# SAGE: A Secure Agent Governance Ecosystem for Organizational Learning

Darrell L. Young dlyoung@smu.edu Jason Teske jason@wave3ai.com

May, 2025

**Abstract**

This paper proposes SAGE, a Secure Agent Governance Ecosystem, as a universal three-tier architecture for agent-based systems adaptable to various organizational contexts. The architecture consists of client-facing agents at the lower tier, supervisory agents overseeing clusters of clients at the middle tier, and managerial agents setting abstract policies at the upper tier. We outline a rigorous experimental framework, including longitudinal A/B testing and user evaluations, to empirically validate learning efficiencies and user satisfaction.

# 1 Introduction

As AI-enabled services proliferate, the need for adaptive, scalable organizational models grows. The proposed SAGE architecture provides a flexible structure, capable of continuous learning at all operational levels. While broadly applicable, the Caring-Call.com elder care service exemplifies a specific implementation scenario. This builds upon foundational work in generative AI for agile operations [2] and prompt engineering for adaptive decision-making [3]. Moreover, as Dhinakaran et al. [1] highlight in *The Future of Well-Being*, integrating a holistic approach that combines multiple privacy-preserving techniques—summarized in their Table 14.1—is essential for secure, AI-powered health management.

# 2 SAGE Learning Architecture

## Client-Centered Management

Each client within the SAGE ecosystem is managed through a personalized and evolving understanding maintained by a Tier 1 Client Agent. These agents engage in regular check-ins, track behavioral patterns, monitor sentiment and wellness indicators, and adapt follow-up questions to the client's communication style and changing needs. The agent's role includes:

- Recording and analyzing conversation transcripts for signs of distress, medical changes, or emotional shifts.

- Summarizing client well-being in structured reports for supervisory review.
- Coordinating tool use (e.g., sentiment analyzers, topic models, symptom checkers) to enhance interaction quality without requiring escalation.
- Requesting guidance or support from the Agent Pool when specialized insight is warranted.
- Maintaining continuity in engagement to build trust and rapport over time.
- Managing the client's calendar of upcoming wellness checks, medication routines, or follow-up conversations.
- Automatically sending calendar-integrated reminders via SMS and email to ensure the client is aware of scheduled video calls, with links embedded securely.
- When granted permission, integrating with the client's prescription refill calendar to provide medication adherence support and reminders for upcoming pharmacy pick-ups or auto-refills.
- Tracking response to reminders and escalating to supervisors if appointments are repeatedly missed or ignored.
- Optionally integrating with smart devices such as smart watches, fitness trackers, or in-home health monitors—when granted permission by the client—to passively collect biometric, mobility, and behavioral data that enhance early detection of health issues and support proactive intervention.

When new symptoms, behaviors, or concerns are detected, the Client Agent applies embedded decision logic to determine whether it can respond autonomously, escalate to a Supervisor Agent, or request input from a specialized agent. This client-centered management approach ensures the experience is not only automated but also empathetically tailored to each individual's context and risk profile. Each client within the SAGE ecosystem is managed through a personalized and evolving understanding maintained by a Tier 1 Client Agent. These agents engage in regular check-ins, track behavioral patterns, monitor sentiment and wellness indicators, and adapt follow-up questions to the client's communication style and changing needs. The agent's role includes:

- Recording and analyzing conversation transcripts for signs of distress, medical changes, or emotional shifts.
- Summarizing client well-being in structured reports for supervisory review.
- Coordinating tool use (e.g., sentiment analyzers, topic models, symptom checkers) to enhance interaction quality without requiring escalation.
- Requesting guidance or support from the Agent Pool when specialized insight is warranted.

- Maintaining continuity in engagement to build trust and rapport over time.

When new symptoms, behaviors, or concerns are detected, the Client Agent applies embedded decision logic to determine whether it can respond autonomously, escalate to a Supervisor Agent, or request input from a specialized agent. This client-centered management approach ensures the experience is not only automated but also empathetically tailored to each individual's context and risk profile.

The architecture comprises:

- **Tier 1: Client Agents** – Direct interaction with end-users, personalized learning from interactions.
- **Tier 2: Supervisor Agents** – Oversee client clusters, aggregate interaction data, synthesize insights.
- **Tier 3: Manager Agents** – Strategic decision-making based on high-level performance data and feedback from supervisors and clients.

