

# Risk Consulting as a Software Platform: A Governed Multi-Agent Architecture for Enterprise Risk Intelligence and Decision Support

Vijay Morampudi

*International Institute of Information Technology Hyderabad*

**DOI:** 10.37819/Jaike.26.0020

## Article history:

Received : 15.03.2026

Revised : 29.08.2026

Accept : 30.08.2026

Published : 21.10.2026

## Corresponding author:

### E-mail:

m.vijay.choudhary@gmail.com

**Abstract:** This paper presents Risk Consulting as a Software Platform (RCaS), a governed multi-agent architecture for enterprise risk intelligence and decision support in regulated environments. RCaS separates agent autonomy, agency, and authority; constrains unsafe capability combinations through a Capability-Separation Constraint (CSC); and governs institutional memory across episodic, semantic, and procedural knowledge, with runtime identity, scoped authorization, evidence provenance, and Expert-in-the-Loop (EITL) oversight enforced throughout. The paper contributes (1) a governance-by-design architecture integrating these mechanisms into a single operating model for bounded delegation, and (2) observational deployment evidence: the platform was developed and progressively transitioned to production over approximately six months, followed by approximately six months of steady-state production operation across approximately 7,000 enterprise risk assessment workflows, processing 16,500 engineering documents and 2,100 multimodal evidence files with 950 expert-reviewed recommendations. Production observations included ~70% automation of routine advisory activity, up to 90% reduction in assessment delivery time, ~95% document extraction accuracy, two- to three-fold gains in consultant throughput, and 90–97% expert acceptance of recommendations. Because the evaluation was observational and single-organization, results are reported as deployment-associated observations rather than causal or generalizable performance claims.

**Keywords:** Generative AI, agentic AI, compound AI, expert-in-the-loop, multi-agent systems, risk intelligence, decision support systems

## 1. INTRODUCTION

Enterprises operate in increasingly interconnected risk environments, where operational disruption, cyber threats, regulatory change, climate events, and supply-chain volatility can interact across organizational boundaries and produce cascading consequences. Traditional risk consulting addresses this through human-led advisory work: experienced professionals collect evidence, assess exposure, and recommend mitigations. This model provides deep domain expertise but is episodic and labor-intensive, and organizational knowledge tends to remain

distributed across individual practitioners rather than being systematically reusable.

Generative AI and multi-agent systems create an opportunity to augment this model — specialized agents can perform evidence extraction, multimodal analysis, retrieval, simulation, and coordination, freeing risk professionals to focus on judgment and consequential decisions. But enterprise risk advisory introduces requirements beyond model accuracy: agents may process untrusted information, touch proprietary institutional knowledge, and influence recommendations that carry financial and regulatory

weight. These risks cannot be addressed through model-level guardrails or output validation alone.

The central architectural challenge is therefore controlled delegation: which decisions and actions can be delegated to agents, what capability that requires, what authority is granted in context, and how that authority should evolve as reliability and conditions change. This paper treats identity, authorization, evidence quality, institutional knowledge, human oversight, and revocability as first-class components of the agent operating model, not as controls bolted on after execution.

We present Risk Consulting as a Software Platform (RCaS), a governed multi-agent architecture that integrates interactive assessment, continuous risk intelligence, institutional memory, and governance into a common operating model. Its defining principle is that an agent's capability to make a decision is architecturally distinct from its authority to execute one.

**Contributions.** First, a governance-by-design architecture for enterprise multi-agent risk advisory, built around the separation of autonomy, agency, and authority; the Capability-Separation Constraint; governed institutional memory; and runtime identity, authorization, provenance, and EITL oversight. Second, observational evidence from a production deployment of this architecture: the platform was developed and progressively transitioned to production over approximately six months, followed by approximately six months of steady-state production operation, from which the reported evaluation is drawn. Findings are reported as deployment-associated observations rather than causal estimates, given the single-organization, non-experimental evaluation design (Section 7).

This paper is a systems and architecture contribution: it does not propose a new foundation model, planning algorithm, or agent communication protocol. Its contribution is the integration of existing capabilities — agent orchestration, memory, identity, authorization, and human oversight — into a coherent, auditable model for bounded delegation in regulated, high-stakes enterprise workflows.

**Organization.** Section 2 reviews related work and positions the architectural gap RCaS addresses. Section 3 describes the evidence foundation. Section 4 presents the architecture. Section 5 distills deployment lessons. Section 6 reports production

results. Section 7 discusses limitations. Section 8 concludes.

## 2. RELATED WORK

### 2.1. Multi-Agent Systems and Agentic AI

Multi-agent systems research has long explored decomposing complex tasks across autonomous or semi-autonomous agents; agentic AI extends this by combining foundation models with planning, tool use, and iterative execution. Yao et al. demonstrated the interleaving of reasoning and action in language-model agents (ReAct), and subsequent frameworks explored structured multi-agent coordination — Li et al.'s CAMEL through role-based communication, and Wu et al.'s AutoGen through conversational orchestration. Guo et al. survey this space along the dimensions of agent roles, communication, and collective problem solving.

These frameworks establish agents as controllers of multi-step workflows rather than text generators, and Liu et al.'s AgentBench and related evaluation work emphasize that agent capability spans planning, tool use, memory, and reliability, not output quality alone. However, task coordination does not by itself define an enterprise delegation model. Production agents may access proprietary information or trigger consequential actions, so the technical capability to perform an action must be distinguished from the organizational authority to perform it — a distinction is not the primary focus of these frameworks.

### 2.2. Agent Evaluation, Security, and Controlled Delegation

Agents that consume tools, external information, and persistent memory carry a different risk profile than conventional LLM applications. Debenedetti et al.'s AgentDojo shows that indirect prompt injections where untrusted external data alter an agent's subsequent reasoning and tool calls — remains a practical vulnerability even for capable agents. Zhang et al.'s Agent Security Bench extends this to vulnerabilities in instruction handling, tool use, and memory retrieval, arguing that agent security cannot be reduced to output filtering; the capabilities, information sources, and actions available to an agent are themselves part of its security boundary. Survey work by Yehudai et al. and Mohammadi et

al. similarly frames evaluation around planning, tool use, memory, and safety as system-level properties rather than static benchmarks.

This literature motivates treating delegation as a *lifecycle* rather than a one-time grant based on offline evaluation — capabilities should be evaluated before production use, progressively expanded as reliability is demonstrated, and revocable when conditions change. RCaS operationalizes this lifecycle in Section 4.1.

