Universal Library of Engineering Technology Research Article

ISSN: 3064-996X | Volume 2, Issue 4

Open Access | PP: 106-114

DOI: https://doi.org/10.70315/uloap.ulete.2025.0204018

Production-Grade LLM Orchestration for Revenue-Critical Workflows in Regulated Industries

Vadym Shashkov

Co-Founder & Chief Technology Officer (CTO), Superagent AI Inc., San Francisco, CA, USA.

Abstract

Insurance distribution is a large, agent-mediated commercial system with persistent workforce-replacement and knowledge-transfer constraints. General-purpose AI can assist discrete tasks, but production use in U.S. insurance distribution additionally requires domain knowledge, state-sensitive controls, carrier-system interoperability, auditable workflow execution, and tenant data separation. This study examines a production-grade multi-agent large language model orchestration platform across a multi-agency U.S. deployment base. The research combines literature review, comparative capability analysis, architecture decomposition, public product chronology, confidential operational records, and a client attestation. The platform is analyzed as six operational agents—Inbound, Outbound, Training, Live Call, Retention, and Quoting—coordinated through a shared orchestration and context layer and supported by common analytics, compliance, integration, security, and data-governance services. In a three-week, 203-call client pilot, average producer ramp-up decreased from seven to nine months to approximately 2.5 weeks, while the newest producer's win rate increased from 28% to 69%, an approximately 146% relative increase. The study proposes a Compliance-Autonomous Deployment (CAD) model that links domain initialization, deterministic controls, staged integration, controlled improvement, and isolation-first security. The findings indicate that production value in regulated workflows depends less on the base model alone than on orchestration, evidence, permissions, evaluation, and operational integration.

Keywords: *LLM Orchestration, Multi-Agent Systems, Autonomous AI, Insurance Distribution, Domain-Specific Fine-Tuning, Compliance-By-Design, Closed-Loop Learning, Multi-Tenant Security, Insurtech, Production AI.*

INTRODUCTION

Total global insurance premiums reached USD 7.799 trillion in 2024, approximately USD 7.8 trillion [1]. Bain & Company projects that the global premium pool could approach USD 10 trillion by 2030 as the range of insurable risks expands and insurers add prevention and service components to conventional protection [2]. In the United States, independent agencies placed 61.5% of property-and-casualty premiums and 87.2% of commercial-lines premiums in 2024; total direct written premiums reached USD 1.05 trillion [3].

Workforce data show both demand and replacement pressure. The Urban Institute's Finance and Insurance Industry Profile reports that the insurance sector was projected to lose approximately 400,000 workers to retirement and attrition by 2026 [4]. The Bureau of Labor Statistics projects approximately 47,000 insurance sales-agent openings per year over 2024–2034, many from occupational transfers and labor-force exits, and reports a median annual wage of USD 60,370 in May 2024 [5]. Jacobson Group and Aon data for Q3 2024 found that 52% of carriers planned to increase staff, 34% planned to maintain headcount, and 14% planned reductions; Q1 2025 findings showed 88% planned to increase or maintain staffing and 81% were most likely to recruit experienced personnel [6, 7].

Prior technology interventions in this domain have fallen short for three reasons. First, insurance sales are regulated state by state, with each jurisdiction imposing distinct licensing requirements, required disclosures, and carrier-approved language; systems not designed to encode this variation produce non-compliant outputs. Second, the carrier portal ecosystem consists of dozens of independent systems that were not built for third-party programmatic access, making automated quoting technically non-trivial. Third, insurance sales conversations require the kind of contextual judgment, objection handling, and tonal calibration that general-purpose language models cannot reliably produce at the level required for live client interactions involving licensed advice. As a result, the three broad categories of AI currently adjacent to insurance distribution, namely

Citation: Vadym Shashkov, "Production-Grade LLM Orchestration for Revenue-Critical Workflows in Regulated Industries", Universal Library of Engineering Technology, 2025; 2(4): 106-114. DOI: https://doi.org/10.70315/uloap.ulete.2025.0204018.

customer experience and back-office automation tools, sales CRM and enablement platforms that assist human agents, and industry-agnostic multi-agent platforms, each address some part of the problem without addressing all of it simultaneously.

Insurance labor-market statistics also reveal the conditions under which industry-specific competencies are formed. Insurance knowledge develops through the study of products, familiarity with carrier procedures, compliance with established communication protocols, and repeated interaction with clients. Traditional personnel onboarding therefore requires considerable time and continuous supervision by experienced employees. Artificial intelligence may partially reduce this burden. The engineering challenge involved in constructing a production-grade system, however, extends well beyond the generation of coherent text. Such a system must coordinate multiple operations, preserve process state, regulate the use of tools, comply with established rules, record evidentiary information, and maintain strict separation between client datasets.

Production agentic systems must do more than generate coherent text. They must decompose and route tasks, preserve state, control tool permissions, incorporate changing knowledge from governed sources, log material events, isolate tenant data, and recover safely from failure. Research on LLM-based multi-agent systems identifies role allocation, coordination, memory, and evaluation as central design dimensions [8]. MARCO demonstrates that specialized real-time orchestration can achieve high task-completion accuracy and reduce latency and cost in evaluated retail and restaurant dialogues, but it does not establish incident-response parameters or zero-variance behavior in regulated sales [9].

