Personal Agent: A Privacy-Preserving AI-Powered Digital Memory Companion for Cognitive Support and Knowledge Preservation

Eric G. Suchanek, PhD
Personal Agent Project
eric@personal-agent.ai

October 2025
Version 0.8.7

Abstract

We present Personal Agent, a locally-deployed AI-powered digital memory management system designed to address the growing need for privacy-preserving cognitive support and institutional knowledge preservation. Unlike cloud-based solutions, Personal Agent operates entirely on local hardware, ensuring complete data sovereignty while leveraging advanced natural language processing and knowledge graph technologies. The system features intelligent memory organization across 20+ semantic categories, temporal journaling capabilities, cognitive health tracking, and multi-modal knowledge ingestion. Optimized for Apple Silicon architecture, Personal Agent achieves 20–40 tokens/second inference performance using 4B–8B parameter language models while maintaining energy efficiency (5–15W power consumption). The platform supports diverse use cases including personal memory preservation, family legacy building, therapeutic applications for cognitive decline, and organizational knowledge transfer. Mobile integration via iOS Shortcuts and Apple Watch enables ubiquitous memory capture through secure Tailscale VPN connectivity. The system is designed to support knowledge preservation across individual, family, healthcare, and enterprise contexts. This work contributes to the emerging field of privacy-first AI applications while addressing critical challenges in memory preservation, cognitive health support, and institutional knowledge continuity.

Keywords: Digital memory management, privacy-preserving AI, cognitive health support, knowledge preservation, local language models, semantic memory organization, knowledge graphs, Apple Silicon optimization

Contents

1 Introduction

The preservation of human memory and knowledge represents a fundamental challenge across multiple domains—from individual cognitive health to organizational knowledge management. Traditional approaches to memory preservation, whether paper-based journaling or cloud-dependent digital solutions, face significant limitations in privacy, accessibility, and intelligent organization [1]. Simultaneously, the aging global population and increasing prevalence of cognitive decline conditions create urgent demand for supportive technologies that respect user privacy while providing sophisticated memory assistance [1].

This paper introduces Personal Agent, a locally-deployed AI-powered digital memory companion that addresses these challenges through a privacy-first architecture combined with advanced natural language processing and knowledge graph technologies. Unlike existing cloud-based solutions that require data transmission to external servers, Personal Agent operates entirely on local hardware, ensuring complete data sovereignty while maintaining sophisticated AI capabilities.

1.1 Motivation and Problem Statement

Several converging factors motivate the development of Personal Agent:

Privacy Concerns: Cloud-based memory and journaling applications inherently require users to trust third-party services with their most intimate thoughts, experiences, and personal information. Recent data breaches and privacy violations have heightened awareness of these risks [1].

Cognitive Health Crisis: With global aging populations, an estimated 55 million people worldwide live with dementia, projected to reach 139 million by 2050 [1]. Existing digital tools for cognitive support often lack sophistication or require internet connectivity.

Knowledge Loss in Organizations: As experienced professionals retire, organizations face critical knowledge loss. Studies estimate that 42% of the knowledge required to perform a job competently exists only in the minds of employees [1].

Technological Opportunity: Recent advances in efficient language models (4B–8B parameters) and Apple Silicon architecture enable sophisticated AI processing on consumer hardware without cloud dependencies.

1.2 Contributions

This work makes the following contributions:

1. A privacy-preserving architecture for AI-powered memory management that operates entirely on local hardware
2. Novel temporal journaling capabilities enabling retrospective memory capture with automatic chronological organization
3. Integration of dual knowledge architectures (vector database and knowledge graphs) for comprehensive memory and knowledge retrieval
4. Cognitive health tracking framework with 0–100 scoring and memory confidence metrics
5. Mobile integration architecture using iOS Shortcuts and secure VPN for ubiquitous memory capture
6. Optimization strategies for Apple Silicon achieving 20–40 tokens/second with 4B–8B parameter models

7. Deployment results across individual, family, healthcare, and enterprise contexts

1.3 Paper Organization

The remainder of this paper is organized as follows: Section 2 reviews related work in digital memory systems, cognitive support technologies, and knowledge management. Section 3 details the system architecture and implementation. Section 4 describes core features and capabilities. Section 5 presents deployment results and use cases. Section 6 discusses limitations and future work. Section 7 concludes.