Although a single general-purpose agent with tool-calling capabilities is sufficient for many tasks, SAGE includes a dynamic agent pool to support specialization when needed. Specialized agents may be activated in response to particular topics, health issues, or complex decisions requiring domain-specific reasoning. Overuse of agent proliferation is discouraged to maintain simplicity and traceability.

## Agent Pool and Specialization

SAGE includes an agent pool available to all tiers, where any agent can request assistance from a specialized peer. Examples of specialized agents include:

- Diabetes Management Agent
- Cardiovascular Risk Monitor Agent
- Fall Detection and Mobility Agent
- Cognitive Decline Observer Agent
- Depression and Mood Analytics Agent
- Chronic Pain and Inflammation Tracker
- Medication Adherence Agent
- Nutrition and Hydration Monitor
- Sleep and Circadian Rhythm Advisor
- Respiratory Symptom Tracker (e.g., COPD, asthma)

- Cancer Survivorship Support Agent
- Caregiver Support Advisor

These specialized agents can be consulted by general-purpose client, supervisor, or manager agents depending on context, and their outputs are routed through the originating agent for continuity.

### Emergency Coordination Agent

SAGE includes a dedicated Emergency Agent responsible for managing high-priority, time-sensitive interventions. This agent is designed with protocols for:

- Triggering direct 911 service alerts if a medical emergency is detected.
- Notifying caregivers and supervisors in borderline or ambiguous cases.
- Consulting with healthcare professionals (e.g., registered nurses or physicians) when required.

All SAGE agents are trained in escalation protocols, including threshold scoring on distress indicators, uncertainty classification, and caregiver override routes. Agents are provided with context-aware decision trees to determine:

- When to escalate to a caregiver versus a medical professional
- How to defer to a manager in non-clinical policy matters
- What constitutes a true emergency requiring external intervention

Standard operating procedures are uniformly distributed and reinforced across the agent system, ensuring consistency and accountability.

## 3 Diagram

## 4 Learning Combinations Across Tiers

The architecture supports extensive learning interactions:

- **Tier 1 from Tier 2:** Supervisors coach client agents.
- **Tier 1 from Tier 3:** Managers implement global policies.
- **Tier 2 from Tier 1:** Supervisors learn client interaction patterns.
- **Tier 2 from Tier 3:** Managers inform supervisors' operational strategies.
- **Tier 3 from Tier 1:** Aggregate data from client-level informs strategic policy.
- **Tier 3 from Tier 2:** Supervisor summaries inform managerial policies.
- **Intra-tier learning:** Peer benchmarking within tiers.

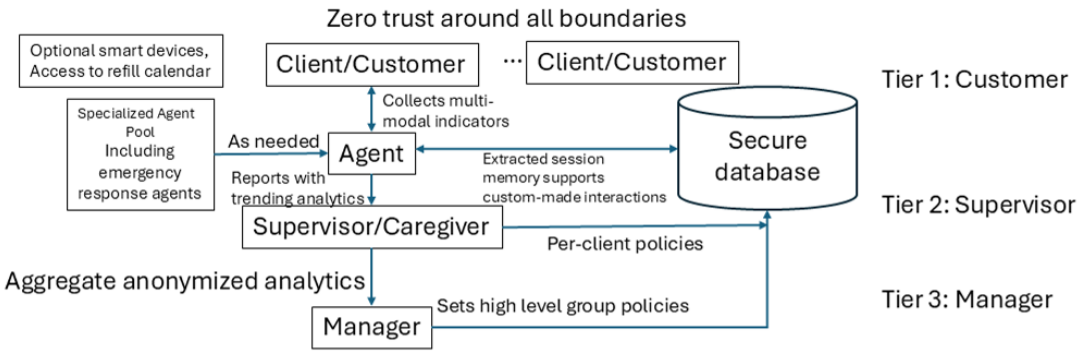


Figure 1: SAGE architecture showing inter- and intra-tier feedback loops, enabling both bottom-up data-driven learning and top-down policy refinement.