A separate distinction is required between mechanisms for incorporating knowledge and instruments for governing system behavior. Ovadia et al. found that retrieval-augmented generation consistently outperformed uncontrolled fine-tuning when factual information, including recently introduced knowledge, had to be incorporated into model outputs [10]. In regulated production environments, changing factual and procedural materials are therefore more appropriately obtained through controlled retrieval from versioned sources. Tone, response format, operational sequencing, and workflow rules may instead be adjusted through instructions, evaluation procedures, controlled adaptation, and other governable methods. The appropriate combination depends on technical conditions and empirical results; consequently, no single implementation pattern can be regarded as universally applicable across all systems.

The scientific gap this article addresses is the absence of published, empirically grounded analysis of production-grade LLM orchestration in environments where regulatory compliance is mandatory, carrier system integration is non-standard, and outcomes are measured in direct revenue. **This study contributes a novel conceptual** framework for compliance-by-design autonomous AI deployment in regulated industries, distinguishing it from general-purpose orchestration approaches through specific architectural mechanisms and verified outcome metrics.

The research objective is to analyze the architectural requirements, implementation mechanisms, and measurable outcomes of a domain-specific, autonomously orchestrated LLM system deployed in U.S. insurance distribution, and to propose an original framework for replicating this approach in other regulated industries.

The scientific novelty of this study lies in the first documented synthesis of domain-specific LLM fine-tuning, compliance-by-design architecture, closed-loop continual learning, and multi-tenant security in a single production deployment with verified outcome data from regulated insurance sales workflows.

This study has several limitations worth noting at the outset. The case data are drawn from a single platform currently in production, which limits cross-vendor generalizability. The study covers the U.S. insurance distribution channel, which has specific regulatory characteristics that differ from other jurisdictions. Outcome metrics such as conversion lift and ramp-up time reduction derive from live deployment data and agency-reported figures rather than randomized controlled trial settings. Finally, the closed-loop fine-tuning cycle is ongoing, and performance metrics may continue to compound over time beyond what this analysis captures.

For the U.S. technology sector and national economic competitiveness, findings of this kind carry practical relevance beyond the insurance vertical. The United States currently leads in AI-enabled enterprise software development, and demonstrating that LLM orchestration can be deployed safely and effectively in regulated industries directly supports the case for domestic AI adoption in financial services, healthcare, and legal sectors. The professional trajectory represented by this work, building autonomous AI systems that expand economic capacity in industries facing demographic constraints, illustrates how domain expertise combined with AI engineering can address structural economic problems rather than merely optimizing existing processes.

MATERIALS AND METHODS

The research methodology is based on a mixed-method thematic approach combining a targeted review of peer-reviewed scholarly publications and official materials, a structured comparison of adjacent artificial intelligence categories, a technical decomposition of an operational platform, and an examination of documented operational outcomes. This research design is appropriate to the question under consideration because it permits architectural characteristics and production performance indicators to be examined concurrently. The available performance evidence,

however, is predominantly observational in nature and was not generated through a randomized experiment.

Before the materials were synthesized, a hierarchy of evidence was established. General analytical propositions were grounded in peer-reviewed publications, official regulatory documents, and industry reports. Public-company materials and independent publications were used solely to describe product composition and functional capabilities disclosed in the public domain. Confidential client testimonials, production telemetry, a penetration-test report prepared by an independent third party, and SOC 2 audit-engagement correspondence were treated as case-specific materials, with their unpublished or confidential status expressly identified.

The comparative taxonomy encompasses three categories of external analogues together with the platform examined in this study. It includes customer-engagement and back-office automation solutions, CRM systems and specialist-support tools, general-purpose agentic platforms, and the specialized orchestrated platform that constitutes the subject of the present research. The comparison was conducted across seven criteria: industry-specific safeguards, integration into operator workflows, accommodation of jurisdictional requirements, controlled improvement mechanisms, isolation of client environments, end-to-end workflow coverage, and the availability of production-grade voice or other communication capabilities.

The empirical unit of analysis is SUPERAGENT AI, a production platform that operates across a multi-agency deployment base in the United States. Public materials document the platform's evolution from an autonomous-agent roadmap to the launch of the Inbound and Outbound AI Agents, the Retention AI Agent, the Quoting AI Agent, and the unified SUPERAGENT platform [13–17]. To ensure analytical consistency, the study considers six operational roles: inbound, outbound, training, direct calling, retention, and quotation, while analytics, compliance, customer relationship management connectivity, integration, security, and management are considered as shared services.

The RightSure pilot project began implementation in October 2025. Over the course of three weeks, during which 203 training calls were conducted, the percentage of successful deals for new employees increased from 28% to 69%. The average time required for newly hired sales professionals to reach operational productivity decreased from seven to nine months to approximately two and a half weeks [17, 18].