### 2.3. Memory-Augmented AI and Institutional Knowledge

Lewis et al.'s Retrieval-Augmented Generation established the standard mechanism for coupling generation with externally retrieved knowledge, and subsequent work on long-term agent memory — Zhong et al.'s MemoryBank and the survey by Zhang et al. extended retention and reuse across interactions. Enterprise risk advisory, however, generates heterogeneous knowledge (assessments, expert judgments, taxonomies, regulatory mappings, procedures) that differs in authority and governance requirements; information from one interaction should not automatically become reusable organizational guidance. This literature addresses retrieval mechanics but gives limited treatment to *which* knowledge should be promoted, validated, versioned, or retired — the gap RCaS addresses through governed episodic, semantic, and procedural memory (Section 4.6).

### 2.4. AI Governance, Safety, and Human Oversight

Governance frameworks such as the NIST AI Risk Management Framework, its Generative AI Profile, and ISO/IEC 42001 provide organizational approaches to managing AI risk across the system lifecycle, and the EU AI Act codifies analogous obligations — risk management, documentation, human oversight, monitoring — for high-risk systems. Novelli et al. and Laux argue for accountability and meaningful human oversight as core governance requirements rather than post-hoc controls.

These frameworks establish *what* governance should achieve but generally provide less architectural detail on runtime enforcement. Recent work on Agentic AI governance similarly distinguishes an agent's technical capabilities from the autonomy it

is authorized to exercise, emphasizing that permitted autonomy should be determined by risk, oversight, accountability, reversibility, and organizational readiness rather than by technical capability alone. Agentic systems add a further concern: risk depends not only on individual agent behavior but on which capabilities are combined within a single agent — an agent that both processes untrusted input and can execute consequential actions carries materially higher risk than one restricted to a single capability class, a concern reinforced by the agent-security findings in Section 2.2. RCaS builds on these governance foundations through capability separation, least-privilege authorization, provenance, and EITL escalation, implemented as a runtime architecture (Section 4) rather than a set of standalone controls.

RCaS is also compatible with established enterprise security principles such as least-privilege access, separation of duties, and zero-trust enforcement, but applies them at the level of agent capability composition and delegated execution, where recent work highlights the need to preserve authorization across delegation, data retrieval, synthesis, and changing authorization boundaries. The distinction is that RCaS treats the combination of capabilities within an agent as an explicit governance boundary: access to untrusted inputs, sensitive institutional intelligence, and consequential actions is not treated as an undifferentiated permission set. This capability-composition view complements resource-level authorization by constraining how high-risk capabilities may be combined before runtime authorization is granted.

### 2.5. AI-Enabled Risk Intelligence

AI has been applied across enterprise risk tasks, including operational risk, compliance, and business risk, with persistent challenges around data quality, explainability, and governance. As AI systems increasingly evolve toward agentic operation, these challenges also extend to autonomous planning, action, coordination, and human oversight within enterprise processes. Existing AI-enabled risk applications are typically organized around a single risk domain or analytical capability, whereas enterprise risk *consulting* spans evidence collection, multimodal analysis, retrieval, scenario analysis, expert review, and organizational learning as a continuous lifecycle. The distinction between AI-enabled risk analytics and a continuously operating

advisory capability — one requiring orchestration, institutional knowledge, and governed human decision-making rather than an additional isolated model — motivates RCaS’s platform framing.

2.6. Positioning of This Work

Table 1 summarizes the gap: existing research addresses agent coordination, evaluation and security, persistent memory, governance, and risk analytics

as largely separate layers. RCaS’s contribution is their integration into a single governed operating architecture for high-stakes enterprise delegation — treating delegation as a lifecycle in which capability separation, institutional memory, runtime authorization, evidence provenance, and human oversight operate as coordinated components, rather than as independent mechanisms layered after the fact.

The distinction is summarized in Table 1.

Table 1. Research landscape and architectural gap addressed by RCaS

| Existing work            | Primary focus                                                               | Gap addressed by RCaS                                                                                                                                            |
|--------------------------|-----------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Multi-agent frameworks   | Agent coordination, communication and task execution                        | Limited separation of agent capability, decision authority, and execution authority                                                                              |
| Agent/tool-use research  | Planning, reasoning, and tool invocation, reliability, and agent evaluation | No enterprise lifecycle for delegation, authorization, monitoring, revocation                                                                                    |
| Agent-security research  | Prompt injection, tool-use attacks, memory attacks, and agent robustness    | Evaluates threats/defenses; no integrated operating model for governed delegation                                                                                |
| Memory-augmented AI      | Persistent context, retrieval, and long-term memory                         | Limited treatment of organizational knowledge promotion, validation, versioning, authority, and retirement                                                       |
| AI governance frameworks | Risk management, accountability, transparency and oversight                 | Primarily provide governance principles and lifecycle controls rather than an integrated runtime architecture for multi-agent delegation                         |
| AI-enabled risk systems  | Domain-specific risk analytics                                              | Focus on individual risk capabilities rather than integrating continuous intelligence, institutional memory, governance, and Expert-in-the-Loop decision support |

3. EVIDENCE FOUNDATION AND PROVENANCE

Enterprise risk assessments reason over heterogeneous evidence — structured operational data, unstructured documentation, multimodal observations, interaction history, and external intelligence — that varies in format, provenance, and confidence. RCaS treats evidence management as a core pipeline component rather than a preprocessing step.

**Sources:** Structured data (asset inventories, loss records, risk registers, inspection schedules) supplies operational and quantitative context. Unstructured documentation (engineering reports, audit findings, mitigation plans, contracts, regulatory publications) supplies contextual and expert reasoning. Multimodal

evidence (photographs, walkthrough video, sensor data) supplies visual and physical context, analyzed jointly with other evidence rather than in isolation. Interaction history (conversational responses, expert annotations, escalation and remediation records) contributes context for subsequent assessments. External intelligence (climate, regulatory, environmental signals) is incorporated when it affects assessed exposure.

**Unified Risk Intelligence Layer (URIL):** URIL normalizes, enriches, validates, and governs evidence before it reaches reasoning agents, giving analytical workflows a consistent interface rather than requiring each agent to interpret raw enterprise data independently. Each evidence object carries metadata on source, acquisition time, ownership, confidence, and processing history, and relationships

across documents, images, and assessment sessions are retained to preserve cross-modal provenance.

URIL distinguishes three evidence states: raw evidence (the original, immutable document, image, or observation), derived observations (intermediate analytical outputs, revisable as new evidence arrives), and validated knowledge (analytical conclusions that have cleared expert and governance review and are eligible for institutional reuse). Separating validated knowledge from source evidence prevents analytical outputs from silently altering the underlying record.

