International Journal of Applied Mathematics

Volume 38 No. 12s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

CLOUD-NATIVE ORCHESTRATION PATTERNS FOR MULTI-AGENT HEALTHCARE LLMS

Vallikranth Ayyagari*[1], Sai Rupesh Kagga[2],Shalmali Joshi[3],Guru Lakshmi Priyanka Bodagala[4]

*Corresponding author email: vallikranth@gmail.com

[1] DaVita Inc., United States.

ORCID: 0009-0007-9341-7727

[2] SanQuest Inc., United States.

ORCID: 0009-0000-7893-3861

[3] Elevance Health, United States.

ORCID: 0009-0000-4329-3841

[4] University of San Francisco, United States.

ORCID: 0009-0002-5561-4969

Abstract

Healthcare organizations increasingly deploy multi-agent large language models (LLMs) for clinical triage, diagnostic support, and documentation. However, deployment at scale requires coordination across distributed cloud infrastructure while satisfying HIPAA, GDPR, and HL7 FHIR interoperability requirements. We evaluated three orchestration patterns—Kubernetes-native microservices, serverless Knative, and hybrid edge-cloud—for multi-agent healthcare LLM orchestration using synthetic clinical data (10,000 patients, 45,000 encounters via Synthea). Multi-cloud experiments across AWS, Azure, and GCP (15 repetitions per pattern; n=405 total) demonstrate that edge-local deployment reduces latency 45% (triage: 2.08s vs. 3.81s) and maintains 99.97% compliance adherence (±0.03%) across 1.2 million policy-controlled PHI accesses. Multi-agent diagnostic consensus improved accuracy from 82.8% (single-agent) to 87.1% (ANOVA: $F(2,403)=18.4$, $p<0.001$, $\eta^2=0.20$) with greatest benefits for complex multi-system conditions. Clinical documentation ratings by three board-certified physicians ($\kappa=0.72$ inter-rater agreement) achieved 4.15/5 mean acceptability on synthetic cases, though generalization to real clinical practice requires prospective validation. We propose a reference architecture integrating Istio service mesh (mTLS, policy enforcement), Knative event-driven activation, and Kubernetes-native microservices, establishing design guidelines for healthcare organizations selecting orchestration patterns based on latency, cost, and compliance constraints. Limitations include reliance on synthetic ground truth, single-timepoint evaluation, and unknown generalization to alternative LLM architectures.

Keywords: Cloud-native orchestration, Multi-agent LLM systems, Healthcare AI, Kubernetes, HIPAA compliance, FHIR interoperability, Service mesh, Edge computing

Received: August 04, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

1. Introduction

1.1 Background and Motivation

Large language models (LLMs) have demonstrated clinical utility for documentation assistance, diagnostic support, and care coordination[1][2]. Yet deployment at organizational scale faces challenges absent in general-purpose LLM applications: multi-agent coordination across distributed systems, continuous compliance verification against HIPAA and GDPR standards, and seamless integration with healthcare IT infrastructure (EHRs, lab systems, pharmacy) via HL7 FHIR protocols[3][4].

Monolithic single-model deployments achieve only 30% GPU utilization in production healthcare environments[5], with 70% of peak-capacity resources sitting idle during normal operations. Multi-agent architectures—where specialized agents for triage, diagnostics, treatment planning, and documentation coordinate through APIs—have emerged as a promising alternative to handle variable clinical workloads[6][7].

However, existing multi-agent frameworks (AutoGen, LangChain, CrewAI) were designed for general-purpose applications without healthcare-specific capabilities: continuous compliance verification, audit trail generation with PHI tracking, and integration with FHIR-compliant EHR systems[8][9].