# 5 Planned Experiments

To scientifically evaluate the architecture, we propose:

1. A/B tests assessing user satisfaction and resolution efficiency with and without the three-tier structure.
2. Time-series analyses tracking knowledge acquisition and strategy shifts at supervisor and managerial levels.
3. Longitudinal studies examining client adaptability in diverse domains (elder care, customer service, education).
4. Real-time performance tracking evaluating managerial interventions.

These evaluation methods parallel techniques used in prompt engineering frameworks for emergent behavior modeling [3].

# 6 Universal Applicability

The architecture provides a generalizable framework beneficial across departments and entire organizations:

- **Inter-departmental:** Client agents handle departmental tasks; supervisors coordinate departmental clusters; managers set inter-departmental policies.
- **Cross-organizational:** Facilitates communication, strategic alignment, and agile response to operational changes, similar to LLM-enabled agile contracting methods described by Young et al. [2], enhancing team dynamics, capacity management, and cross-organizational collaboration.

- **Sector-specific applications:** Education (tutors, curriculum mentors, policy overseers), Healthcare (patient triage, symptom synthesis, care pathways), Retail (customer service, sales optimization, supply chain management), Government (citizen guidance, service alignment, regulatory evolution).

# 7 Privacy and Security

The three-tier agent architecture is built upon strong principles of privacy, confidentiality, and trust. Role-Based Access Control (RBAC) ensures that agents only access data strictly necessary for their function:

- **Supervisors (Tier 2)** can only access information related to clients directly under their care. They are prohibited from viewing any other client data.
- **Managers (Tier 3)** receive only abstracted, de-identified analytics to guide high-level decisions. No individual client data is exposed unless explicit, revocable consent is granted by the client.

In addition to access controls, data security is rigorously enforced:

- **Encryption at Rest and in Transit:** All data is encrypted using AES-256 or higher while stored in the database and during communication between system components.
- **Zero-Trust Architecture:** The system is deployed in a cloud-native environment that implements Zero-Trust security models to prevent lateral movement during potential breaches. Each agent and service must authenticate and be continuously verified.
- **Secure Destruction of Call Records:** Voice or chat interactions are automatically purged from storage after key analytics are extracted. This destruction is performed using NIST-approved zeroization methods to guarantee irreversible data deletion and uphold client privacy.

The initial list of indicators—such as mood, distress flags, physical state, and key topics—is extracted from the transcript immediately following the session using secure local or cloud inference. After scoring and analysis, the original call data is never retained. This multi-layered privacy and security strategy ensures regulatory compliance (e.g., HIPAA, GDPR) and builds end-user trust through transparent, privacy-preserving design.

# 8 Wellness as a Specific Case

Caring-Call.com illustrates a targeted application where Client Agents manage daily elder check-ins, Supervisor Agents detect behavioral trends, and Manager Agents adjust care strategies based on aggregated insights.

# 9 Conclusion

The proposed three-tier agent architecture represents a transformative organizational design paradigm, enabling adaptive learning and strategic policy creation. Its scalability and adaptability make it universally applicable, validated by rigorous planned experiments.

# References


[1] D. Dhinakaran, S. Edwin Raja, J. Jeno Jasmine, P. Vimal Kumar, and R. Ramani. The future of well-being: Ai-powered health management with privacy at its core. In Bharat Bhushan, Akib Khanday, Khursheed Aurangzeb, Sudhir Kumar Sharma, and Parma Nand, editors, *Advances in Computational Intelligence and Communication Systems*, chapter 14. John Wiley & Sons, Hoboken, NJ, December 2024.

[2] Darrell L. Young, Perry Boyette, James Moreland Jr., and Jason Teske. Generative ai agile assistant. In Jessie B. Walker and Rebekah E. Jansen, editors, *Disruptive Technologies in Information Sciences VIII*, volume 13058, page 1305809. SPIE, International Society for Optics and Photonics, 2024. Presented at SPIE Defense + Commercial Sensing, National Harbor, Maryland, United States.

[3] Darrell L. Young, Eric C. Larson, and Mitchell A. Thornton. Prompt engineering for detecting phishing. In *Assurance and Security for AI-enabled Systems 2025*, volume 13476, page 1347609. SPIE, 2025.