**Provenance and quality:** Agent outputs, transformations, expert overrides, and policy decisions are linked to originating evidence, supporting three properties: *traceability* (recommendations link to supporting evidence), *reproducibility* (analytical states can be reconstructed), and *accountability* (interventions remain identifiable). Because evidence quality is uneven — incomplete document fields, ambiguous imagery, noisy sensor data, transcription error — RCaS applies evidence-type-specific validation (multi-pass document processing, multimodal consistency checks, transcription-confidence assessment) before information enters

A)

### Risk Consulting as a Software Platform



B)

### Platform Overview

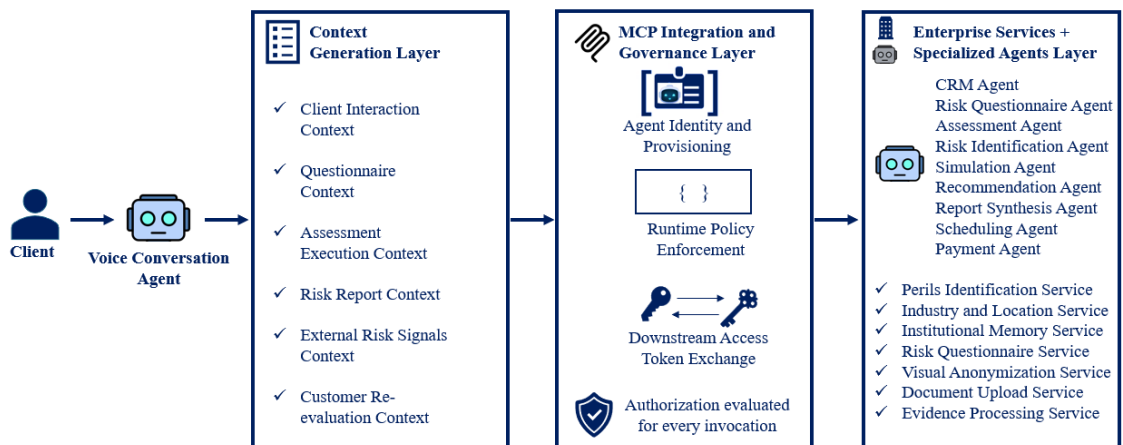


**Figure 1.** A) (End-to-End Architecture) and (B) (Governed Enterprise Agent Execution) are original diagrams prepared by the author to illustrate the four-layer architecture and the runtime authorization flow described below; they are not reproduced or adapted from any external source.

downstream reasoning, routing low-confidence or conflicting evidence to targeted EITL review rather than propagating it silently. The resulting lifecycle — *acquisition* → *normalization* → *enrichment* → *validation* → *governed retrieval* → *agent reasoning* → *expert validation* → *decision trace* → *institutional reuse* — preserves provenance and governance state throughout, and underpins the governance mechanisms described in Section 4.

4. PLATFORM ARCHITECTURE

RCaS coordinates specialized agents for evidence collection, multimodal analysis, risk reasoning, simulation, institutional learning, and decision support under explicit governance policy, across four integrated layers:

- 1. **Interactive Risk Decision Support** — conducts assessments, collects and analyzes evidence, generates recommendations, supports expert decision-making.
- 2. **Continuous Risk Intelligence** — monitors emerging risks, regulatory developments, and customer-specific changes outside active assessments.
- 3. **Institutional Memory** — preserves validated historical experience and reusable procedures.
- 4. **Governance and Control** — provides identity, authentication, authorization, policy enforcement, provenance, auditability, and Expert-in-the-Loop (EITL) oversight across the platform.

These layers operate as a coordinated system: interactive assessments consume institutional knowledge and continuously generated intelligence, while the Governance and Control Layer mediate all access to evidence, knowledge, and enterprise capability. The architecture’s organizing principle — carried through every mechanism below — is that agent capability is not equivalent to agent permission: an agent may be technically capable of an operation without being authorized to perform it.

4.1. Governance Foundations

**Autonomy, agency, and authority:** RCaS separates three dimensions of agent operation. *Autonomy* is an agent’s ability to interpret evidence and generate a decision without immediate human review — it does not imply unrestricted operation, as output remains subject to escalation on confidence, policy, and materiality. *Agency* is an agent’s ability to execute an action that changes an enterprise system or external environment (updating records, sending external notifications). *Authority* is the policy-granted permission to exercise autonomy, agency, or both within a defined scope, determined by organizational policy, operational risk, action materiality, and reversibility — not by model capability alone. In RCaS, autonomy and agency describe capabilities available to an agent, whereas authority determines whether those capabilities may be exercised in a particular context. This separation lets an organization grant substantial analytical autonomy without granting equivalent system-

Increase Autonomy and Agency Gradually

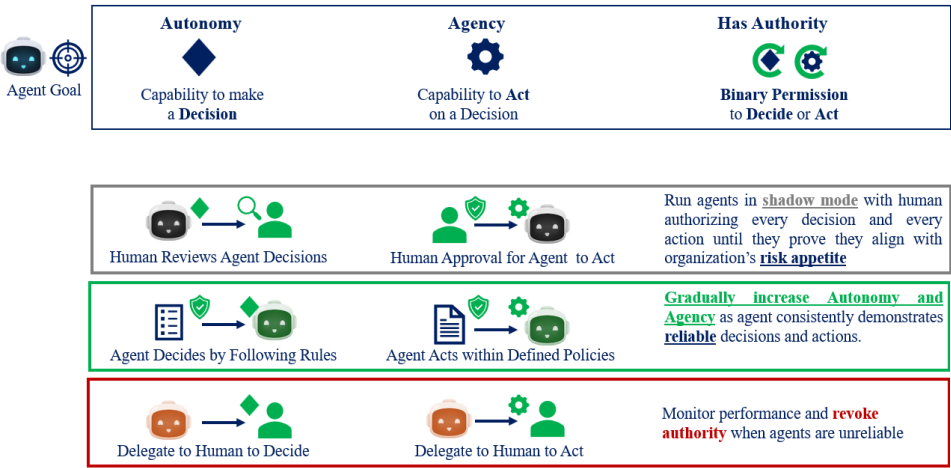


Figure 2. (Progressive Delegation and Revocation) is an original diagram prepared by the author

modifying authority and conversely allows narrowly scoped agency for low-risk, reversible actions while consequential decisions remain subject to human approval.