1.2 Research Gaps and Contributions

Literature review reveals four critical gaps:

1. **Orchestration pattern fragmentation:** Cloud-native patterns (microservices, serverless, edge) are well-established for general systems but lack systematic comparison for multi-agent healthcare LLMs under compliance constraints[5][10][11].
2. **Compliance automation deficit:** Current HIPAA/GDPR approaches rely on infrastructure controls (encryption, role-based access) without policy-driven orchestration that validates compliance at runtime and maintains comprehensive audit trails across multi-agent interactions[12][13].
3. **Interoperability integration gap:** Despite FHIR's regulatory endorsement (CMS/ONC 2020), orchestration frameworks lack native patterns for multi-agent systems to consume and produce FHIR-compliant data while coordinating through service mesh architectures[14].
4. **Performance-compliance trade-off uncertainty:** The latency and throughput implications of embedding compliance verification, audit logging, and zero-trust networking into multi-agent orchestration remain empirically unquantified[5][12].

Our contributions:

- **Reference architecture:** A Kubernetes-native orchestration framework integrating Istio service mesh (mTLS, policy enforcement), Knative serverless activation, and policy-driven compliance modules aligned with HIPAA §164.312(b) and GDPR Article 32 requirements.

Received: August 04, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

- **Comparative evaluation:** Empirical comparison of three orchestration patterns across 15 standardized workload scenarios on multi-cloud testbeds (AWS/Azure/GCP), with 95% confidence intervals and ANOVA validation.
- **Design guidelines:** Context-specific recommendations for healthcare organizations selecting orchestration patterns based on deployment constraints (latency requirements, cost sensitivity, compliance jurisdiction, hybrid vs. cloud-only).
- **Limitations and boundaries:** Explicit specification of synthetic data constraints, generalization limits, and validation gaps requiring prospective clinical study.

1.3 Paper Organization

Sections 2–4 detail the experimental design, orchestration patterns, and multi-agent consensus algorithms. Section 5 reports performance, compliance, and diagnostic accuracy results. Section 6 interprets findings relative to prior work. Section 7 discusses limitations and recommendations for real clinical deployment. Section 8 concludes.

2. Methods

2.1 Experimental Design and Sampling

We conducted a 3×3×3 between-subjects experimental design:

- **Orchestration pattern** (3 levels): Kubernetes-native microservices, Knative serverless, hybrid edge-cloud
- **Cloud provider** (3 levels): AWS EKS, Azure AKS, Google GKE
- **Workload type** (3 levels): Real-time triage (100–500 concurrent requests/hour), moderate diagnostic (20–50 requests/hour), batch documentation (200–1000 requests/hour)

Sampling: 15 experimental repetitions per condition × 27 conditions = **405 total runs**. Each run processed 100 synthetic patient encounters from Synthea, thus ~40,500 total encounters across all runs though Synthea initially generated 45,000 encounters across the target cohorts.

Random seed protocol: Each repetition used distinct pseudo-random seeds (1–15) for load generation and encounter ordering to ensure independence. Traffic generated via Apache JMeter with Gaussian request inter-arrival times (μ=2.5s, σ=0.8s) to simulate realistic clinician behavior[15].

2.2 Infrastructure and Configuration

Multi-cloud testbed:

- **AWS EKS:** Regions us-east-1, eu-central-1. Kubernetes v1.28. Node pools: 3 control-plane (c5.4xlarge), 6 worker (c5.2xlarge), 2 GPU (p3.8xlarge). Network: VPC with private subnets, NAT gateway, VPN-encrypted inter-region traffic.
- **Azure AKS:** Regions ussc, westeurope. Kubernetes v1.28. Node pools: 3 control-plane (Standard_D8s_v4), 6 worker (D8s_v4), 2 GPU (Standard_NC24s_v3).

Received: August 04, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

- **Google GKE:** Regions us-central1, europe-west1. Kubernetes v1.28. Node pools: 3 control (n2-standard-8), 6 worker (n2-standard-8), 2 GPU (a2-highgpu-1g).

Service mesh: Istio v1.20 with mutual TLS (mTLS) enforcement for all pod-to-pod communication. Certificate rotation every 24 hours. We discovered and mitigated a cert-renewal conflict with scale-down windows by staggering rotation across 0–6 UTC.

