A PERSONALIZED REAL-TIME PROACTIVE VOICE MEMORY ASSISTANT

Hao Zhou[1,2,†], Md Mahbubur Rahman[1], Simon A. Lee[1,3,†], Baiying Lu[1,4,†], Juhyeon Lee[1,5,†], Cyrus Tanade[1], Megha Thukral[1,6,†], Md. Sazzad Hissain Khan[7], Samsad Ul Islam[7], Subramaniam Venkatraman[1], Sharanya Arcot Desai[1]

[1]Samsung Research America, [2]The Pennsylvania State University, [3]University of California, Los Angeles, [4]Dartmouth, [5]University of Massachusetts Amherst, [6]Georgia Institute of Technology, [7]Samsung Research Bangladesh

ABSTRACT

Timely recall of details is crucial for collaboration, decision-making, and planning, especially for people with dementia (57 million globally) or memory impairments (10% of adults over 65 have dementia and 22% have mild impairment in U.S.). Existing assistive technologies often rely on manual queries or lack awareness of conversational ownership, leading to irrelevant recall and privacy concerns, and limiting their effectiveness in dynamic, multi-party settings. In this paper, we introduce *MemoryAids*, a proactive voice memory assistant that seamlessly operates during live conversations. *MemoryAids* only focuses on owner speech, detects missing details in real time, and provides concise summaries by proposing a low-latency owner detection module and leveraging in-context learning. Our evaluations show accurate owner detection (a recall of 90.7%), recall moment detection (92.7% accuracy with 5.8% word error rate), and sub-second latency, highlighting its potential to benefit people with memory impairments.

***Index Terms*—** proactive memory aids, speaker identification, in-context learning LLM

1. INTRODUCTION

Timely recall of conversational details is essential for effective collaboration, decision-making, and planning. When people forget or miss key information (e.g., due to interference by others, lack of sleep, excessive stress [1]), they are forced to pause and consult notes or digital records, which disrupts the natural flow of dialogue. This problem is even worse for people who live with dementia (57 million globally [2]) or memory impairments (10% of adults over 65 have dementia and 22% have mild impairment [3, 4] in the U.S.).

Prior Works. With the advances of automatic speech recognition, natural language processing, and retrieval augmented generation, several systems have been proposed to assist users in accessing information. MemPal [5] supports query-based memory cues given visual cues from a wearable camera. LLAMAPIE [6] supports proactive assistance by providing related information during dialogue, but it relies on pre-summarized events as input rather than operating on raw conversations. Moreover, it accepts voice from all participants without distinguishing the intended owner, thereby providing no privacy protection and risking the leakage of sensitive information.

Requirements. For a memory assistant to be practical and effective, we aim to satisfy two requirements. (1) *Owner-Awareness:* The assistant should attend only to the owner's speech during conversations. By ignoring all non-owner voices, the system prevents outsiders from triggering memory retrieval or responses, thereby ensuring that private information tied to the owner cannot be accessed by others. In addition, owner-awareness strengthens contextual grounding in multi-turn dialogues where speaker identity may be implicit. Capturing what the owner has previously said is also important, especially for people with memory impairments, such as dementia or Alzheimer's disease, where recalling recent conversations is a core difficulty [7]. (2) *Minimal User Intervention:* The assistant proactively anticipates the owner's information needs and delivers concise, relevant cues in real time, avoiding explicit queries or manual interaction to maintain the natural conversation flow and improve communication efficiency.

Solutions. In this paper, we propose *MemoryAids*, a voice memory assistant that operates seamlessly during live conversations. In contrast to prior works, *MemoryAids* proactively summarizes ongoing dialogues, detects moments where forgotten or missing details hinder communication (*recall moments*), and delivers the relevant information in real time. To realize it, we propose

[†]WORK DONE DURING INTERNSHIP AT SAMSUNG RESEARCH AMERICA

the following modules. (1) *Low-Latency Owner Voice Detection via Preamble:* Inspired by synchronization preambles in modern wireless communication [8], we introduce a one-time initialization phase where the owner provides a short speech preamble (≈3 seconds). This preamble serves as a lightweight reference for matching incoming speech segments, enabling accurate filtering of non-owner voices during live conversations. Moreover, the stored preamble is iteratively refined over time as more conversations occur, further improving robustness in owner-specific voice detection. This allows *MemoryAids* to obtain the awareness of its owner, thus providing owner-centric assistance. (2) *Recall Moment Detection via In-Context Learning:* We use LLMs using in-context learning [9, 10, 11], where example dialogues of *recall moments* are provided in the prompt at inference time. The generalizability of LLMs allows *MemoryAids* to generalize from examples, and allows owner to customize their own recall moments without retraining. When such recall moments are detected, the system proactively provides the relevant details, eliminating the need for manual query.

