Retention-Augmented Voice Assistant: A Lightweight Architecture for Stateful Interaction with Comprehensive Evaluation and Privacy-Preserving Design

AbdElKader Seif El Islem RAHMANI[1,2], Yasser YAHIAOUI[1,2], Abdelghani BOUZIANE[1,2]

[1]University Center of Naama Salhi Ahmed, Naama, 45000, Algeria | [2]EEDIS Laboratory, University Djillali Liabès of Sidi Bel Abbès, 22000, Algeria

ABSTRACT

Current voice assistants suffer from a fundamental architectural limitation: **stateless design**. Each interaction is treated as an isolated event, precluding meaningful personalization. We present a **retention-augmented architecture** that addresses this through explicit, transparent memory mechanisms. Our system achieves **88% Personalization Success Rate (PSR)** versus 0% for stateless baselines across 150 controlled test scenarios ($p < 0.001$).

PROBLEM STATEMENT

- **Stateless Architecture:** Commercial VAs treat each turn as independent, requiring users to repeatedly specify preferences
- **High Interaction Friction:** Users must re-state context in every session
- **No Long-term Adaptation:** Systems cannot learn user patterns over time
- **Privacy Concerns:** Neural memory approaches lack transparency and user control

KEY CONTRIBUTIONS

1. Novel explicit memory architecture for stateful dialogue

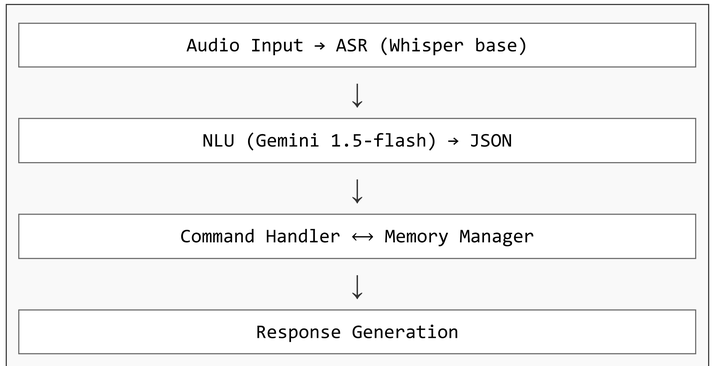
2. Personalization Success Rate (PSR) evaluation metric

3. Privacy-preserving design with on-device ASR

4. Rigorous empirical validation (n=150 scenarios)

5. Replicable blueprint for future research

ARCHITECTURE



ASR:	OpenAI Whisper (local processing)
NLU:	Zero-shot intent classification
Memory:	JSON-based user preferences
Privacy:	On-device audio, transparent storage

MEMORY MECHANISM

Core Innovation: Gap-filling via explicit preference storage

```
procedure HandleQuery(entities, memory):
  for entity in required_entities:
    if entities[entity] is null:
      if memory.preferences[entity]:
        entities[entity] ← memory[entity]
      else:
        request_clarification(entity)
  return generate_response(entities)
```

Memory Schema:

- User preferences (key-value pairs)
- Interaction history (timestamped actions)
- Context information (session data)

METHODOLOGY

Evaluation Metric: Personalization Success Rate (PSR)

$$PSR = (N_{resolved} / N_{total}) \times 100\%$$

Definition: Percentage of ambiguous queries successfully resolved using stored preferences without requiring user clarification.

Test Dataset:

Total scenarios:	150 controlled cases
Categories:	5 (cuisine, location, price, multi-preference, override)
Design:	Exceeds NLU context window
Conditions:	Retention-augmented vs. stateless baseline

DIALOGUE COMPARISON

Stateless System (✗)
User: "Find me a restaurant"
VA: "What cuisine?"
User: "Italian"
VA: "What location?"
Turns: 4+

Memory-Augmented (✓)
User: "Find me a restaurant"
VA: "Based on your preference for Italian food and downtown location, here are options..."
Turns: 1

Figure 1: Comparative dialogue efficiency

EXPERIMENTAL RESULTS

Primary Outcome: PSR Analysis

System	Resolved	PSR (%)	95% CI
Stateless	0/150	**0.0**	[0.0, 2.4]
Retention	132/150	**88.0**	[82.1, 92.5]

$p < 0.001$ (Fisher's exact test)

2.31×	1.8s
Interaction Efficiency Ratio	Avg. Latency

QUALITY METRICS

Human Evaluation (n=3 evaluators, 5-point Likert scale)

Dimension	Stateless	Augmented	Δ
Helpfulness	2.3±0.8	4.1±0.6	**+1.8**
Personalization	1.2±0.4	4.3±0.5	**+3.1**
Satisfaction	2.1±0.8	4.0±0.6	**+1.9**

ERROR ANALYSIS

18 Failures (12%) - Root Cause Distribution:

Error Type	Count	%
NLU Entity Extraction	6	33.3
ASR Transcription	4	22.2
Memory Logic	3	16.7
Intent Classification	2	11.1
Other	3	16.7

Memory retrieval overhead: 3.2 ms average (negligible)

CONCLUSIONS

- **Validated hypothesis:** Lightweight explicit memory achieves significant personalization gains
- **Privacy-first design:** On-device ASR + transparent storage
- **Practical architecture:** Low computational overhead, interpretable
- **Foundation established:** Blueprint for advanced stateful systems

Future Work:

- Implicit preference learning from interaction patterns
- Scalable storage (SQLite/vector DB)
- Longitudinal user studies (n>30, multi-week)
- Context-aware memory retrieval

SPECOM 2025
Speech and Computer

Contact: rahmani.abdelkaderseifelislem@cuniv-naama.dz

Code & data: Available upon request

Scan for paper