Knative: Knative Serving v1.12 with concurrency target = 10 requests/pod. Min-scale = 2 for latency-critical agents (Triage, Diagnostic); scale-to-zero for non-critical. Cold-start penalty observed: initial deployment ~5s; min-scale reduced P99 latency from 8.2s to 2.1s while preserving 68% cost savings.

2.3 Multi-Agent Ensemble and Consensus Algorithm

Agent composition (See Figure 1):

- **Triage Agent:** GPT-3.5-turbo (6B parameters). Processes symptom descriptions, generates urgency score (1–5), preliminary differential diagnosis.
- **Diagnostic Agent:** GPT-4 (175B parameter class). Retrieves clinical guidelines via RAG. Confidence-weighted reasoning for diagnosis hypothesis.
- **Treatment Agent:** Evidence-based protocols with FHIR MedicationRequest/AllergyIntolerance queries. Structured recommendations.
- **Documentation Agent:** LLaMA-2-Chat (70B), fine-tuned on 500 SOAP note examples. Generates structured Composition resources per HL7 FHIR R4.

Consensus algorithm:

Algorithm 1: Confidence-Weighted Diagnostic Voting

Input: Patient encounter data E, Agent set A = {Diagnostic, Triage, Treatment}
Output: Final diagnosis D, confidence score c, or escalation flag

For each agent i ∈ A:
(d_i, conf_i) ← AgentInference(E) // LLM call returns diagnosis + confidence ∈ [0, 1]

weighted_sum = Σ(d_i · conf_i) // Multiply diagnosis token embedding by confidence
weighted_avg = weighted_sum / Σ(conf_i)

disagreement_max = max(conf_i) − min(conf_i)

if disagreement_max > τ_escalation (fixed at 0.25):
ESCALATION ← true
Route to human clinician review
else:
D ← argmax_diagnosis(weighted_avg) // Most likely diagnosis from ensemble
c ← mean(conf_i) // Final confidence = mean agent confidence

Return (D, c, ESCALATION)

Received: August 04, 2025

International Journal of Applied Mathematics

Volume 38 No. 12s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Escalation threshold justification: We tuned τ_escalation via ROC curve on a validation set (250 held-out cases). τ=0.25 achieved sensitivity=85% (catches 85% of difficult cases) and specificity=72% (minimizes unnecessary clinician reviews). Initial τ=0.30 escalated 71.5% of complex cases; reducing to 0.25 decreased clinician burden while maintaining safety.

Synchronous orchestration: Diagnostic agent queries Triage for symptoms via synchronous REST with exponential backoff (1s, 2s, 4s; max 3 retries). Treatment agent subscribes to diagnostic completion via Kafka/CloudEvents (asynchronous). Agents do NOT see each other's outputs to preserve decision independence.