Evaluation. We validate *MemoryAids* across multiple dimensions. Our system achieves a word error rate of 5.8% and recall moment detection with 92.7% accuracy on the LLAMAPIE dataset [6]. It runs with a 927 ms latency on average per inference and with a recall of 90.7% for owner detection. We summarize our contributions:

- To the best of our knowledge, *MemoryAids* is the first real-time memory assistant that proactively decides when to provide missing information exclusively for the owner during conversations.
- We provide a comprehensive evaluation demonstrating the effectiveness, efficiency, and usability of *MemoryAids*, with implications for populations with hearing or memory impairments.

2. RELATED WORKS

Table 1: Comparisons of Memory Assistants.

System	Proactive Assistant	Owner-awareness Privacy	In-conversation Summarization
Memoro [12]	✔	✖	✖
MemPal [5]	✖	✖	✖
Mirai [13]	✔	✖	✖
LLAMAPIE [6]	✔	✖	✖
MemoryAids (Ours)	✔	✔	✔

Memory Assistants: Several memory assistants have been proposed recently [12, 5, 13, 6]. Yet, these systems lack owner-awareness and do not explicitly address privacy concerns or owner-centric in-conversation summarization (Table 1). In contrast, *MemoryAids* is the first memory assistant that protects privacy through owner-aware voice detection and supports owner-centric in-conversation summarization.

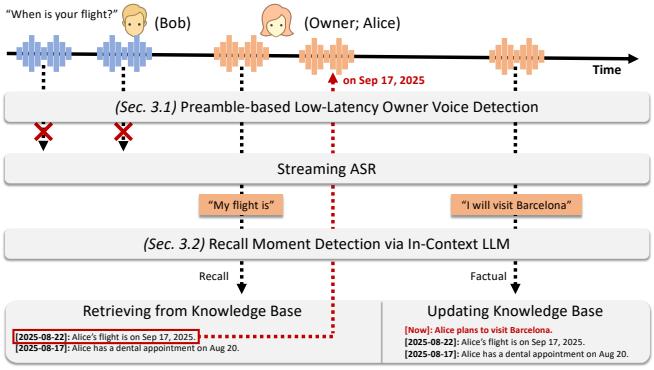

Fig. 1: Overall pipeline of *MemoryAids*.

conversation summarization.

LLM In-Context Learning: LLMs have demonstrated remarkable capabilities with in-context learning [9, 10, 11, 14], where they adapt at inference time using conversational histories or annotated examples. While these techniques have been applied to general dialogue or QA systems, *MemoryAids* incorporates in-context learning for proactive memory assistance with owner-awareness, ensuring personalization and preserving privacy.

3. *MemoryAids*

Fig. 1 depicts the overall pipeline of *MemoryAids*. Incoming audio from two or more speakers (e.g., from earbuds) is first processed by a low-latency owner voice detector to ensure that only the owner's speech is retained. The retained speech is then transcribed in real time (we used a variant of Whisper [15]), after which an LLM with in-context learning determines whether each spoken sentence represents a *recall moment* (i.e., where users have forgotten any specific details) that requires memory assistance or a factual update. Depending on this decision, the system either retrieves missing details from the knowledge base or updates it with newly gathered information, providing proactive and privacy-preserving memory cues during conversations. The cues are delivered as concise text prompts on the owner's phone screen or smart glasses, providing a practical and unobtrusive feedback channel, in contrast to audio playback that could interfere with the ongoing conversation. Next, we detail our key designs.

3.1. Preamble Based Low-Latency Owner Detection

To achieve low-latency operation, *MemoryAids* proposes a lightweight preamble-based owner identification module[1]. At initialization (one-time), the owner records a short (≈3s) speech preamble, which is converted into an embedding $\mathbf{e}_0 \in \mathbb{R}^d$ using Pyannote.audio [18]

[1]We omit the usage of speaker diarization algorithms [16, 17] since they generally provide generic speaker labels.

and stored as a reference. For each incoming speech segment x_t, we compute an embedding $\mathbf{e}_t \in \mathbb{R}^d$ and measure cosine similarity against the current reference: $S(\mathbf{e}_t, \mathbf{e}_0) = \frac{\mathbf{e}_t \cdot \mathbf{e}_0}{\|\mathbf{e}_t\| \|\mathbf{e}_0\|}$. If the similarity, $S(\mathbf{e}_t, \mathbf{e}_0)$, is larger than 0.2 (empirically set), we identify the owner and add the embedding to a candidate set $\mathcal{E} = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$. To refine the reference, we select the most representative candidate that has the highest correlation (>0.8) with all the other candidates: $\mathbf{e}_0 \leftarrow \arg\max_{\mathbf{e}_i \in \mathcal{E}} \left(\frac{1}{|\mathcal{E}|-1} \sum_{\substack{j=1 \\ j \neq i}}^{|\mathcal{E}|} \frac{\mathbf{e}_i \cdot \mathbf{e}_j}{\|\mathbf{e}_i\| \|\mathbf{e}_j\|} \right)$.