The technical examination covered the orchestration layer, the routing of states and contextual information, the allocation of specialized roles, the management of retrieval processes and industry knowledge, the application of deterministic workflow rules, integration with operator systems, the generation of audit logs, access segregation within individual client environments, and procedures governing model and workflow updates.

The security and regulatory-compliance assessment was confined to the scope of the available evidence. In 2025, the architecture underwent an independent penetration test performed by third-party specialists. The legal analysis addresses architectural characteristics and does not offer conclusions concerning the permissibility of particular solutions from the standpoint of any specific regulatory authority. The platform is considered in relation to the relevant groups of obligations, including insurance licensing and disclosure requirements, state-level insurance data-protection rules, the provisions of NYDFS Part 500 where applicable, and the confidentiality and information-security obligations established under the GLBA [24–27].

RESULTS AND DISCUSSION

The U.S. insurance distribution market presents a set of structural characteristics that make it simultaneously one of the most valuable targets for AI-driven automation and one of the most resistant to general-purpose AI deployment. The market size and workforce dynamics are summarized here to ground the subsequent architectural and outcome analysis.

Table 1 compares the three external categories with the subject platform across seven capability dimensions. The entries are qualitative because a common, independently validated quantitative benchmark does not exist across all four categories. "Yes" for the subject platform means that the capability is documented in public materials or confidential production evidence; it does not imply independent market-wide superiority.

Table 1. Qualitative capability comparison for AI in U.S. insurance distribution (compiled by the author based on [6, 9, 13, 14]).

Capability dimension	Customer-experience / back-office AI	CRM / human-agent enablement	Industry-agnostic agent frameworks	Domain-specific orchestrated platform
Insurance-domain grounding	Partial	Partial	Not built in	Documented
Carrier workflow / portal integration	Limited	Usually human-mediated	Requires custom implementation	Documented
Jurisdiction-sensitive controls	Workflow dependent	Workflow dependent	Requires custom implementation	Rule and workflow controls documented
Controlled improvement loop	Varies	Varies	Framework dependent	Documented evaluation and update loop

Tenant data separation	Vendor dependent	Vendor dependent	Implementation dependent	Tenant-scoped controls documented
End-to-end workflow coverage	Limited workflows	Human remains execution layer	Technically possible; domain work required	Multiple specialized workflows
Production voice / communication support	Often narrow	Assistance-oriented	Framework dependent	Documented voice and communication agents

The comparison indicates that the production challenge is combinatorial. Customer-experience tools can automate bounded interactions but may not execute carrier-specific workflows. CRM systems can improve visibility and coaching while leaving the human producer as the primary execution layer. General agent frameworks supply orchestration primitives but require domain data, rules, integrations, testing, and security design before regulated production use.

A separate distinction is required between factual knowledge and behavioral adaptation. Retrieval from versioned sources is generally better suited to changing facts and procedures, while controlled parameter adaptation can be used for stable domain language and response patterns when supported by evaluation and rollback. Ovadia et al. provide evidence on retrieval versus fine-tuning for knowledge injection, and LoRA provides a parameter-efficient adaptation method; neither source alone proves that fine-tuned compliance behavior is immune to adversarial drift [10, 11]. Prompt-injection research reinforces the need for tool isolation, source controls, and authorization boundaries around model outputs [12].

Figure 1 shows the architectural diagram of the six-agent system developed by the author, which includes an orchestration layer, agent specialization, a shared context store, and a closed feedback loop for fine-tuning.