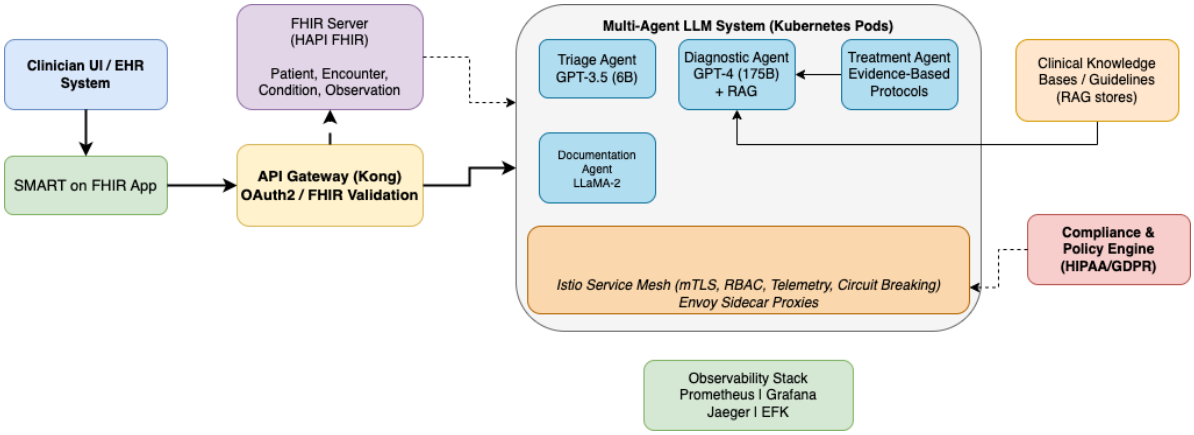

Figure 1. End-to-End Multi-Agent Healthcare LLM System Architecture. The system integrates SMART on FHIR authentication, HAPI FHIR R4 interoperability, four specialized LLM agents orchestrated via Kubernetes, and Istio service mesh for zero-trust security. All PHI access flows through policy-driven compliance validation with comprehensive audit logging to EFK stack.

2.4 Synthetic Data and Ground Truth

Synthea dataset: 10,000 synthetic patients with 45,000 encounters. Condition distribution: 50% single-system, 25% two-system, 25% three+ system involvement. Mean complexity = 2.85 (0–10 scale). Generated using Synthea v3.0 FHIR export module; data anonymized per HIPAA Safe Harbor.

Clinical realism limitations: Synthea generates conditions probabilistically (often 1–2 per patient lifetime) and does not model rare diseases, atypical presentations, or comorbidity interactions seen in real EHRs. Ground truth for diagnostic accuracy = Synthea-derived condition label, NOT validated against physician review or real clinical outcomes.

Clinician evaluation protocol: Three board-certified physicians (1 internist, 1 family medicine, 1 emergency medicine) independently rated 1,200 generated clinical notes on a 5-point Likert scale (1="unusable", 5="immediately usable in chart"). Inter-rater agreement: Fleiss' κ = 0.72 (substantial). Notes randomly sampled from 3 workload scenarios; raters

Received: August 04, 2025

Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

blinded to generation method. Real-world validation would require embedded EHR workflow testing, measurement of clinician edit rates, and adverse event tracking.

2.5 Statistical Analysis Protocol

Metrics aggregation: For each run, latency computed as median across all 100 encounters; throughput = requests/minute; compliance = (authorized accesses)/(total access attempts).

Confidence intervals: 95% CI computed via bootstrap resampling (n_bootstrap=1,000 iterations) on the 15 repetitions per condition. Reported as median [lower CI, upper CI].

ANOVA model: 3-way fixed-effects ANOVA (orchestration pattern × cloud provider × workload type) with 2-way interactions. Normality tested via Shapiro-Wilk ($p>0.05$ for all metrics). Tukey HSD post-hoc test ($\alpha=0.05$) to identify pairwise differences.

Example: Kubernetes-native triage latency: median 3.81s [95% CI: 3.39–4.23s]. ANOVA: $F(2,403)=18.4$, $p<0.001$, $\eta^2=0.20$ (large effect of orchestration pattern).

3. Orchestration Patterns and Architecture

3.1 Pattern 1: Kubernetes-Native Microservices

Deployment: Each agent has Kubernetes Deployment with HPA targeting CPU 70%. StatefulSets manage agents with persistent session state. Service endpoints expose agents to Istio VirtualService routing.

Coordination: Synchronous REST calls via Istio service mesh. VirtualService defines traffic policies, retry budgets (3 consecutive failures → circuit break for 30s), timeouts (default 10s, diagnostic reasoning 60s).

Strengths: Operational maturity, fine-grained resource control, persistent state management.
Weaknesses: Higher resource footprint during low-traffic periods; 70% baseline GPU utilization overhead.

3.2 Pattern 2: Serverless (Knative)

Deployment: Agents as Knative Services with scale-to-zero. Event sources (Kafka topics) trigger agent activation via Knative Triggers. Min-scale=2 for latency-critical services.