Benefits: This preamble-based approach enables accurate and low-latency owner identification. By discarding non-owner speech, it preserves the privacy as the assistant stops responding to other parties and ensures that retrieval and summarization remain strictly owner-specific. The refinement mechanism further improves robustness as usage continues, allowing *MemoryAids* to adapt to natural variations in the owner's voice while maintaining reliability.

3.2. Recall Moment Detection via In-Context LLM

Listing 1: Example 1

```
Input:
Alice: My flight is.
Prior context:
[2025-08-22]: Alice's flight is on Sep
       17, 2025.
Output:
{ "speaker": "Alice",
  "type": "recall",
  "fact": "on Sep 17"}
```

Listing 2: Example 2

```
Input:
Alice: I will visit Barcelona.
Prior context:
[2025-08-22]: Alice's flight is on Sep
       17, 2025.
Output:
{ "speaker": "Alice",
  "type": "factual",
  "fact": "Alice will visit Barcelona"}
```

Fig. 2: Embedding examples to *MemoryAids*'s prompt.

While owner detection ensures that only the user's speech is processed, the assistant must also determine when to intervene. Inspired by previous works [10, 11], *MemoryAids* employs LLMs with in-context learning to perform this task. Specifically, example dialogues (Fig. 2) annotated with "recall" or "factual" are embedded directly in our prompt at inference. Given a live transcript stream, the LLM generalizes from these examples to identify 1) moments where forgotten or missing information is likely to pause communication, or 2) moments where the new knowledge from the owner is gathered. Once detected, the LLM generates a concise cue based on the relevance of the query in the database via similarity checking (we use Sentence BERT [19] and cosine similarity), or the system updates the database based on the detected new knowledge about the owner.

Benefits: By unifying recall moments detection, factual information summarization, and content generation in a single model, our design eliminates the complexity and error propagation associated with multi-stage pipelines [6]. More importantly, it also enables rapid adaptation to personalized dialogue styles without retraining, since additional examples can simply be included in the prompt, whereas multi-stage approaches require retraining classifiers for each new style.

4. PERFORMANCE EVALUATION

We validate *MemoryAids* by answering questions below.

4.1. Can *MemoryAids* detect the owner's voice?

Accurate identification of the owner's voice is a fundamental requirement for *MemoryAids*. Without this capability, the system processes speech from other participants, raising privacy concerns and leading to irrelevant memory retrieval. To evaluate the effectiveness of our preamble-based module (Sec. 4.1), we conducted a controlled usability study with 12 healthy participants (Our study was IRB exempt since only anonymous usability feedback was collected. Participants' voices were temporarily processed by the system but not recorded or stored). Each participant engaged in a scripted dialogue that included numerous factual details, such as flight timings, locations, conferences, hotels, sports schedules, and restaurants. The participant and the study member take turns to read the script. Each turn contains 24.3 words on average, and 15 turns for the participant. At the end of the conversation, the participant was asked to answer questions about the conversation details. For every sentence spoken by the participant (owner), we logged a) whether *MemoryAids* successfully recognized owner's voice or b) the number of times that the other speaker was wrongly recognized as the owner, potentially leaking private information of the owner. Fig. 3a shows *MemoryAids* detects owner's voice with a recall of 90.7%, and Fig. 3b depicts that *MemoryAids* protects owner's privacy by preventing the other speaker from being recognized as the owner. These results validate the design choice of using a lightweight embedding similarity approach instead of full diarization pipelines, which only provides generic speaker labels, and confirm that *MemoryAids* can reliably suppress non-owner speech in real time, thereby preserving privacy by preventing retrieval from other speakers.

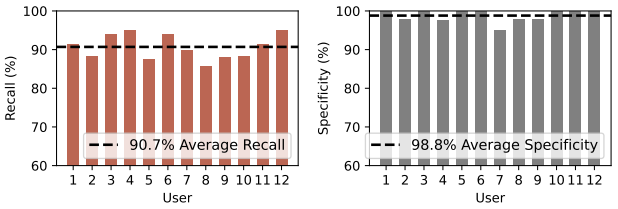

Fig. 3: *MemoryAids* (a) accurately detects owner's voices, and (b) prevents the other speaker being recognized as the owner, protecting the owner's privacy.

