Model	Method	Code		Database		Actions		Math		Average	
		$\bar{P}$	R	$\bar{P}$	R	$\bar{P}$	R	$\bar{P}$	R	$\bar{P}$	R
GPT-4o-mini	Full	74.2	76.0	92.5	93.5	93.7	92.4	87.2	70.9	86.9	83.2
	Sharded	51.4	57.0	52.5	54.2	45.5	60.0	64.9	45.6	53.6	54.2
	w/ Ours	66.9	63.6	65.3	59.8	85.7	81.2	77.7	70.4	73.9	68.8
	Gain	+15.5	+6.6	+12.8	+5.6	+40.2	+21.2	+12.8	+24.8	+20.3	+14.6
GPT-5.2	Full	83.2	84.0	96.3	95.9	90.2	93.2	94.5	89.4	92.7	85.4
	Sharded	39.4	44.0	49.4	48.0	35.6	46.6	69.6	48.6	48.5	46.8
	w/ Ours	69.1	70.1	64.5	56.7	76.2	65.2	80.6	62.0	72.6	63.5
	Gain	+29.7	+26.1	+15.1	+8.7	+40.6	+18.6	+11.0	+13.4	+24.1	+16.7
DeepSeek-v3.2-Thinking	Full	98.3	95.9	94.4	88.8	92.2	88.6	94.0	81.6	94.7	88.7
	Sharded	78.4	65.7	43.6	54.2	42.3	48.6	78.8	56.3	60.8	56.2
	w/ Ours	86.1	84.8	67.3	55.9	88.0	71.6	86.3	67.3	81.9	69.9
	Gain	+7.7	+19.1	+23.7	+1.7	+45.7	+23.0	+7.5	+11.0	+21.1	+13.7

Table 1: **Main Results.** Comparison of average performance ($\bar{P}$) and reliability (R) across three LLM backbones. All experiments are averaged over 5 runs. R is calculated as the mean of $1 - (S_{\max} - S_{\min})$ across all instances. *Full* represents ideal instructions, while *Sharded* represents ambiguous user inputs. *w/ Ours* denotes the performance when using our Experience-Driven Mediator. The best improvements are highlighted in green.

sort to ungrounded guessing to fill the information gap, resulting in performance collapses similar to those observed in smaller baselines. Crucially, even reasoning-enhanced models like **DeepSeek-v3.2-Thinking** fail to overcome this hurdle. Contrary to the expectation that intrinsic Chain-of-Thought (CoT) might infer missing context, DeepSeek's performance in the *Sharded* setting remains severely compromised. This result exposes a critical limitation: reasoning capabilities are ineffective against information ambiguity. While CoT excels at logical deduction when premises are clear, it cannot "reason" its way out of an information vacuum. Without the explicit context reconstruction provided by our Mediator, the model is forced into ungrounded speculation, regardless of its reasoning depth.

Efficacy of Mediator. Implementing our proposed Mediator yields substantial and consistent improvements across all backbones and domains. On average, our method recovers performance by approximately 20% in $\bar{P}$ and 15% points in R. Notably, even for the reasoning-enhanced DeepSeek-v3.2-Thinking, our approach secures a 21.1% increase, demonstrating that the Mediator provides value complementary to advanced internal reasoning. By reconstructing ambiguous contexts into explicit, self-contained instructions ($\hat{U}$), it collapses the uncertainty space for each individual instance. This ensures that the Assistant receives a low-entropy input, leading to consistently high-quality execution regardless of the random seed, effectively converting a probabilistic guessing game into a deterministic reasoning task.

Correlation between Reliability and Performance. We observe substantial improvements in both Performance ($\bar{P}$) and Reliability (R). This parallel growth demonstrate that multi-turn generation is not inherently prone to high variance. Instead, the data suggests that the previously observed instability stemmed primarily from intent ambiguity rather than intrinsic model stochasticity. By effectively aligning the Assistant with the intended constraints, our Mediator proves that reliability can be enhanced just as effectively as performance. This confirms that once the intent is grounded, the model's generation process transitions from a volatile guessing game to a robust and reproducible execution.

5.2 Intent Alignment vs. Factual Memory

To identify the root cause of the performance degradation in multi-turn dialogues, we compare our approach against two representative context utilization strategies (Table 2): a naive summarization method (**w/ Sum**) and a mainstream RAG-based memory framework **Mem0** to persist user facts.

Memory is not Understanding. A prevailing hypothesis suggests that models fail because they forget constraints or details from previous turns. If this were true, Mem0 should significantly restore performance by retrieving relevant historical facts. However, the results in Table 2 contradict this assumption. **w/ Mem0** yields only marginal gains over the baseline (53.6% → 56.5%), and **w/ Sum**

Method	C	D	A	M	Avg.
Full	74.2	92.5	93.7	87.2	86.9
Sharded	51.4	52.5	45.5	64.9	53.6
w/ Sum	57.1	44.9	54.6	62.0	54.7
w/ Mem0	52.8	56.9	49.4	66.9	56.5
w/ Ours	66.9	65.3	85.7	77.7	73.9

Table 2: **Comparative Analysis.** We compare our method against Summarization (Sum) and a RAG-based memory framework (Mem0) on GPT-4o-mini. The marginal gains of Mem0 indicate that simply recalling factual context is insufficient for multi-turn robustness. In contrast, our method succeeds by explicitly reconstructing the user's intent.

shows high variance, even degrading performance in some domains.