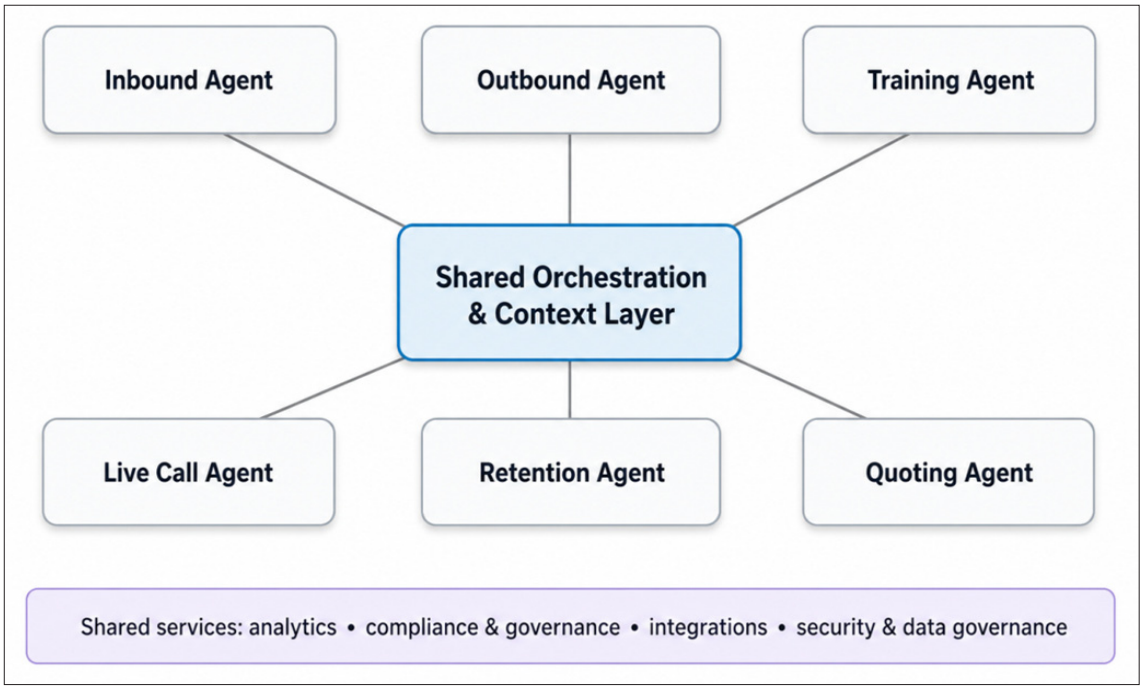

Figure 1. Six-agent production architecture with shared orchestration and platform controls (compiled by the author based on [13-17]).

The architecture assigns discrete operational scopes to six agents while using a shared orchestration layer for routing, context, state, and policy enforcement. The Inbound and Outbound agents handle lead conversations and scheduled outreach. The Training agent delivers structured simulations and performance feedback. The Live Call agent provides real-time assistance during producer calls. The Retention agent coordinates renewal-related outreach without a quantified outcome claim in the present evidence base. The Quoting agent captures structured risk data and interacts with carrier forms.

Public product materials report support for more than 50 carrier forms and approximately 30-second form submission, compared with a manual baseline described as 20–45 minutes depending on the workflow measured. Using the narrower 20–30-minute baseline applied in the case documentation, 30 seconds represents approximately a 40–60-fold acceleration, or a time reduction exceeding 97%. These values are company-reported product metrics, not an independently benchmarked market average.

Domain knowledge is incorporated through two governed paths. Versioned retrieval supplies current product, carrier, and procedural information. Controlled adaptation may encode stable vocabulary, conversational structure, and task patterns, including parameter-efficient techniques such as

LoRA, but changes require evaluation, approval, and rollback criteria [10, 11]. This separation avoids treating model weights as the source of truth for frequently changing facts.

The regulatory environment for insurance distribution in the United States imposes requirements that generic AI architectures cannot satisfy at the point of sale. Each of the 50 states maintains its own licensing requirements, approved carrier forms, and mandatory disclosure rules. An AI system operating in this environment must enforce these rules in real time, at the sentence level, without relying on human review. This requirement fundamentally shapes the architectural design.

Figure 2 illustrates the compliance-by-design stack developed by the author, showing the five functional layers from customer interaction through to carrier system integration and the data flow between them.

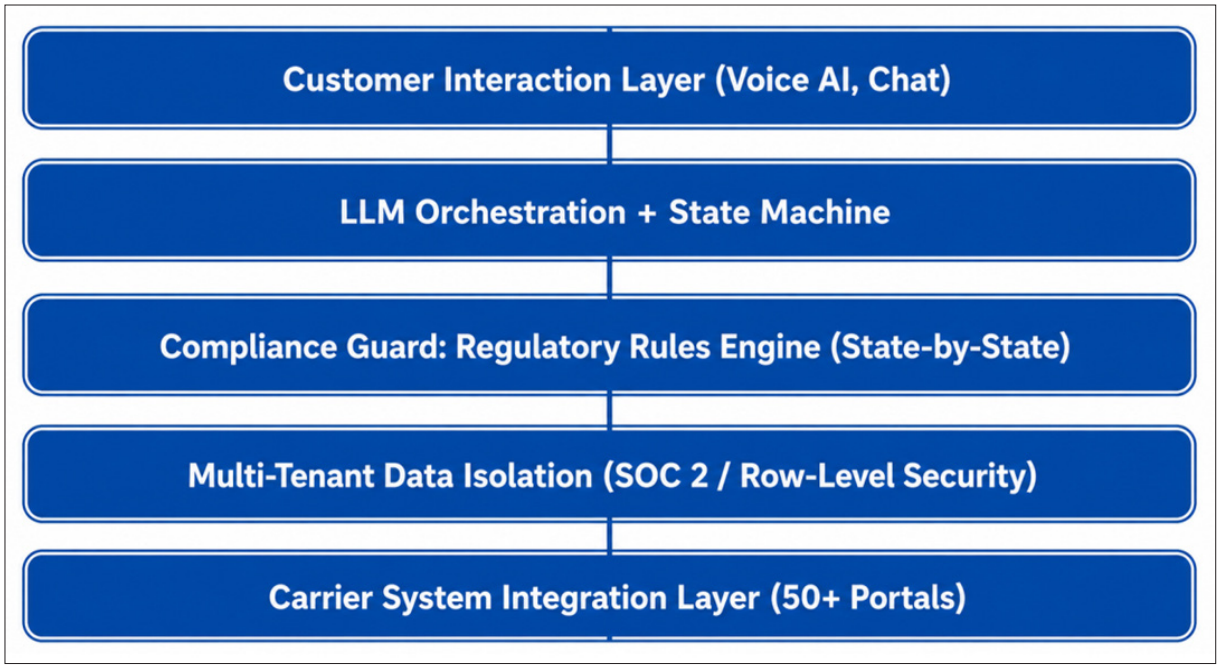

Figure 2. Compliance-by-design deployment stack for regulated insurance AI, showing mandatory functional layers from customer interface to carrier integration (developed by the author based on [11, 12, 16]).