**Progressive delegation:** Capabilities are configurable runtime permissions, not permanent agent properties. A candidate capability typically enters shadow mode — evaluated against human decisions or predefined criteria without affecting production outcomes — before progressing through: *Observe* → *Shadow* → *Recommend* → *Execute with Approval* → *Bounded Autonomous Execution*. Transitions depend on demonstrated reliability, materiality, and reversibility; capabilities can equally be reduced or revoked as performance, policy, or conditions change (Figure 2). Authority allocation follows the same risk logic: high analytical confidence does not automatically authorize a consequential action, while a narrowly scoped low-risk action may be delegated even while the broader workflow stays under human review.

The deployment implications of progressive delegation are discussed separately in Section 5.

## 4.2. Capability-Separation Constraint (CSC)

Beyond bounding what a single agent may do, RCaS constrains which high-risk capabilities may be combined within one agent. The CSC identifies three capability classes — (1) processing untrusted external inputs, (2) accessing sensitive institutional

intelligence, and (3) executing consequential external actions — and no single agent is granted unrestricted access to all three simultaneously. Formally, if  $U$  denotes unrestricted processing of untrusted inputs,  $S$  denotes unrestricted access to sensitive institutional intelligence, and  $C$  denotes unrestricted consequential-action execution, the constraint is  $\neg (U \wedge S \wedge C)$  for a single agent. This is a capability-composition constraint rather than a replacement for least-privilege authorization: an agent may still be restricted to a single capability class, or to a bounded combination, according to policy and risk. Workflows requiring multiple classes are decomposed across independently governed agents connected through controlled interfaces (Figure 3): for example, an agent that extracts and classifies an externally supplied document does not simultaneously receive unrestricted institutional-memory access or system-modification authority; a separately governed agent performs any resulting retrieval or approved action, subject to its own authorization. Transitions between capability domains are protected by policy evaluation, least-privilege authorization, provenance tracking, and EITL approval where required — making capability separation a structural, architecture-level control rather than a prompt- or output-filtering safeguard. The CSC is intended as a risk-reduction design constraint rather than a complete security model: it does not by itself address identity compromise, authorization errors, tool-level vulnerabilities, or cross-agent attack paths, and additional controls (Sections 4.3 and 4.7) remain necessary alongside it.

### Agent Risk Dimensions and Controls

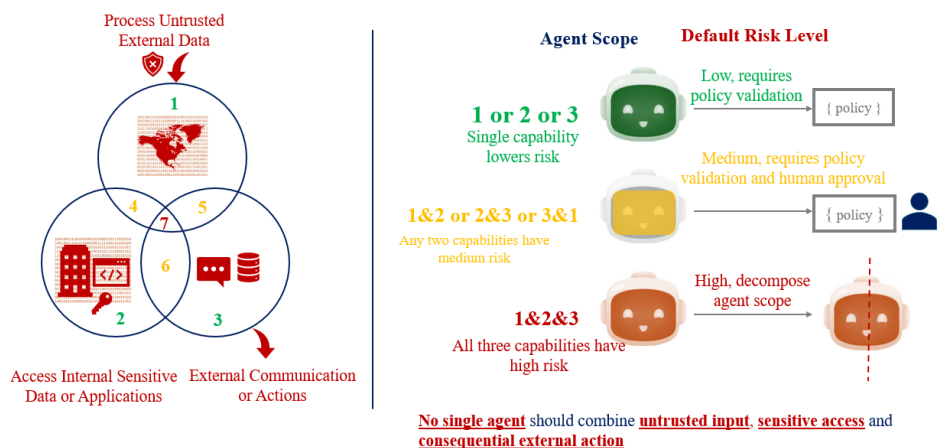


**Figure 3.** (Capability-Separation Constraint) is an original diagram prepared by the author

### 4.3. Runtime Identity, Authentication, and Authorization

Governance is enforced through a zero-trust runtime control architecture (Figure 4): agents hold no persistent, unrestricted privileges, and access is evaluated per protected operation. On launch, an agent receives an enterprise workload identity authenticated through the enterprise identity infrastructure; the Governance and Control Layer then evaluates whether the requested operation is

permitted in context and, if so, issues a short-lived, narrowly scoped authorization credential for that specific action, scope, and validity window, which the receiving enterprise service independently validates before executing. The resulting control sequence — *identity* → *authentication* → *policy evaluation* → *scoped authorization* — can be re-evaluated whenever tenant, resource, policy, risk, or operating context changes.

#### Agent Identity and Authorization

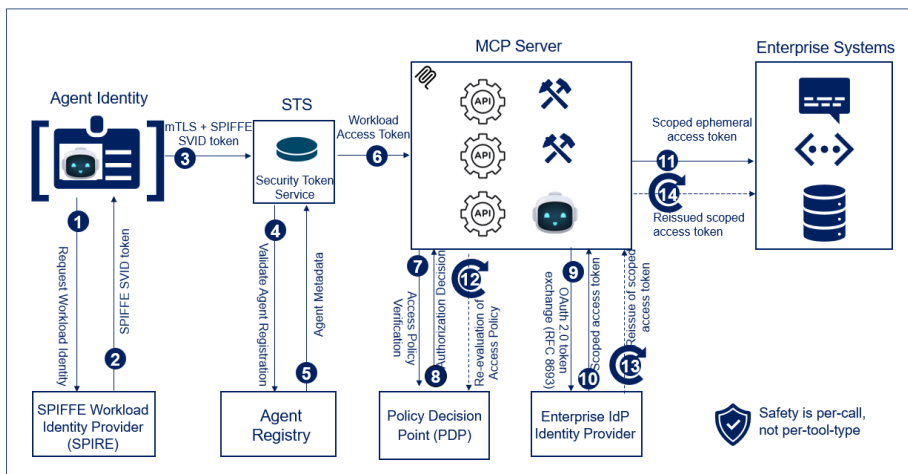


**Figure 4.** (Runtime Agent Identity, Authentication, Authorization, and Scoped Token Exchange) is an original diagram prepared by the author

### 4.4. Interactive Risk Decision Support Layer

This layer is the primary operational interface between users and the platform, coordinating agents across assessment execution, evidence collection, multimodal analysis, simulation, and reporting. Rather than a fixed questionnaire, the workflow adapts to enterprise context, available evidence, and remaining uncertainty, with governance controls determining whether outputs proceed automatically or require EITL intervention.

**Assessment and context.** The Risk Assessment Execution Agent manages adaptive questioning, response validation, preliminary risk identification, and confidence-aware escalation, producing analytical observations rather than final decisions. The Industry- and Location-Aware Context Service establishes the contextual envelope (industry, asset profile, jurisdiction, applicable regulation,

environmental hazard). The Customer Interaction Memory Agent retrieves relevant historical responses and rationale, ranked by relevance, recency, and governance policy.