This finding highlights a critical distinction: retrieving context is not equivalent to resolving intent. In the *Sharded* setting, the model usually has access to the raw information (facts); the failure stems from its inability to determine how the user intends to apply those facts to the current ambiguous request. For example, Mem0 might successfully recall that "the user wants a Python script," but it fails to clarify which specific logic from the conversation history should be applied now.

In contrast, by explicitly synthesizing an unambiguous instruction, our Mediator bridges the gap between raw context and actionable intent. This results in a decisive improvement, outperforming Mem0 by 17.4%. This confirms that simply "remembering" is insufficient to solve the LiC problem, the system should explicitly align with the user's intent.

Why Refiner? We further investigate the necessity of the Refiner module by comparing it with an In-Context Learning (ICL) baseline, which embeds raw contrastive pairs directly into the Mediator's prompt, and the Oracle baseline. As shown in Figure 4, our method significantly outperforms the Oracle while incurring only a negligible increase in token usage. In contrast, direct ICL achieves comparable accuracy to our method, but it incurs a 3.6× increase in token consumption due to verbose context. This demonstrates that Refiner distills verbose interaction logs into concise guidelines, ensuring practical inference efficiency without compromising performance.

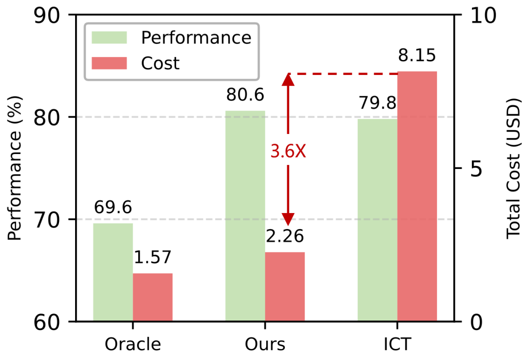

Figure 4: **Comparison with In-Context Learning.** We compare our method against the Oracle baseline and a Direct ICL approach. Our method delivers a substantial performance boost over the Oracle with only a marginal cost increase. While ICL achieves comparable accuracy to Ours, it consumes 3.6× more tokens.

6 Conclusions

In this paper, we revisited the "Lost in Conversation" (LiC) phenomenon, identifying it not as a fundamental deficit in model capability, but as an intent alignment gap between user expression and model understanding. We theoretically demonstrated that simply scaling up LLMs is insufficient to resolve this issue, necessitating an architectural intervention. To this end, we proposed a Mediator-Assistant framework equipped with an experience Refiner. By distilling historical contrastive trajectories into concise guidelines, our method enables the Mediator to explicate ambiguous inputs into explicit instructions, then significantly mitigates performance degradation in multi-turn interactions.

7 Limitations

Constrained by the limited scale of existing benchmarks, our current Refiner operates in a few-shot, non-parametric manner, extracting explicit and heuristic-level guidelines. While this design ensures efficiency, it captures only coarse-grained interaction patterns. Future work could leverage larger-scale datasets to transition towards parameterized training, enabling the Mediator to internalize more nuanced alignment strategies via fine-tuning rather than relying solely on in-context summaries. Furthermore, the multi-turn settings in current benchmarks exhibit relatively homogeneous user logic. Developing more comprehensive benchmarks that mirror complex user behaviors remains a critical direction to further validate and evolve our framework.

References

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and 1 others. 2024. Mt-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7421–7454.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. 2025. Mem0: Building production-ready ai agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*.

DeepSeek-AI. 2025. Deepseek-v3.2-exp: Boosting long-context efficiency with deepseek sparse attention.

Christine Herlihy, Jennifer Neville, Tobias Schnabel, and Adith Swaminathan. 2024. On overcoming miscalibrated conversational priors in llm-based chatbots. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, pages 1599–1620.

Jeff Huang and Efthimis N Efthimiadis. 2009. Analyzing and evaluating query reformulation strategies in web search logs. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 77–86.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. 2025. Llms get lost in multi-turn conversation. *arXiv preprint arXiv:2505.06120*.

Ming Li. 2025. Verifiable accuracy and abstention rewards in curriculum rl to alleviate lost-in-conversation. *arXiv preprint arXiv:2510.18731*.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating very long-term conversational memory of llm agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13851–13870.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. Ambigqa: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797.

Daan Odijk, Ryen W White, Ahmed Hassan Awadallah, and Susan T Dumais. 2015. Struggling and success in web search. In *Proceedings of the 24th ACM international on conference on information and knowledge management*, pages 1551–1560.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Lingling Xu, Haoran Xie, Si-Zhao Joe Qin, Xiaohui Tao, and Fu Lee Wang. 2023. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. *Preprint*, arXiv:2312.12148.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. 2025. A-mem: Agentic memory for llm agents. *arXiv preprint arXiv:2502.12110*.

J Diego Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. 2023. Why johnny can't prompt: how non-ai experts try (and fail) to design llm prompts. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pages 1–21.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19724–19731.

Wanjun Zhong, Duyu Tang, Jiahai Wang, Jian Yin, and Nan Duan. 2021. Useradapter: Few-shot user learning in sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1484–1488.

George Kingsley Zipf. 1949. *Human behavior and the principle of least effort: An introduction to human ecology*. Addison-Wesley Press.

A Benchmark Details

A.1 Dataset Details

We utilize the pre-constructed conversational data provided by Laban et al. (2025). The original benchmark encompassed six diverse tasks, which can be categorized into two distinct types based on their evaluation criteria. The first category, Binary Correctness Tasks, consists of Code, Database, Actions, and Math. For these tasks, success is defined by strict execution accuracy or exact constraint satisfaction. The second category, Refinement Tasks, comprises Data-to-text and Summarization, where