PII collection and carrier-system integration typically add data-processing agreements, security review, and carrier approval steps that delay deployment. A staged deployment can begin with training or assistance functions that do not require customer PII, then add integrations after the relevant approvals are completed. This sequencing does not create a regulatory exemption; it reduces the initial scope of data and system access [24, 25, 26, 27].

The multi-tenant design uses tenant-scoped access controls, routing, secrets, and session-context handling to reduce cross-tenant access risk. The architecture underwent independent third-party penetration testing in 2025 [19]. Table 2 provides a structured comparison of production performance across four AI architectural tiers, from single-agent baseline through domain-specific multi-agent system, across six operational dimensions relevant to insurance sales. Data are drawn from platform telemetry, agency reporting, and publicly available insurtech benchmarks.

Table 2. Comparison across six operational dimensions (compiled by the author based on [19-21]).

Architecture Component	Single-Agent (Baseline)	RAG-Enhanced Single Agent	General MAS Platform	Domain-Tuned MAS (Production)
Quote retrieval speed	20-30 min (human)	8-12 min	4-6 min	~30 sec
Carrier portal coverage	Manually configured	N/A	Limited	50+ portals
Compliance guardrail enforcement	Human review	Prompt-based	Partial	Rule engine + LLM
New-agent ramp-up time	7-9 months	7-9 months	5-7 months	~2.5 weeks
Closed-loop fine-tuning	No	No	No	Yes

The RightSure insurance agency provides the most fully documented case for outcome validation. Before deployment of the Superagent AI Training Agent and Live-Coaching Agent, the agency's average new-agent ramp-up time was approximately seven to nine months, consistent with industry norms documented by the Jacobson Group [4].

Figure 3 presents a side-by-side visualization of the key performance outcomes observed at RightSure, comparing pre-deployment baselines with post-deployment results on the two documented metrics.

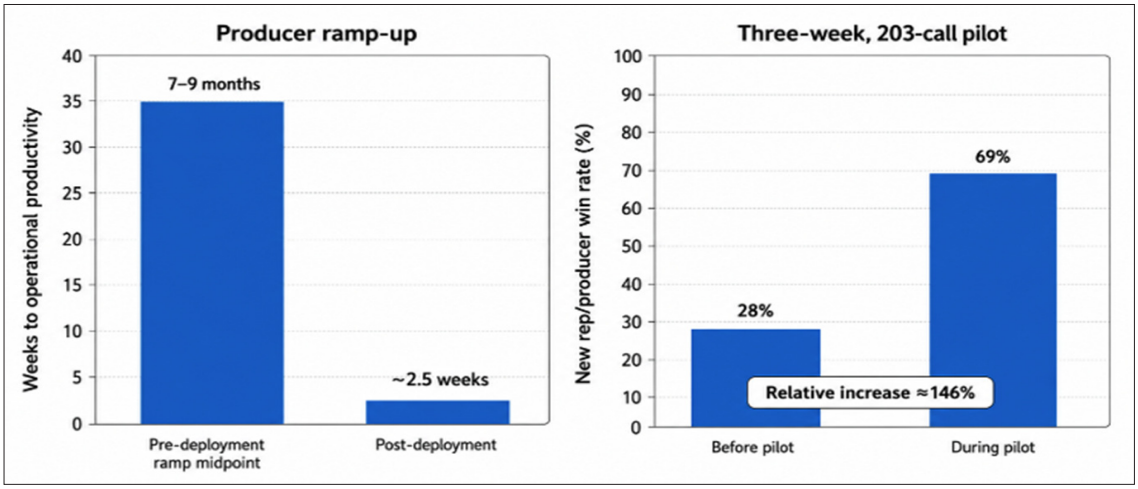

Figure 3. RightSure pilot outcomes on the two documented metrics (compiled by the author based on agency-reported data [17, 18]).

Two mechanisms explain the magnitude of the conversion improvement. First, the Live-Coaching Agent provides real-time objection handling and product recommendation prompts during live calls, which directly addresses the knowledge deficit that characterizes newly licensed agents in their first months of production. Second, the Training Agent compresses what would otherwise be months of scenario-based learning into structured simulations drawn from the actual call corpus, which means new agents arrive at live calls with higher baseline competency. The approximately 146% relative conversion increase should be understood in this context: it reflects the compounding of improved readiness, better in-call support, and more consistent follow-up discipline enforced by the CRM agent, not a single-factor improvement.