**Evidence processing.** The Multi-Pass Document Extraction Agent processes engineering documentation using complementary perception passes with spatial consistency validation; fields meeting stability thresholds are promoted downstream, while uncertain fields are escalated for targeted expert validation. The Evidence Analysis Agent integrates structured, textual, multimodal, and regulatory evidence with citation-enforced generation, preserving traceability and exposing confidence and conflicts rather than resolving them silently.

**Simulation and review.** The Simulation and Trade-Off Evaluation Agent evaluates alternative mitigation strategies (scenario, sensitivity, resilience, cost,

compliance analysis) as advisory output requiring validation before adoption. The Risk Report Synthesis Agent combines validated evidence, analysis, and expert-approved rationale into citation-enforced reports. When EITL intervention is triggered, the platform passes the expert a structured decision context — validated evidence, reasoning trace, confidence, applicable policy, and unresolved uncertainty — rather than raw conversation history.

#### 4.5. Continuous Risk Intelligence Layer

This layer extends the platform beyond point-in-time assessment by monitoring regulatory, environmental, and customer-operational change, initiating targeted reassessment rather than requiring full reassessment when material change is detected. It supports two functions under the same runtime authorization and EITL controls used elsewhere: knowledge evolution, which proposes updates to questionnaires, risk knowledge, and mitigation guidance from external intelligence and operational experience, subject to governance review before reuse; and continuous customer re-evaluation, which compares historical assessments against current conditions and emerging intelligence, updating risk posture and generating alerts where warranted.

#### 4.6. Institutional Memory Layer

The Institutional Memory Layer converts validated organizational experience into reusable knowledge across three forms distinguished by purpose and governance requirement: episodic memory captures historical assessment experience — evidence, expert interventions, outcomes, semantic memory captures stable organizational knowledge — taxonomies, regulatory mappings and domain relationships, and procedural memory captures validated learning about how risk should be assessed and mitigated, including which recommendations or mitigation strategies were effective or ineffective, the conditions under which they were applied, implementation considerations and observed outcomes.

Procedural memory therefore allows the platform to learn not only what was recommended, but what happened after the recommendation was implemented. For example, a mitigation recommendation that repeatedly produces the desired risk reduction under a particular operating condition may become a validated procedural pattern, while a recommendation that is ineffective,

difficult to implement, or produces an unintended outcome can be marked accordingly and prevented from being treated as a preferred future intervention. This creates a governed feedback mechanism from recommendation → mitigation action → observed outcome → validated learning → future recommendation.

Memory is not an autonomous, self-learning mechanism. The Knowledge Curation Agent identifies candidate knowledge from completed assessments, expert interventions, mitigation outcomes, and operational feedback. Candidates are evaluated for evidence quality, consistency, expert agreement, outcome support, and policy alignment before promotion. Procedural knowledge receives the strictest governance because it can directly influence future recommendations and system behavior; changes require appropriate expert approval, version management, audit logging, and rollback capability. Historical observations or agent-generated conclusions therefore do not automatically become authoritative organizational knowledge.

#### 4.7. Coordinated Operation

**The four layers form a closed but governed learning cycle:** *external and customer signals → governed intelligence → assessment → evidence and analysis → expert validation → operational outcomes → governed knowledge curation → institutional memory → future assessments.* Learning therefore occurs through controlled updates to knowledge, procedures, and policy: validated evidence about which mitigation recommendations were effective, ineffective, or condition-dependent can influence future retrieval, recommendation guidance, workflow routing, and escalation criteria. The underlying foundation models remain unchanged in production; adaptation occurs through the governed information and control layers surrounding them. This preserves model version control, auditability, and rollback while allowing the operating environment and organizational knowledge to evolve. Delegated capabilities can consequently expand, constrained, or revoked independently of any model retraining.

### 5. DESIGN PRINCIPLES AND DEPLOYMENT LESSONS

Section 4 describes the architectural mechanisms; this section distills how those mechanisms were

operated in production, as deployment lessons rather than causal findings.

**Delegation as a lifecycle, not a ceiling:** Capabilities were introduced below their maximum scope and expanded only as shadow-mode evidence accumulated (*Identify* → *Shadow* → *Validate* → *Delegate* → *Monitor* → *Increase/Reduce/Revoke*). Autonomy and Agency are best treated as an operational control variable, not a fixed configuration.

**Evidence quality bounds delegation:** Reliability depends on the evidence entering a workflow as much as on the delegated capability itself. When evidence falls outside the conditions under which a capability was validated, the appropriate response is targeted review or reduced delegation — greater model reliability does not license greater delegation when the supporting evidence is unreliable.

**Decision lineage is operational, not just an audit artifact:** Reconstructing *how* a decision was reached — evidence, reasoning context, policy, human intervention — was required routinely, not only after the fact, to review, challenge, and reproduce AI-assisted decisions (Section 3).

**Organizational learning stays separate from model adaptation:** Validated experience is incorporated through controlled updates to knowledge, procedure,

and policy, evaluated against evidence and expert judgment before reuse, with stronger controls on changes that can influence subsequent system behavior — allowing organizational learning without uncontrolled online model modification.

**Human oversight is targeted, not comprehensive:** EITL review is concentrated on uncertainty, materiality, policy sensitivity, and novel conditions, with experts given structured context rather than raw interaction logs — automation changes the *allocation* of expert effort rather than removing expert accountability.

**Revocation is a normal state, not a failure mode:** Delegated capability was continuously monitored, and conditions change (data distributions, customer conditions, regulation) after validation. Response options — increased EITL involvement, reduced scope, return to shadow mode, outright revocation, or restoring a prior configuration — were treated as ordinary lifecycle transitions.

**Governance principles generalize; implementation detail does not:** The reusable contribution is the operating model for governed delegation — progressive introduction, evidence-aware evaluation, traceability, targeted oversight, and revocability — while evidence pipelines, taxonomies, thresholds, and escalation criteria remain necessarily domain-specific (Table 2).