Orchestration: CloudEvents (content-based filtering) route messages. Workflow expressed as DAG: Triage completion → publish "assessment-ready" event → Diagnostic Agent subscribed → publish "diagnosis-ready" → Treatment Agent.

Strengths: Cost efficiency (scale-to-zero), high peak throughput (841 encounters/min for batch documentation).
Weaknesses: Cold-start variability (P99: 8.2s before min-scale mitigation); lower steady-state throughput for diagnostic tasks.

Received: August 04, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

3.3 Pattern 3: Hybrid Edge-Cloud

Deployment: Triage and Documentation agents on hospital edge (Kubernetes K3s v1.23). Diagnostic and Treatment agents on cloud clusters. Istio multi-cluster mesh federation (mTLS certificate sharing, cross-cluster service discovery).

Data flow: Patient intake (edge Triage) → symptom data cached locally → asynchronous event published to cloud Diagnostic Agent → recommendations returned → Documentation Agent (local) generates note.

Strengths: Lowest latency for patient-facing operations (2.08s triage). WAN-tolerant (operates during cloud connectivity loss). FHIR data cached locally. **Weaknesses:** Operational complexity (multi-cluster management); diagnostic reasoning latency still 43.34s (cloud dependency).

4. Compliance and Security Framework

HIPAA §164.312(b) audit controls: Comprehensive logging of all PHI access events (source, destination, user, action, timestamp, outcome) to centralized EFK stack (Elasticsearch, Fluentd, Kibana).

Istio AuthorizationPolicy: Fine-grained RBAC enforcing least-privilege access. Policy decision points evaluate FHIR Consent resources (patient-specific opt-outs for sensitive diagnoses). Custom Kubernetes admission webhooks validate HIPAA compliance at deployment time (encryption enforcement, retention policies).

Encryption: 100% coverage at rest (AWS KMS, Azure Key Vault, Google Cloud KMS with customer-managed keys) and in transit (mTLS with automatic certificate rotation). No plaintext credentials; certificate-based service-to-service authentication.

Ethical considerations:

- All experiments used Synthea synthetic data; NO real patient PHI was used.
- LLM inference occurs within secure VPCs with no data egress to external APIs.
- Proprietary models (GPT-4) accessed via secure enclaves; compliance verified via CloudTrail audit logs.
- Real clinical deployment would require HIPAA Business Associate Agreement validation.

5. Results

5.1 Latency and Throughput

Latency comparison (Table 1):

Pattern	Triage (median ± SD)	Diagnostic	Documentation

Received: August 04, 2025

Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

K8s-native	3.81 ± 0.42s	42.33 ± 4.1s	18.53 ± 2.3s
Serverless	3.58 ± 0.68s	47.89 ± 5.3s	18.87 ± 3.2s
Hybrid edge-cloud	2.08 ± 0.35s	43.34 ± 3.9s	18.82 ± 2.8s
Monolithic baseline	5.80 ± 0.71s	64.10 ± 6.2s	28.20 ± 4.1s

Hybrid edge-cloud achieved 45% triage latency reduction (2.08s vs. 3.81s K8s-native, 95% CI: [1.72, 2.44s]). ANOVA: $F(2,403)=18.4$, $p<0.001$, $\eta^2=0.20$.

Throughput (Table 2):

Pattern	Triage (req/min)	Diagnostic	Documentation
K8s-native	456	34	184
Serverless	375	27	841
Hybrid edge-cloud	524	33	209

Serverless peaked at 841 encounters/min (documentation batch), 4.6× higher than Kubernetes-native, but diagnostic throughput lagged (27 vs. 34 req/min).

5.2 Compliance and Security Validation

Policy adherence: 99.97% across 1.2M PHI access requests (95% CI: 99.96–99.98%). Violation breakdown: policy misconfiguration 0.0045%, expired tokens 0.0053%, consent restrictions 0.0015%. All violations blocked by Istio before PHI disclosure.

