S. Madikeri, P. Motlicek, K. Pandia, K. Hacioğlu, and A. Stolcke, "Unifying global and near-context biasing in a single trie pass," in *Text, Speech, and Dialogue*, K. Ekštein, M. Konopík, O. Pražák, and F. Pártl, Eds. Cham: Springer Nature Switzerland, 2026, pp. 170–181.

[24] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[25] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan *et al.*, "The Llama 3 herd of models," *CoRR*, vol. abs/2407.21783, 2024.

[26] S. Burdisso, S. Madikeri, and P. Motlicek, "Dialog2Flow: Pre-training soft-contrastive action-driven sentence embeddings for automatic dialog flow extraction," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 5421–5440. [Online]. Available: https://aclanthology.org/2024.emnlp-main.310/

[27] Gemma Team, "Gemma 3," https://goo.gle/Gemma3Report, Kaggle, Tech. Rep., 2025.

[28] M. Abdin, J. Aneja, H. Behl, S. Bubeck, R. Eldan, S. Gunasekar, M. Harrison, R. J. Hewett, M. Javaheripi, P. Kauffmann, J. R. Lee, Y. T. Lee, Y. Li, W. Liu, C. C. T. Mendes, A. Nguyen, E. Price, G. de Rosa, O. Saarikivi, A. Salim, S. Shah, X. Wang, R. Ward, Y. Wu, D. Yu, C. Zhang, and Y. Zhang, "Phi-4 technical report," 2024. [Online]. Available: https://arxiv.org/abs/2412.08905

[29] D. Guo, D. Yang, H. Zhang, and et al., "Deepseek-R1 incentivizes reasoning in LLMs through reinforcement learning," *Nature*, vol. 645, no. 8081, p. 633–638, Sep. 2025. [Online]. Available: http://dx.doi.org/10.1038/s41586-025-09422-z

[30] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, C. Zheng, D. Liu, F. Zhou, F. Huang, F. Hu, H. Ge, H. Wei, H. Lin, J. Tang, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Zhou, J. Lin, K. Dang, K. Bao, K. Yang, L. Yu, L. Deng, M. Li, M. Xue, M. Li, P. Zhang, P. Wang, Q. Zhu, R. Men, R. Gao, S. Liu, S. Luo, T. Li, T. Tang, W. Yin, X. Ren, X. Wang, X. Zhang, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Zhang, Y. Wan, Y. Liu, Z. Wang, Z. Cui, Z. Zhang, Z. Zhou, and Z. Qiu, "Qwen3 technical report," 2025. [Online]. Available: https://arxiv.org/abs/2505.09388

[31] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *Proc. Intl. Conf. on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

[32] J. Li, A. Sun, J. Han, and C. Li, "A survey on deep learning for named entity recognition : Extended abstract," in *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, 2023, pp. 3817–3818.