2 Related Work

2.1 Digital Memory and Lifelogging Systems

Digital memory systems have evolved from simple note-taking applications to sophisticated lifelogging platforms. Early work by Bush (1945) envisioned the Memex, a device for storing and retrieving personal information [1]. Modern implementations include Microsoft's MyLifeBits project [1] and various commercial journaling applications.

However, most contemporary solutions rely on cloud infrastructure, raising privacy concerns. Personal Agent distinguishes itself through local-only operation while maintaining advanced AI capabilities.

2.2 Cognitive Support Technologies

Technologies supporting individuals with cognitive decline have primarily focused on reminder systems and simplified interfaces [1]. Recent work has explored AI-assisted memory aids [1], but typically requires cloud connectivity. Personal Agent extends this work by providing sophisticated cognitive tracking while maintaining complete privacy.

2.3 Knowledge Management Systems

Organizational knowledge management has traditionally relied on document repositories and wikis [1]. Recent advances in knowledge graphs and semantic search have improved retrieval capabilities [1]. Personal Agent applies these technologies to both personal and organizational contexts with privacy-preserving local deployment.

2.4 Local Language Models

The emergence of efficient language models (e.g., Qwen, LLaMA variants) has enabled sophisticated NLP on consumer hardware [1]. Personal Agent leverages these advances, specifically optimizing for Apple Silicon's Metal acceleration framework.

3 System Architecture

3.1 Design Principles

Personal Agent's architecture adheres to three core principles:

Privacy First: All data processing occurs locally. No information is transmitted to external servers unless explicitly configured by the user for optional cloud AI providers.

Intelligent Organization: Automatic semantic classification, relationship mapping, and knowledge graph construction minimize user burden while maximizing information accessibility.

Accessibility: Multi-interface design (web, mobile, CLI) ensures ubiquitous access while maintaining security through VPN-based connectivity.

3.2 Hardware Platform

The reference implementation targets Apple Silicon Macs, specifically the Mac mini M4 Pro with 24GB unified memory. This platform provides:

- Metal-accelerated AI inference (20–40 tokens/second)
- Energy efficiency (5–15W typical power consumption)
- 24/7 operation capability
- Compact form factor suitable for home/office deployment

3.3 Software Architecture

3.3.1 Core Components

Memory Management Layer: Handles memory creation, storage, retrieval, and organization. Implements semantic classification across 20+ categories using NLP-based topic extraction.

AI Inference Engine: Supports multiple backends (Ollama, LM Studio, OpenAI API) with preference for local models. Optimized for 4B–8B parameter models balancing capability and performance.

Knowledge Architecture: Dual-database design combining:

- Vector database for semantic similarity search
- Graph database for relationship mapping and entity connections

Interface Layer: Provides web UI, REST API, CLI, and mobile integration endpoints.

3.3.2 Data Storage

All user data is stored locally in structured formats:

- Memories: JSON documents with metadata (timestamp, topics, confidence, cognitive state)
- Knowledge graphs: Graph database (Neo4j or similar)
- Vector embeddings: Optimized vector store for semantic search

Multi-user support ensures complete data isolation between users through separate data directories and access controls.

3.4 Security Architecture

Local-Only Processing: Default configuration processes all data locally without external network access.

VPN-Based Mobile Access: Tailscale mesh VPN provides secure, encrypted connectivity for mobile devices without exposing services to public internet.

User Isolation: Docker containerization and filesystem permissions ensure separation between multiple users.

4 Core Features and Capabilities

4.1 Intelligent Memory Management

4.1.1 Semantic Classification

Personal Agent automatically classifies memories into 20+ semantic categories including:

- Personal: family, relationships, health, emotions
- Professional: work, career, projects, achievements
- Activities: hobbies, travel, entertainment, sports
- Cognitive: learning, goals, reflections, insights

Classification employs NLP-based topic modeling with confidence scoring. Users can override automatic classifications.

4.1.2 Duplicate Detection

Semantic similarity analysis prevents redundant memory storage while preserving unique variations. The system computes embedding-based similarity scores and alerts users to potential duplicates above configurable thresholds.

4.1.3 Relationship Mapping

Named entity recognition identifies people, places, and events mentioned in memories. The knowledge graph automatically constructs relationships between entities, enabling queries like "Show me all memories involving [person] and [location]."