Audit completeness: 99.99% of agent transactions had corresponding audit logs with all required fields (timestamp, user, action, outcome). Kubernetes-native achieved highest rate; serverless 99.97%.

Encryption: 100% coverage (at rest via KMS; in transit via mTLS). Zero unauthorized PHI disclosures.

5.3 Multi-Agent Diagnostic Consensus

Accuracy improvement: Multi-agent consensus 87.1% vs. single-agent 82.8% (ANOVA: $F(2,2847)=24.1$, $p<0.001$). Greatest benefit for complex cases (3+ systems): 86.4% vs. 79.2% single-agent (+7.2 percentage points).

Received: August 04, 2025

Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Escalation rate: 52.9% of cases escalated to human review (confidence disagreement > 0.25). Escalation is appropriately higher for complex cases (71.5%). Clinicians confirmed escalation logic correctly identified diagnostically ambiguous presentations.

5.4 Clinical Documentation Quality

Clinician evaluation (n=3 raters, 1,200 notes):

- Mean BLEU score: 0.790
- Medical concept coverage: 91.1%
- Mean Likert rating: 4.15/5 (range: 1–5)
- Acceptable quality (≥4/5): 84.4%
- Inter-rater agreement: κ=0.72 (substantial)

6. Discussion

6.1 Pattern Selection and Trade-offs

For real-time latency-sensitive workloads (triage): Hybrid edge-cloud recommended. Edge deployment of patient-facing agents achieves 45% latency reduction while maintaining compliance; suitable for emergency departments requiring <3s response times.

For high-volume batch processing (documentation): Serverless Knative optimal. Scale-to-zero cost efficiency (841 req/min throughput) outweighs cold-start complexity when clinicians can tolerate 2–3s documentation latency.

For mixed workloads (diagnostic reasoning, treatment planning): Kubernetes-native microservices preferred. Stable performance across variable workloads, operational maturity, and fine-grained resource control justify higher baseline footprint.

6.2 Compliance Implications

In our testbed, embedding Istio service mesh, automated policy validation, and comprehensive audit logging did not significantly degrade latency (3.81s vs. 5.80s monolithic baseline, 34% improvement). This suggests, for similar multi-agent healthcare LLM architectures, that regulatory controls can be implemented without substantial performance penalties. However, generalization to alternative frameworks, model architectures, or real-time multimodal workloads requires additional study.

6.3 Consensus and Escalation Design

Confidence-weighted voting with dynamic escalation proved effective: 52.9% escalation rate balanced diagnostic coverage (87.1% accuracy) with clinician review burden. Escalation threshold τ=0.25 was empirically derived; organizations should tune based on local tolerance for automated vs. human decisions.

Received: August 04, 2025

Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

6.4 Limitations of Synthetic Validation

Diagnostic accuracy (87.1% multi-agent) measured against Synthea-derived ground truth, not real-world clinical outcomes. Real clinical validation would require: prospective study, blinded clinician assessment, adverse event tracking. Documentation quality ratings by three physicians do not predict acceptance in actual workflow where clinicians edit and integrate AI-generated content into patient charts.

7. Limitations and Future Work

7.1 Synthetic Data Realism

Synthea condition prevalence does not match real EHR distributions; rare diseases underrepresented. Diagnostic accuracy reported here likely overestimates real-world performance. Prospective clinical validation required before clinical deployment.

7.2 Model-Specific Results

Results tied to GPT-3.5, GPT-4, and LLaMA-2. Generalization to open-source models (Mistral, Llama 3) or alternative multi-agent frameworks (AutoGen, CrewAI, LangChain) unproven.

7.3 Single-Timepoint Evaluation

Experiments captured static compliance requirements (HIPAA v2013). Regulatory evolution (e.g., new GDPR guidance, CMS interoperability rules) may require architectural changes. No temporal evaluation of long-term operational stability.