**Table 2.** Summary of design principles

| Design principle                        | Deployment implication                                                                                             |
|-----------------------------------------|--------------------------------------------------------------------------------------------------------------------|
| <b>Progressive delegation</b>           | Validate capabilities in shadow mode before increasing autonomy or agency                                          |
| <b>Evidence-aware delegation</b>        | Consider evidence quality and operating conditions when determining delegation scope.                              |
| <b>Decision lineage</b>                 | Preserve the evidence, system context, policies, and human interventions underlying decisions.                     |
| <b>Governed organizational learning</b> | Update knowledge, procedures, and policies through controlled validation rather than uncontrolled model adaptation |
| <b>Focused human oversight</b>          | Direct expert attention toward uncertainty, exceptions, professional judgment, and consequential decisions.        |
| <b>Revocation and recovery</b>          | Treat reduction, rollback, and revocation as normal lifecycle states when conditions change.                       |
| <b>Architectural reuse</b>              | Standardize governance principles while allowing domain-specific knowledge, workflows, and thresholds to vary.     |

## 6. RESULTS AND ANALYSIS

This section reports production outcomes from approximately six months of steady-state production

operation that followed a six-month development-to-production rollout (Section 6.1), covering ~7,000 risk assessment workflows, 16,500 engineering

documents, 2,100 multimodal evidence files, and 950 expert-reviewed recommendations. Measurements come from production workflow logs, governance records, document-processing statistics, expert review, and operational monitoring; because the deployment was not conducted under controlled experimental conditions, findings are reported as deployment-associated observations rather than causal estimates.

### 6.1. Operational Performance

The platform was developed and progressively transitioned into production over approximately six months (pilot validation → controlled client expansion → scaled production), with governance policy, institutional knowledge, and workflow configuration refined through expert-approved changes during this period. This was followed by approximately six months of steady-state production operation, from which the results in this section are drawn. No online modification of foundation-model parameters occurred in production (Table 3).

**Table 3.** Operational performance

| Metric                       | Legacy Workflow | Production Platform | Observed Improvement  |
|------------------------------|-----------------|---------------------|-----------------------|
| Document Extraction Accuracy | 70%             | 95%                 | +25 percentage points |
| Assessment Delivery Time     | 30-40 hours     | 2-4 hours           | Up to 90% reduction   |
| Consultant Throughput        | 1x              | 2x-3x               | 100%-200% increase    |
| Routine advisory automation  |                 | 70%                 |                       |

**Note:** Deployment-associated observations, not causal estimates. Consultant throughput is reported as a relative index normalized to the legacy workflow (1×); the production observation represents 2–3× the corresponding legacy throughput.

Approximately 70% of routine advisory activity (evidence ingestion, document processing, preliminary risk identification, report preparation, workflow coordination) was automated, calculated as the share of predefined routine activities executed without manual intervention against the production workflow taxonomy. Judgment-dependent and consequential activities remained under EITL review. The denominator is the set of predefined routine activities in the production workflow taxonomy, rather than all advisory activities. These gains reflect the integrated operation of automated evidence

processing, adaptive workflows, retrieval-supported reasoning, and selective escalation collectively, not any single model or agent.

### 6.2. Evidence-Processing Quality

**Document extraction:** Extraction accuracy was assessed by comparing extracted field values against expert-reviewed reference values for the evaluated document sample, with a field counted as correct when its extracted value matched the reviewed reference. Accuracy improved across three successive processing configurations:

**Table 4.** Extraction accuracy across evidence processing configurations

| Configuration                   | Extraction Accuracy |
|---------------------------------|---------------------|
| OCR only                        | 70%                 |
| Single-pass extraction pipeline | 82%                 |
| Multi-pass extraction pipeline  | 95%                 |

*Values reflect production observations showing the progression across successive configurations, not a controlled benchmark.* Conventional OCR

underperformed on engineering documents with complex layouts, tables, and annotations; deterministic multi-pass processing with spatial/

consistency validation produced the highest observed accuracy, with the added benefit that uncertain fields remain linked to source location for targeted expert verification rather than silent propagation.

**Multimodal evidence and traceability:** Production observations indicated improved evidence completeness and fewer inconsistencies between reported conditions and visual evidence when multimodal observations shared a common evidence foundation with textual and structured data — interpreted as evidence of operational usefulness rather than a controlled accuracy estimate. Evidence provenance was maintained end-to-end, supporting traceability, reproducibility, and accountability as governance outcomes rather than performance metrics.

### 6.3. Recommendation Quality and Governance

Across 950 expert-reviewed recommendations, acceptance (recommendations accepted without material modification, excluding formatting changes) ranged from 90–97%, higher where evidence was rich and institutional knowledge mature, and lower for novel or under-documented environments — an indicator of usefulness and expert agreement, not of realized risk reduction or absolute correctness.

Selective, confidence- and materiality-based EITL escalation reduced the proportion of recommendations requiring comprehensive expert review by ~50–70% relative to the prior workflow, with the reduction calculated against the corresponding review proportion under the prior process. Higher-risk or lower-confidence cases remained subject to comprehensive review — making this a workflow-efficiency measure, not evidence that expert review is no longer needed. Continuous runtime policy validation, in turn, was associated with a 30–50% relative increase in identified policy non-conformance events compared with the prior validation process, calculated on the same production validation population used for the comparison. This reflects more deviations identified during execution, not a measured change in the underlying prevalence of violations.

### 6.4. Institutional Memory and Human Impact

Validated outcomes, rationales, and mitigation strategies were curated into episodic, semantic, and

procedural memory following multi-expert review. The deployment observed a 25–40% reduction in new risk-engineer onboarding time, measured as time to reach the established operational-readiness criterion used by the deployment, and reduced redundant evidence collection for recurring engagements via customer interaction memory — suggestive of value beyond retrieval efficiency, though the evaluation does not isolate memory's independent causal contribution.

Consistent with the governance model, the deployment shifted — rather than eliminated — expert effort: routine evidence processing and preliminary analysis moved to agents, while risk professionals concentrated more on exception handling, governance, and consequential judgment. This is a descriptive observation of the deployment, not a causal productivity estimate.

### 6.5. Summary (Table 5)

**Table 5.** Summary of production observations

| Dimension                                | Production observation    |
|------------------------------------------|---------------------------|
| Assessment workflows                     | ~7,000                    |
| Engineering documents processed          | 16,500                    |
| Multimodal evidence files                | 2,100                     |
| Expert-reviewed recommendations          | 950                       |
| Routine advisory automation              | ~70%                      |
| Document extraction accuracy             | ~95%                      |
| Assessment delivery time                 | Up to 90% reduction       |
| Consultant throughput                    | 2–3×                      |
| Expert recommendation acceptance         | 90–97%                    |
| Comprehensive expert review              | ~50–70% reduction         |
| Identified policy non-conformance events | ~30–50% relative increase |
| New risk-engineer onboarding time        | ~25–40% reduction         |

The document-extraction comparison (Table 4) offers the clearest component-level evidence within the deployment, since successive processing configurations were compared directly; the remaining results should be read as outcomes of the integrated architecture rather than attributable to any single mechanism, and do not constitute a controlled ablation study (Section 7.1).