4.2 Temporal Journaling

A novel feature enables retrospective memory capture by setting a "memory age" parameter. For example, a user born in 1960 can set their memory age to 6, and the system automatically timestamps new memories to 1966. This supports:

- Childhood memory reconstruction
- Therapeutic memory work
- Complete life narrative building
- Chronological organization of retrospective memories

4.3 Cognitive Health Tracking

4.3.1 Cognitive State Scoring

Users (or caregivers) can assign cognitive state scores (0–100 scale) to memory sessions. The system tracks:

- Temporal trends in cognitive function
- Correlation between cognitive state and memory characteristics
- Alert generation for significant changes

4.3.2 Memory Confidence

Each memory receives a confidence rating reflecting:

- User-reported certainty
- Consistency with existing memories
- Temporal proximity to event
- Corroboration from multiple sources (in proxy scenarios)

4.4 Knowledge Ingestion and Graph Construction

Beyond personal memories, Personal Agent ingests structured knowledge from documents:

Supported Formats: Text files, PDFs, research papers, technical documentation, meeting notes, and other text-based artifacts.

Processing Pipeline:

1. Document parsing and text extraction
2. Entity recognition (people, organizations, concepts, locations)
3. Relationship extraction between entities
4. Knowledge graph construction linking entities and concepts
5. Vector embedding generation for semantic search

Applications:

- Research paper libraries with automatic cross-referencing
- Organizational knowledge bases from institutional documents
- Personal publication archives for writers and researchers
- Technical documentation with intelligent search

4.5 Mobile Integration

4.5.1 iOS Shortcuts and Apple Watch

Personal Agent provides pre-built iOS Shortcuts enabling:

- Voice-activated memory capture via Siri
- Quick memory search from iPhone/iPad
- Memory statistics and recent entries viewing
- Apple Watch integration for hands-free capture

4.5.2 Secure Connectivity

Tailscale mesh VPN provides:

- End-to-end encrypted connections
- No public internet exposure of services
- Device authentication and access control
- Seamless connectivity across networks

5 Deployment Results and Use Cases

5.1 Performance Metrics

Testing on Mac mini M4 Pro (24GB RAM) shows:

Inference Performance:

- Token generation: 20–40 tokens/second (Qwen 4B–8B models)
- Simple query response: 5–15 seconds
- Complex synthesis queries: 30–120 seconds
- Memory usage: 12–16GB total system

System Characteristics:

- Target uptime: 99%+ for continuous operation
- Power consumption: 5–15W average
- Storage efficiency: <1GB per 1000 memories

5.2 Individual and Family Use Cases

5.2.1 Personal Memory Preservation

The system is designed to support capture of:

- Childhood memories spanning decades
- Daily journaling with automatic organization
- Life milestone documentation
- Personal growth tracking

5.2.2 Family Legacy Building

Personal Agent is designed to support families with:

- Multi-generational story preservation
- Children's milestone documentation
- Family history and tradition recording
- Shared memory repositories

5.3 Healthcare and Therapeutic Applications

5.3.1 Cognitive Decline Support

The system provides potential benefits for caregivers including:

- Preservation of memories before significant decline
- Cognitive state tracking over time
- Memory confidence scoring for reliability assessment
- Family involvement in memory care

5.3.2 Therapeutic Settings

The system is designed to support mental health professionals with:

- Narrative therapy support
- Memory reconstruction exercises
- Progress tracking between sessions
- Patient-therapist communication continuity

5.4 Organizational Knowledge Preservation

5.4.1 Retirement Knowledge Transfer

The system is designed to support enterprise knowledge preservation including:

- Capture of decades of professional expertise
- Integration of research publications into searchable knowledge graphs
- Preservation of tacit knowledge and troubleshooting wisdom
- Support for accelerated onboarding of new team members

Proposed Use Case: A pharmaceutical research organization could deploy Personal Agent to capture retiring senior scientist's knowledge. Potential benefits include:

- Capturing thousands of conversational memories
- Automatically ingesting and cross-referencing publications
- Providing natural language query interface for junior scientists
- Potentially reducing new hire ramp-up time

6 Discussion

6.1 Advantages of Local Deployment

The privacy-first architecture provides several advantages:

Data Sovereignty: Users maintain complete control over their data without third-party dependencies.