4.2. Can *MemoryAids* identity sentence types and how accurate are the responses?

We then evaluate the effectiveness of the proposed *Recall Moment Detection via In-Context LLM* (Sec. 4.2). In the same study, we logged the detection accuracy of two

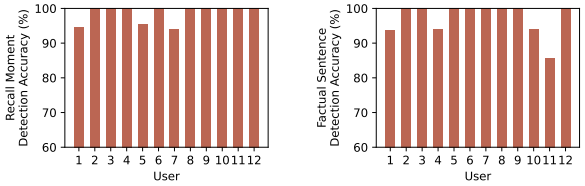

Fig. 4: *MemoryAids* accurately detects different types of sentences during live conversations.

types of sentences: a) *recall moment*, where the system proactively returns information to the user, and b) *factual* sentence, where the system updates its knowledge base about the user. On average, each participant contributed around 20 recall and factual sentences. Fig. 4 shows the overall detection accuracy, demonstrating the effectiveness of in-context learning. We note that for the scripted sentences, the accuracy of recall moment detection is not the same for all users. This is because of the errors from speech recognition, where some sentences fail to be fully recognized due to accents.

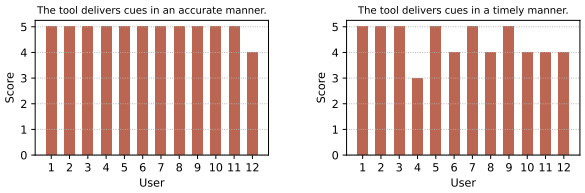

Fig. 5: Participants' feedback on *MemoryAids*.

We also asked participants to evaluate the subjective aspects of *MemoryAids* along two dimensions: 1) whether the tool's assistance was delivered accurately (**Accuracy**), and 2) whether it was delivered in a timely manner (**Timeliness**). Participants rated these statements on a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree). Fig. 5 summarizes the results, showing consistently high satisfaction with both accuracy and timeliness. These results highlight *MemoryAids*'s effectiveness in supporting real-time conversations and motivate future extensions to populations with memory impairments and cognitive decline.

4.3. Comparing against existing work

We employ LLAMAPIE dataset [6] to compare (1) if the system could detect the moments where the owner needs cues and (2) how accurate the responses are, i.e., word error rate [20] between ground truth and responses. The dataset contains ≈3128 conversations, and each speaker has ≈22 words per turn. Table 2 depicts *MemoryAids*[2] achieves comparable accuracy to its counterpart, where a specialized model is to detect the recall moments and another model for generating responses.

4.4. Can *MemoryAids* run in real time?

Table 3 reports the end-to-end latency of *MemoryAids* and the latency breakdowns across the three major components, owner detection, transcription, and retrieval-augmented generation. Our owner detection module operates with low latency (≈53 ms), confirming its suitability for real-time filtering. Streaming ASR introduces latency (≈80 ms). By contrast, the RAG module accounts for the largest processing time (≈793 ms). This includes both the sentence embedding similarity search over the knowledge base and the time needed to generate responses from Gemini 2.5 Flash-Lite. Similar to prior work [6], we offload all computations to a server with a 4-core AMD CPU and an NVIDIA L40S. Overall, the system achieves a sub-second end-to-end latency, which is acceptable for live conversational assistance [6].

[2]We are restricted to use Gemini 2.5 [21].

Table 2: *MemoryAids* is compared against baselines.

System	Detection Accuracy (↑)	Response Word Error Rate (↓)
LLAMAPIE [6]	93.5%	7.8%
MemoryAids$_{\text{-Gemini-2.5-flash}}$	90.8%	6.3%
MemoryAids$_{\text{-Gemini-2.5-pro}}$	99.1%	5.9%
MemoryAids$_{\text{-Gemini-2.5-flash-lite}}$	92.7%	5.8%

Table 3: Latency breakdown. Values in milliseconds (ms).