## 7. LIMITATIONS AND THREATS TO VALIDITY

**Observational design:** The evaluation is a production deployment, not a controlled experiment; reported improvements are deployment-associated observations, and concurrent factors (process maturity, user familiarity, organizational adoption) may have contributed alongside the platform itself. No randomized control or full ablation study isolates the independent contribution of individual agents, governance mechanisms, or memory components — the document-extraction comparison (Table 4) is the closest available component-level evidence, but it is not a controlled ablation across otherwise identical conditions.

**Scope:** The deployment spans a single enterprise and primarily property/business risk workflows; organizational process, data characteristics, and regulatory context vary across organizations, industries, and jurisdictions, so the transferable contribution should be read as the governance and

operating architecture, not as evidence that specific agents, thresholds, or pipelines transfer unchanged.

**Measurement scope:** Reported metrics (automation, extraction accuracy, delivery time, throughput, acceptance, review volume) measure operational performance, not realized risk outcomes. Expert acceptance is not equivalent to recommendation correctness or realized risk reduction; an increase in identified policy non-conformance reflects detection under production validation, not a measured change in underlying violation prevalence. Longitudinal measures of mitigation effectiveness, claims experience, and regulatory outcomes are needed to evaluate downstream impact directly.

**Expert validation and bias:** Expert review is not an independent ground truth — reviewers may share organizational assumptions — and the author is affiliated with the organization that built and deployed RCaS, a source of potential bias mitigated by relying on workflow logs, governance records, and structured review rather than subjective assessment alone (see Conflict of Interest statement). Independent evaluation by uninvolved reviewers would strengthen these findings.

**Data availability:** Production datasets contain confidential client and organizational information and cannot be released, limiting external replication of the extraction and multimodal results using the original evidence distribution; privacy-preserving

or synthetic benchmarks could support broader validation.

**Drift and technology dependence:** Capability reliability is conditional on the data, policy, and operating conditions under which it was evaluated; the architecture addresses this through shadow evaluation, progressive delegation, and revocability, but the current evaluation does not establish these mechanisms' effectiveness over long periods or major environmental shifts. The platform also depends on foundation models, retrieval, and orchestration infrastructure that will evolve, and version control and audit trails mitigate but do not eliminate this dependence.

**Institutional memory governance:** Governed memory can still propagate historical assumptions if validation is weak; a poorly validated procedural update could influence assessments at scale. The reported onboarding and consistency observations should not be read as evidence that institutional memory is universally beneficial — longitudinal evaluation is needed to confirm it improves decision quality without reinforcing outdated practice.

**Boundaries of interpretation:** This is a systems and deployment evaluation. It does not establish causal superiority over conventional workflows, generalizable performance across organizations, realized long-term risk reduction, universal delegation criteria, elimination of expert judgment, or immunity to drift.

## 8. CONCLUSION

This paper presented RCaS, a governed multi-agent architecture for enterprise risk intelligence and decision support in regulated environments, built on the premise that model capability, autonomy, agency, and organizational authority are not equivalent — capabilities should be bounded, evaluated before delegation, authorized in context, and revocable as conditions change. Governance is embedded throughout architecture via the separation of autonomy/agency/authority, the Capability-Separation Constraint, runtime identity and authorization, evidence provenance, EITL oversight, and governed institutional memory (Section 4), operationalized through the deployment lessons in Section 5.

Following a six-month development-to-production rollout, approximately six months of steady-state

production operation across ~7,000 risk assessment workflows provides observational evidence that this architecture operates at production scale within the evaluated enterprise setting, with the efficiency, accuracy, and acceptance outcomes summarized in Section 6. These results support operational feasibility within the evaluated setting; consistent with Section 7, they are not causal estimates, evidence of realized long-term risk reduction, or proof of generalizable performance.

The broader contribution is an operating model for governed delegation — enabling routine analytical work to be progressively automated while preserving human accountability for exceptions, judgment, and consequential decisions, with organizational learning occurring through governed knowledge and policy evolution rather than uncontrolled model adaptation.

Future work should prioritize, in order of importance to the claims above: (1) controlled, matched-workflow comparison and ablation to isolate the contribution of individual mechanisms; (2) longitudinal evaluation of realized risk outcomes (mitigation effectiveness, claims experience, regulatory outcomes) beyond operational efficiency; (3) evaluation of the delegation lifecycle under changing conditions (data drift, regulatory change) to establish evidence-based revocation criteria; (4) validation across additional organizations, domains, and jurisdictions to separate broadly reusable elements from domain-specific configuration; and (5) independent, longitudinal evaluation of institutional memory governance as foundation models and infrastructure evolve.

## Conflict of Interest

The author is affiliated with the organization that designed, developed, and deployed RCaS. Reported outcomes derive from an internal production deployment and may be subject to organizational or researcher bias; to mitigate this, the evaluation relied primarily on production workflow logs, governance records, quality-assurance reviews, and structured EITL validation rather than subjective assessment, and validation samples were reviewed by domain experts not directly involved in platform implementation. Findings are presented as observational deployment outcomes, not causal performance claims. Independent evaluation across additional organizations is encouraged.

### Acknowledgements

The author thanks the Marsh AI Center of Excellence for its multidisciplinary support in developing and deploying RCaS, and the property risk engineering professionals, governance specialists, and early-adopter teams whose feedback and EITL validation informed the platform's architecture and governance mechanisms.

### Ethics Approval and Consent to Participate

This study did not involve human participants, clinical intervention, or animal experimentation. The research evaluated an enterprise AI platform using production workflow logs, governance records, quality-assurance reviews, document-processing statistics, and structured Expert-in-the-Loop (EITL) validation under applicable organizational governance and confidentiality requirements. Accordingly, formal research ethics committee approval and participant consent were not applicable.

### Human and Animal Rights

This research did not involve human subjects or animal experimentation. No identifiable personal information or client-identifiable information is disclosed in the article. The production evaluation was conducted using enterprise-governed information and anonymized operational records in accordance with applicable organizational controls.

### Consent for Publication

Not applicable. This article does not contain identifiable personal information, individual participant data, or other material requiring consent for publication.

### Data Availability Statement

Production datasets contain confidential client and proprietary information and are not publicly available. Reported metrics were derived from anonymized production workflow logs, governance records, quality-assurance reviews, and EITL validation processes, analyzed under applicable organizational governance and confidentiality requirements. Requests for further methodological information may be considered subject to organizational approval.

### Funding

No external funding was received for this research.