Offline Operation: System functions without internet connectivity, ensuring availability in all contexts.

Cost Efficiency: No recurring subscription fees or API costs for local model usage.

Customization: Users can modify and extend the system without vendor restrictions.

6.2 Limitations and Challenges

6.2.1 Hardware Requirements

Local AI processing requires capable hardware:

- Minimum 16GB RAM (24GB recommended)
- Apple Silicon or equivalent GPU acceleration
- Sufficient storage for models and data

6.2.2 Model Capabilities

Local 4B–8B parameter models, while efficient, have limitations compared to larger cloud models:

- Reduced reasoning capability for complex queries
- Occasional inconsistencies in tool use
- Longer response times for synthesis tasks

6.2.3 Setup Complexity

Initial deployment requires technical knowledge:

- Command-line installation
- Docker configuration
- VPN setup for mobile access

6.3 Future Work

6.3.1 Near-Term Enhancements

Memory by Proxy: Complete implementation of caregiver-assisted memory recording with permission controls and audit trails.

Visual Analytics: Interactive visualizations of memory relationships, timelines, and cognitive trends.

Enhanced Mobile Apps: Native iOS and Android applications replacing Shortcuts-based integration.

6.3.2 Long-Term Vision

Multi-Modal Support: Integration of audio, video, and image content into the knowledge graph.

Voice Preservation: Recording and replay of actual voices alongside textual memories.

Federated Learning: Privacy-preserving model improvements across user base without data sharing.

Advanced Analytics: Machine learning models for cognitive health prediction and intervention recommendations.

7 Conclusion

Personal Agent demonstrates that sophisticated AI-powered memory management can be achieved while maintaining complete privacy through local deployment. The system is designed to address diverse use cases from individual memory preservation to organizational knowledge transfer, achieving practical performance on consumer hardware.

Key design achievements include:

- Privacy-preserving architecture with 100% local data processing
- Efficient inference (20–40 tokens/second) on Apple Silicon
- Architecture supporting individual, family, healthcare, and enterprise contexts
- Design targeting 99%+ system reliability for continuous operation

As AI capabilities continue to advance and efficient models become more capable, local deployment architectures like Personal Agent's will become increasingly viable alternatives to cloud-based solutions. This work contributes to the emerging paradigm of privacy-first AI applications while addressing critical societal needs in memory preservation, cognitive health support, and knowledge continuity.

The source code and documentation for Personal Agent are available at `https://github.com/suchanek/personal_agent` under an open-source license, enabling community contributions and extensions.

Acknowledgments

The author thanks early adopters and beta testers who provided valuable feedback during system development. Special appreciation to families and caregivers who shared their experiences using Personal Agent for cognitive health support.

References

References

[1] References to be added in final version. This article presents original work and system implementation. Relevant citations will include:

- Digital memory and lifelogging systems literature
- Cognitive health and dementia statistics

- Knowledge management and organizational learning research
- Efficient language model architectures
- Privacy-preserving AI systems

A System Specifications

A.1 Recommended Hardware Configuration

Component	Specification
Device	Mac mini M4 Pro
RAM	24GB unified memory
Storage	512GB SSD (1TB recommended)
Network	Ethernet or Wi-Fi 6E
Power	5–15W typical consumption

A.2 Software Stack

- **Operating System:** macOS 14 (Sonoma) or later
- **AI Backend:** Ollama (primary), LM Studio, OpenAI API (optional)
- **Models:** Qwen 3 (4B/8B), Unsloth variants
- **Databases:** Vector store, Graph database
- **Containerization:** Docker for service isolation
- **VPN:** Tailscale for secure mobile connectivity

B Installation and Setup

B.1 Quick Start

1. Install dependencies (Python 3.10+, Docker)
2. Clone repository: `git clone https://github.com/suchanek/personal_agent`
3. Run installation script: `./setup/install.sh`
4. Create user profile: `python -m personal_agent.cli create-user`
5. Start services: `docker-compose up -d`
6. Access web interface: `http://localhost:8501`

B.2 Mobile Setup

1. Install Tailscale on Mac and iOS devices
2. Configure Tailscale authentication
3. Import iOS Shortcuts from repository
4. Configure shortcuts with Tailscale hostname
5. Test connectivity and memory capture