A comparative reading of Table 1 and the findings of the illustrative analysis supports the central argument: in regulated industries, the competitive advantage of production-grade AI does not derive primarily from the quality of the underlying language model. Rather, it emerges from the combined effect of domain-specific fine-tuning, a robust regulatory and governance architecture, and seamless integration with operator workflows and enterprise systems. Together, these elements enable greater AI autonomy while ensuring that its deployment remains operationally effective, legally compliant, and reliable within the relevant regulatory context.

Figure 4 presents an author-developed positioning matrix for AI solutions relevant to insurance distribution. The matrix maps representative products along two dimensions: domain specificity, ranging from general-purpose to domain-specific functionality, and mode of execution, ranging from human-assisted operation to high execution autonomy. ChatGPT and Claude are positioned as relatively autonomous but general-purpose systems, whereas Lemonade Maya represents a generalist, human-assisted solution. Grounded CRM tools occupy the domain-specific but assisted quadrant. By contrast, Superagent AI is located in the upper-right target zone, combining high domain specificity with a high degree of execution autonomy. The figure therefore indicates that this strategically important quadrant remains comparatively underdeveloped and is occupied primarily by purpose-built solutions capable of integrating industry-specific knowledge, regulatory controls, and autonomous workflow execution.

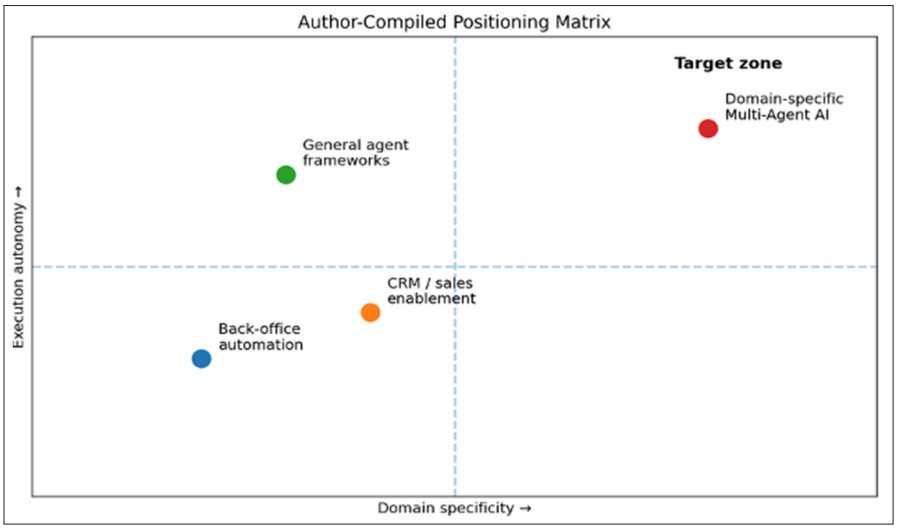

Figure 4. Author-compiled qualitative positioning matrix for insurance AI categories. The matrix is an analytical assessment and does not imply independently measured market-wide superiority.

This positioning analysis leads to an original framework for regulated AI deployment that the present study proposes as a contribution to both research and practice. The author terms this framework the Compliance-Autonomous Deployment (CAD) model, defined by five conditions that must be satisfied simultaneously for a multi-agent LLM system to operate in a revenue-critical, regulated workflow without human oversight:

(1) Domain initialization: the base model must be fine-tuned on proprietary domain data to a level that embeds regulatory vocabulary, tone calibration, and product logic as model weights rather than prompt instructions. (2) Compliance architecture: a deterministic rule engine must sit between LLM outputs and customer-facing responses, enforcing disclosures and escalation triggers regardless of conversation direction. (3) Carrier or system integration: the autonomous system must interface natively with the data systems that govern the regulated activity, in this case, carrier portals, without requiring human mediation for each transaction. (4) Closed-loop improvement: every interaction must feed back into a quality-filtered fine-tuning pipeline so that the system improves continuously in proportion to deployment volume. (5) Isolation-first security: multi-tenant data handling must implement physical or cryptographic isolation at every layer where tenant data enters the serving pipeline, not merely at the application layer.

Table 3 maps each CAD model principle to the specific technical mechanism that implements it in the Superagent AI platform, the regulatory outcome that mechanism produces, and an observable metric by which that outcome can be verified. This mapping illustrates that the CAD model is not a theoretical construct but an operational specification.

Table 3. Compliance-Autonomous Deployment (CAD) model: principles, mechanisms, and verifiable outcomes (developed by the author based on [9, 12, 15, 17, 18]).