Component	Mean ± Std	Median	95%-tile
Owner Detection	53.2 ± 14.7	51.0	78.7
Transcription (ASR)	79.8 ± 20.2	80.2	112.9
RAG (Embed+LLM)	793.9 ± 82.6	793.6	929.5
Total	926.9 ± 87.7	926.4	1070.2

5. DISCUSSION

Future work will expand the study to people with memory impairment, hearing loss, and the elderly population in free-living settings, as these groups would benefit most from proactive conversational support in their daily lives. We also envision seamless integration of *MemoryAids* into everyday devices such as earbuds, smartphones, smart glasses, or smartrings [22, 23] to understand the users' intents for a comprehensively proactive assistant and enhance usability in daily life. From a privacy perspective, we plan to secure the entire pipeline by encrypting both transcriptions and embeddings (e.g., via homomorphic encryption [24]), coupled with locally deployed LLMs. This would ensure that user data remains fully private, with encrypted representations only decodable on the user's device, while still supporting retrieval and real-time recall. In addition, because human voices naturally evolve due to aging, illness, or environments, our preamble-based owner voice detection module will be extended to continuously adapt to these changes over time. By refining its embeddings with the most recent speech samples, the system can maintain robustness without repeated manual calibration. Together, these directions chart the path toward a highly private, adaptive, and practical memory assistant.

6. REFERENCES

[1] OT Wolf, P Atsak, et al., "Stress and memory: a selective review on recent developments in the understanding of stress hormone effects on memory and their clinical relevance," *Journal of neuroendocrinology*, 2016.

[2] World Health Organization, "Dementia fact sheet," 2021.

[3] Brenda L Plassman et al., "Prevalence of dementia in the united states: the aging, demographics, and memory study," *Neuroepidemiology*, 2007.

[4] Alzheimer's Association, "2024 alzheimer's disease facts and figures," 2024.

[5] Natasha Maniar, Samantha WT Chan, Wazeer Zulfikar, Scott Ren, Christine Xu, and Pattie Maes, "Mempal: Leveraging multimodal ai and llms for voice-activated object retrieval in homes of older adults," in *IUI*, 2025.

[6] Tuochao Chen et al., "Llamapie: Proactive in-ear conversation assistants," *arXiv preprint arXiv:2505.04066*, 2025.

[7] Eric Grandmaison and Martine Simard, "A critical review of memory stimulation programs in alzheimer's disease," *The Journal of neuropsychiatry and clinical neurosciences*, 2003.

[8] Eleftherios Kofidis, Dimitrios Katselis, Athanasios Rontogiannis, and Sergios Theodoridis, "Preamble-based channel estimation in ofdm/oqam systems: A review," *Signal processing*, 2013.

[9] Tom Brown et al., "Language models are few-shot learners," *NeurIPS*, 2020.

[10] Chen Cheng et al., "Exploring the robustness of in-context learning with noisy labels," in *ICASSP*. IEEE, 2025.

[11] Zhehuai Chen, He Huang, Andrei Andrusenko, Oleksii Hrinchuk, Krishna C Puvvada, Jason Li, Subhankar Ghosh, Jagadeesh Balam, and Boris Ginsburg, "Salm: Speech-augmented language model with in-context learning for speech recognition and translation," in *ICASSP*. IEEE, 2024.

[12] Wazeer Deen Zulfikar, Samantha Chan, and Pattie Maes, "Memoro: Using large language models to realize a concise interface for real-time memory augmentation," in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1–18.

[13] Cathy Mengying Fang, Yasith Samaradivakara, Pattie Maes, and Suranga Nanayakkara, "Mirai: A wearable proactive ai" inner-voice" for contextual nudging," in *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 2025, pp. 1–9.

[14] Yifan Wang et al., "Hint-enhanced in-context learning wakes large language models up for knowledge-intensive tasks," in *ICASSP*. IEEE, 2024.

[15] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28492–28518.

[16] Aonan Zhang et al., "Fully supervised speaker diarization," in *ICASSP*. IEEE, 2019.

[17] Quan Wang et al., "Speaker diarization with lstm," in *ICASSP*. IEEE, 2018.

[18] Hervé Bredin et al., "Pyannote. audio: neural building blocks for speaker diarization," in *ICASSP*. IEEE, 2020.

[19] Nils Reimers and Iryna Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[20] Thilo von Neumann et al., "On word error rate definitions and their efficient computation for multi-speaker speech recognition systems," in *ICASSP*. IEEE, 2023, pp. 1–5.

[21] Gheorghe Comanici et al., "Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities," *arXiv preprint arXiv:2507.06261*, 2025.

[22] Hao Zhou, Taiting Lu, Yilin Liu, Shijia Zhang, and Mahanth Gowda, "Learning on the rings: Self-supervised 3d finger motion tracking using wearable sensors," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 2, pp. 1–31, 2022.

[23] Hao Zhou, Taiting Lu, Kenneth DeHaan, and Mahanth Gowda, "Aslring: American sign language recognition with meta-learning on wearables," 2024.

[24] Vele Tosevski and Glenn Gulak, "Large-scale recurrent neural networks with fully homomorphic encryption for privacy-enhanced speaker identification," in *ICASSP*. IEEE, 2025.