### Author Contributions

**Conceptualization:** Vijay Morampudi.

**Writing:** Vijay Morampudi.

**Reviewing:** Vijay Morampudi.

**Supervision:** Not Applicable.

**Editing:** Vijay Morampudi.

### REFERENCES

- Autio, C., Schwartz, R., Dunietz, J., Jain, S., Stanley, M., Tabassi, E., Hall, P., and Roberts, K., *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*, NIST Trustworthy and Responsible AI, NIST AI 600-1, National Institute of Standards and Technology, Gaithersburg, MD, 2024. <https://doi.org/10.6028/NIST.AI.600-1>.
- Debenedetti, E., Zhang, J., Balunovic, M., Beurer-Kellner, L., Fischer, M., and Tramèr, F., "AgentDojo: A Dynamic Environment to Evaluate Prompt Injection Attacks and Defenses for LLM Agents," *Advances in Neural Information Processing Systems*, vol. 37, pp. 82895–82920, 2024. <https://doi.org/10.52202/079017-2636>.
- European Parliament and Council of the European Union, "Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 2018/858, (EU) No 2018/1139 and (EU) No 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)," *Official Journal of the European Union*, L, 2024/1689, 12 July 2024.
- Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N. V., Wiest, O., and Zhang, X., "Large Language Model Based Multi-agents: A Survey of Progress and Challenges," *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 8048–8057, 2024. <https://doi.org/10.24963/ijcai.2024/890>.
- International Organization for Standardization, *ISO/IEC 42001:2023 — Information technology —*

*Artificial intelligence — Management system*, ISO, 2023.

- Kumar, N., Wei, X., and Zhang, H., “Agentic artificial intelligence as a new frontier in information systems: Promise, peril, and research opportunities,” *Information & Management*, vol. 63, no. 3, Article 104317, 2026. <https://doi.org/10.1016/j.im.2026.104317>.
- Laux, J., “Institutionalised distrust and human oversight of artificial intelligence: towards a democratic design of AI governance under the European Union AI Act,” *AI & Society*, vol. 39, pp. 2853–2866, 2024. <https://doi.org/10.1007/s00146-023-01777-z>.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., and Kiela, D., “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” *Advances in Neural Information Processing Systems*, vol. 33, 2020.
- Li, G., Hammoud, H. A. A. K., Itani, H., Khizbullin, D., and Ghanem, B., “CAMEL: Communicative Agents for ‘Mind’ Exploration of Large Language Model Society,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 51991–52008, 2023. <https://doi.org/10.52202/075280-2264>.
- Liu, X., Yu, H., Zhang, H., Xu, Y., Lei, X., Lai, H., Gu, Y., Ding, H., Men, K., Yang, K., Zhang, S., Deng, X., Zeng, A., Du, Z., Zhang, C., Shen, S., Zhang, T., Su, Y., Sun, H., Huang, M., Dong, Y., and Tang, J., “AgentBench: Evaluating LLMs as Agents,” *International Conference on Learning Representations (ICLR)*, 2024.
- Mohammadi, M., Li, Y., Lo, J., and Yip, W., “Evaluation and Benchmarking of LLM Agents: A Survey,” *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, vol. 2, pp. 6129–6139, 2025. <https://doi.org/10.1145/3711896.3736570>.
- Novelli, C., Taddeo, M., and Floridi, L., “Accountability in artificial intelligence: what it is and how it works,” *AI & Society*, vol. 39, pp. 1871–1882, 2024. <https://doi.org/10.1007/s00146-023-01635-y>.
- Tabassi, E., *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST Trustworthy and Responsible AI, National Institute of Standards and Technology, Gaithersburg, MD, [online], 2023. <https://doi.org/10.6028/NIST.AI.100-1>, [https://tsapps.nist.gov/publication/get\\_pdf.cfm?pub\\_id=936225](https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=936225) (Accessed August 18, 2026).
- Tallam, K., “Authorization Propagation in Multi-Agent AI Systems: Identity Governance as Infrastructure,” *arXiv preprint arXiv:2605.05440*, 2026. <https://doi.org/10.48550/arXiv.2605.05440>.
- Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Awadallah, A., White, R. W., Burger, D., and Wang, C., “AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation,” *Proceedings of the Conference on Language Modeling (COLM)*, 2024.
- Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., and Cao, Y., “ReAct: Synergizing Reasoning and Acting in Language Models,” *International Conference on Learning Representations (ICLR)*, 2023.
- Yehudai, A., Eden, L., Li, A., Uziel, G., Zhao, Y., Bar-Haim, R., Cohan, A., and Shmueli-Scheuer, M., “Survey on Evaluation of LLM-based Agents,” *arXiv preprint arXiv:2503.16416*, 2025. <https://doi.org/10.48550/arXiv.2503.16416>.
- Zhang, H., Huang, J., Mei, K., Yao, Y., Wang, Z., Zhan, C., Wang, H., and Zhang, Y., “Agent Security Bench (ASB): Formalizing and Benchmarking Attacks and Defenses in LLM-based Agents,” *International Conference on Learning Representations (ICLR)*, 2025. <https://doi.org/10.48550/arXiv.2410.02644>.
- Zhang, Z., Dai, Q., Bo, X., Ma, C., Li, R., Chen, X., Zhu, J., Dong, Z., and Wen, J.-R., “A Survey on the Memory Mechanism of Large Language Model-based Agents,” *ACM Transactions on Information Systems*, vol. 43, no. 6, Article 155, pp. 1–47, 2025. <https://doi.org/10.1145/3748302>.
- Zheng, H., Dong, Q., Depena, R. K., Bhatia, J. D., Xiao, F., and Xu, P., “Separating Capability from Permission: A Governance Framework for Agentic AI Autonomy Levels,” *arXiv preprint arXiv:2607.23438*, 2026. <https://doi.org/10.48550/arXiv.2607.23438>.
- Zhong, W., Guo, L., Gao, Q., Ye, H., and Wang, Y., “MemoryBank: Enhancing Large Language Models with Long-Term Memory,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 17, pp. 19724–19731, 2024. <https://doi.org/10.1609/aaai.v38i17.29946>.



Publisher's note: Eurasia Academic Publishing Group (EAPG) and AAKS remain neutral about jurisdictional claims in published maps and institutional affiliations.

**Open Access.** This article is licensed under a Creative Commons Attribution-NoDerivatives 4.0 International (CC BY-ND 4.0) licence, which permits copy and redistribute the material in any medium or format for any purpose, even commercially. The licensor cannot revoke these freedoms as long as you follow the licence terms. Under the following terms you must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorsed you or your use. If you remix, transform, or build upon the material, you may not distribute the modified material. To view a copy of this license, visit <https://creativecommons.org/licenses/by-nd/4.0/>.