CAD principle	Technical mechanism	Intended outcome	Verification evidence
Minimal-data, compliance-aware onboarding	Minimize PII at initial configuration; stage integrations and access approvals	Reduce initial privacy, security, and carrier-integration burden	Approximately one-week enterprise setup in internal onboarding material
Deterministic compliance controls	Rule checks, mandatory disclosure injection, authorization and escalation gates	Support licensing, disclosure, and supervision requirements	Versioned test cases and sampled audit-log review; no unsupported 100% claim
Multi-tenant isolation	Tenant-scoped access, encryption, routing, secrets, and session-context controls	Reduce cross-tenant access risk	Independent third-party penetration test completed in 2025
Controlled interaction audit	Versioned transcripts and event metadata for material workflow steps	Traceability, investigation support, and E&O documentation	Sampled records showing source, rule, tool, version, and escalation path
Domain knowledge and evaluated adaptation	Versioned retrieval sources, workflow examples, evaluations, and controlled model/prompt/tool updates	Improve domain consistency without conflating facts with behavior	Evaluation sets, change records, and rollback criteria

The practical recommendation that follows from the CAD model is that organizations seeking to deploy autonomous AI in regulated industries should treat compliance architecture as an equal-priority design constraint alongside performance optimization, not as a post-deployment compliance review. The evidence from the insurance deployment is that compliance-by-design, when implemented as described, eliminates the adoption delay created by pre-integration regulatory review, because the architecture is designed to operate within existing licensing, disclosure, and supervision requirements.

A further recommendation for technology architects is that the closed-loop fine-tuning pipeline, often treated as a performance optimization mechanism, should be recognized as a regulatory documentation asset. Every interaction in the pipeline constitutes an auditable record of model behavior, training data lineage, and output quality filtering. In environments subject to insurance department examination or errors-and-omissions litigation, this audit trail provides the documentation that would otherwise require manual review. Treating the fine-tuning pipeline as both a performance asset and a compliance asset changes its design priority from a periodic batch process to a continuous, audit-ready system.

For the broader research community, the insurance case offers a template for evaluating autonomous AI deployment in other regulated industries. Healthcare prior authorization workflows, mortgage origination, and legal document execution all share the structural characteristics that made general-purpose AI insufficient in insurance: state or federal regulatory variance, existing system fragmentation, and the requirement that AI-generated outputs meet the same legal standard as human-generated ones. The CAD model framework provides a generalizable design specification for approaching these contexts, with the insurance production evidence serving as proof of concept.

CONCLUSION

This study examined the architecture, implementation mechanisms, and measurable outcomes of a production-grade, multi-agent LLM orchestration system deployed in U.S. insurance distribution, a domain characterized by regulatory complexity, workforce shortages, and fragmented carrier infrastructure that has consistently defeated general-purpose AI deployment. The research objective was to analyze the technical requirements for autonomous AI deployment in a regulated, revenue-critical context and to propose a replicable framework based on verified deployment data.

The RightSure case provides observational evidence on two outcomes: producer ramp-up decreased from seven to nine months to approximately 2.5 weeks, and the newest producer's win rate increased from 28% to 69% during a three-week, 203-call pilot, an approximately 146% relative increase. The article does not claim a client-verified retention effect. The CAD model translates the findings into five implementation principles that can be tested in other regulated workflows. Future research should use pre-registered definitions, longer observation windows, multi-client samples, independent replication, and explicit measurement of error, escalation, and compliance outcomes.

The Compliance-Autonomous Deployment (CAD) model represents the original contribution of this study to both scientific research and practice. The five principles of the CAD model, domain initialization, compliance architecture, system integration, closed-loop improvement, and isolation-first security, define the conditions under which autonomous LLM orchestration is viable in a licensed, regulated distribution context. The mapping of these principles to specific technical mechanisms and verifiable regulatory outcomes in Table 3 provides practitioners with an operational specification rather than a conceptual outline.

The practical significance of these findings for the U.S. insurance market is material. An industry unable to close the productivity gap through conventional training pipelines has a demonstrated technological path to maintaining distribution capacity without proportional workforce growth. At an economic scale, this means that a sector managing $7.8 trillion in annual premium volume can sustain and potentially expand output per licensed agent, which has implications for insurance penetration, consumer access, and the competitiveness of the U.S. distribution model against markets where AI-native insurance platforms are emerging.

Future research should extend this framework to other regulated industries with analogous structural characteristics, including healthcare prior authorization and legal document execution, to test whether the CAD model generalizes beyond insurance. Controlled longitudinal measurement of the closed-loop fine-tuning effect across deployment cohorts of different sizes would enable more precise quantification of the performance compounding rate. Finally, examination of state-level regulatory responses to fully autonomous AI in licensed sales channels would provide valuable context for scaling the model across all 50 U.S. jurisdictions.