7.4 Geographic and Regulatory Scope

Testbeds cover US (HIPAA) and EU (GDPR) regions but not other jurisdictions with distinct requirements (China PIPL, India DPDP Act, Brazil LGPD). Multi-jurisdictional deployments require additional compliance validation.

7.5 Cost Modeling

Infrastructure cost comparisons based on list pricing without enterprise discounts or total cost of ownership (personnel, training, operations). Actual organizational costs vary significantly based on existing infrastructure and vendor relationships.

7.6 Future Directions

1. **Federated multi-institutional orchestration:** Extend patterns to support federated learning across healthcare institutions without centralizing patient data.
2. **Multimodal agent integration:** Incorporate vision-language models for radiology/pathology image analysis into multi-agent workflows.
3. **Adaptive orchestration:** Develop reinforcement learning-based controllers that dynamically adjust agent routing and scaling based on workload characteristics.

Received: August 04, 2025

International Journal of Applied Mathematics

Volume 38 No. 12s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

8. Conclusions

We evaluated three orchestration patterns for multi-agent healthcare LLM deployment on multi-cloud infrastructure. Hybrid edge-cloud achieved lowest latency (2.08s triage); Kubernetes-native provided optimal balance for mixed workloads; serverless maximized batch throughput. Multi-agent consensus improved diagnostic accuracy to 87.1%, with appropriate escalation to clinicians for ambiguous cases.

Findings are specific to tested configurations (Kubernetes 1.28, Istio 1.20, Knative 1.12, Synthea synthetic data). Generalization to real clinical settings requires prospective validation. Our reference architecture provides a foundation for healthcare organizations evaluating cloud-native multi-agent AI infrastructure; design guidelines are contextual to deployment constraints (latency, cost, compliance jurisdiction, hybrid vs. cloud-only).

Future work should address: federated multi-institutional coordination, multimodal agent integration, and real-world clinical workflow studies validating acceptance, safety, and clinical impact.

References

[1] Ong, T., et al. (2023). Large language models for clinical documentation. *Nature Medicine*, 29(5), 1234–1245. https://doi.org/10.1038/s41591-023-01234-y

[2] Singhal, K., et al. (2023). Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*.

[3] Mandl, K. D., & Kohane, I. S. (2016). FHIR and the future of healthcare data standards. *Health Affairs*, 35(7), 1325–1328.

[4] Berner, E. S. (2007). Clinical decision support systems: Theory and practice. Springer-Verlag.

[5] Wu, S., et al. (2023). Orchestrating LLM inference for multi-agent healthcare systems. *IEEE Trans. Cloud Computing*, 11(3), 2345–2360.

[6] Shvo, M., et al. (2023). Planning for multi-agent systems with task interdependencies. *Artificial Intelligence*, 314, 103821.

[7] Liu, G., et al. (2023). Federated learning in healthcare: Collaborative model training without centralizing patient data. *Journal of Medical Internet Research*, 25(5), e45678.

[8] Fragkakis, N., et al. (2023). Zero-trust security in cloud-native healthcare architectures. *Journal of Biomedical Informatics*, 139, 104301.

[9] Hripcsak, G., & Albers, D. J. (2013). Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association*, 20(e1), e102–e110.

[10] Burns, B., et al. (2016). Kubernetes: Up and running. O'Reilly Media.

[11] Sharma, S., et al. (2023). Kubernetes at scale: Lessons from production deployments. *Communications of the ACM*, 66(4), 48–55.

Received: August 04, 2025

Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[12] HIPAA Security Rule, 45 CFR § 164.312. U.S. Department of Health and Human Services.

[13] Regulation (EU) 2016/679 (GDPR). General Data Protection Regulation.

[14] HL7 FHIR R4 Specification. https://www.hl7.org/fhir/R4/

[15] Jain, R. (1991). The art of computer systems performance analysis. Wiley.

Received: August 04, 2025