REFERENCES

1. Swiss Re Institute. (2025, July 9). World insurance in 2025: A riskier, more fragmented world order (sigma No. 2/2025). https://www.swissre.com/institute/research/sigma-research/sigma-2025-02-world-insurance-riskier-fragmented-world.html
2. Schwedel, A., Judah, M., & Goossens, C. (2021, November 8). The future of insurance: As risks mount, insurers aim to augment protection with prevention. Bain & Company. https://www.bain.com/insights/the-future-of-insurance-as-risks-mount-insurers-aim-to-augment-protection-with-prevention/
3. Independent Insurance Agents & Brokers of America. (2025, July 17). Big "I" releases 2025 Market Share Report. https://www.independentagent.com/news/big-i-releases-2025-market-share-report/
4. Urban Institute. (2025, February). Finance and insurance industry profile: National Occupational Frameworks. https://apprenticeships.urban.org/sites/default/files/2025-02/Finance_and_Insurance_Industry_Profile_National_Occupational_Frameworks_2.25.pdf
5. U.S. Bureau of Labor Statistics. (2025). Insurance sales agents: Occupational Outlook Handbook, 2024-2034 projections. https://www.bls.gov/ooh/sales/insurance-sales-agents.htm
6. The Jacobson Group. (2024, August 20). Q3 2024 Insurance Labor Market Study: Recruiting difficulty eases slightly. https://www.jacobsononline.com/about-us/press-releases/q3-2024-insurance-labor-market-study-recruiting-difficulty-eases-slightly/
7. The Jacobson Group. (2025, April 7). Study reflects ongoing labor market stability for 2025. https://www.jacobsononline.com/blog/study-reflects-ongoing-labor-market-stability-for-2025/
8. He, J., Treude, C., & Lo, D. (2025). LLM-based multi-agent systems for software engineering: Literature review, vision, and the road ahead. ACM Transactions on Software Engineering and Methodology, 34(5), Article 124. https://doi.org/10.1145/3712003
9. Shrimal, A., Kanagaraj, S., Biswas, K., Raghuraman, S., Nediyanchath, A., Zhang, Y., & Yenigalla, P. (2024). MARCO: Multi-agent real-time chat orchestration. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track, 1381–1392. https://doi.org/10.18653/v1/2024.emnlp-industry.102
10. Ovadia, O., Brief, M., Mishaeli, M., & Elisha, O. (2024). Fine-tuning or retrieval? Comparing knowledge injection in LLMs. Proceedings of the 2024 Conference on Empirical

Methods in Natural Language Processing, 237–250. https://doi.org/10.18653/v1/2024.emnlp-main.15

11. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. International Conference on Learning Representations. https://arxiv.org/abs/2106.09685
12. Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., & Fritz, M. (2023). Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security, 79–90. https://doi.org/10.1145/3605764.3623985
13. Hemenway, C. (2025, August 13). Insurtech looks to "redefine" insurance with autonomous AI agents by 2026. Insurance Journal. https://www.insurancejournal.com/news/national/2025/08/13/835548.htm
14. Federal Trade Commission. Gramm-Leach-Bliley Act and Safeguards Rule resources. https://www.ftc.gov/business-guidance/privacy-security/gramm-leach-bliley-act
15. SUPERAGENT AI. (2025, October 14). The End of the Insurance Cold Call: SUPERAGENT AI Deploys Autonomous AI Agents That Dial, Qualify, Book, and Quote 24/7.
16. Gort, B., Kibalya, G. M., & Antonopoulos, A. (2025). Agentedge: Agentic ai for service orchestration in the edge-cloud continuum.
17. Debenedetti, E., Zhang, J., Balunović, M., Beurer-Kellner, L., Fischer, M., & Tramèr, F. (2024). AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents. Advances in Neural Information Processing Systems, 37, 82895–82920. https://proceedings.neurips.cc/paper_files/paper/2024/hash/97091a5177d8dc64b1da8bf3e1f6fb54-Abstract-Datasets_and_Benchmarks_Track.html (accessed 12 June 2025).
18. RightSure Insurance Group. (2025, November). Client attestation concerning producer ramp-up and pilot conversion outcomes [Confidential document on file with the author].
19. Independent third-party security assessor. (2025). Penetration-test report for the platform studied [Confidential report on file with the author].
20. Shethiya, A. S. (2023). LLM-Powered Architectures: Designing the Next Generation of Intelligent Software Systems. Academia Nexus Journal, 2(1).
21. Julian, V., & Botti, V. (2019). Multi-agent systems. Applied Sciences, 9(7), 1402.
22. Liu, Y., Lo, S. K., Lu, Q., Zhu, L., Zhao, D., Xu, X., Harrer, S., & Whittle, J. (2025). Agent design pattern catalogue: A collection of architectural patterns for foundation-model-based agents. Journal of Systems and Software, 220, 112278. https://doi.org/10.1016/j.jss.2024.112278
23. Yao Y. et al. Toward super agent system with hybrid ai routers //arXiv preprint arXiv:2504.10519. – 2025.
24. National Association of Insurance Commissioners. (2023). Model Bulletin on the Use of Artificial Intelligence Systems by Insurers. https://content.naic.org/insurance-topics/artificial-intelligence
25. National Association of Insurance Commissioners. (2017). Insurance Data Security Model Law (Model 668). https://content.naic.org/sites/default/files/inline-files/MDL-668.pdf
26. New York State Department of Financial Services. (2024). Cybersecurity Regulation, 23 NYCRR Part 500. https://www.dfs.ny.gov/cybersecurity
27. Federal Trade Commission. (2024). Gramm-Leach-Bliley Act and Safeguards Rule resources. https://www.ftc.gov/business-guidance/privacy-security/gramm-leach-bliley-act