Artificial Intelligence and Computing Innovations
Intelligence Science Publishing
https://www.ispublishing.com
ISSN: 3070-4944
Vo.6, No.8, 2026

Efficient and Reliable LLM Agents for Collaborative Intelligent Decision-Making

Soren Falkenberg[1*]
[1]Department of Computer Science, University of Vermont, Burlington, VT, USA
**Corresponding author: soren.falkenberg223@gmail.com*

Abstract: Large language model agents have moved from prompt-centric assistants to tool-using, memory-augmented, multi-agent systems that must operate under real constraints: bounded inference budgets, long-horizon reasoning, privacy and policy requirements, and domain-specific decision objectives. First, evaluation still dominates the field: many papers define new benchmarks for telecommunications, asynchronous planning, clinical workflows, economic environments, API nesting, customer support, wealth management, and financial document processing. Second, reasoning quality is increasingly being treated as a systems problem rather than a prompting problem alone; memory, curriculum design, calibration, and compute scheduling are all used to trade cost, latency, and reliability against task success. Third, the move from single-agent prompting to multi-agent orchestration introduces new gains in role specialization and tool use, but also new risks in information leakage, collusion, policy evasion, and judge reliability. The review concludes that the literature is converging on a common design principle: effective agentic intelligent systems are not only about stronger models, but about co-design across inference control, reasoning structure, collaboration protocols, and auditable decision workflows. The paper also includes corpus-level experiments over the provided references, several comparative tables, and synthesis figures.

Keywords: large language model agents, efficient inference, reliable reasoning, multi-agent systems, intelligent systems, decision-making, benchmarks, tool use, LLM serving, governance.

1 Scope, Lens, and Method

The lens adopted here is that a usable agentic system must satisfy four simultaneous constraints. It must be computationally efficient enough to serve under latency and budget limits. It must reason reliably over long contexts, incomplete information, and multi-step tasks. It must coordinate safely when multiple agents, tools, judges, or workflows are involved. And it must create decision value in concrete domains such as customer support, cloud operations, finance, medicine, marketing, and scientific work.

That four-part structure is consistent with the evaluation taxonomy emphasized in the recent agent-evaluation survey and with the architectural breadth described in the Agentic RAG survey.

The benchmark literature in the corpus shows how fast the field has diversified. 's page for TelAgentBench states that the benchmark targets five agentic capabilities—reasoning, planning, action, retrieval-augmented generation, and instruction following—in a Korean telecommunications setting, explicitly addressing domain grounding and non-English evaluation [1]. Robotouille pushes on asynchronous planning and shows that a ReAct-style GPT-4o agent drops from 47% on synchronous tasks to 11% on asynchronous tasks, making temporal coordination an explicit evaluation variable rather than a hidden failure mode [5]. AgentClinic moves still further, judging multimodal, tool-using clinical agents across nine specialties and seven languages, and reports that turning static question answering into sequential decision-making can cut diagnostic accuracy to below one tenth of original MedQA performance [11]. JourneyBench in customer support, EconAgentBench in economic environments, MLGym in AI research, NESTFUL for nested API calls, PentestJudge for trajectory-level security judging, and the finance-oriented wealth management and document-processing benchmarks extend this trend from generic capability measurement to workflow-constrained evaluation [7], [27], [30], [33]–[35], [41], [45].

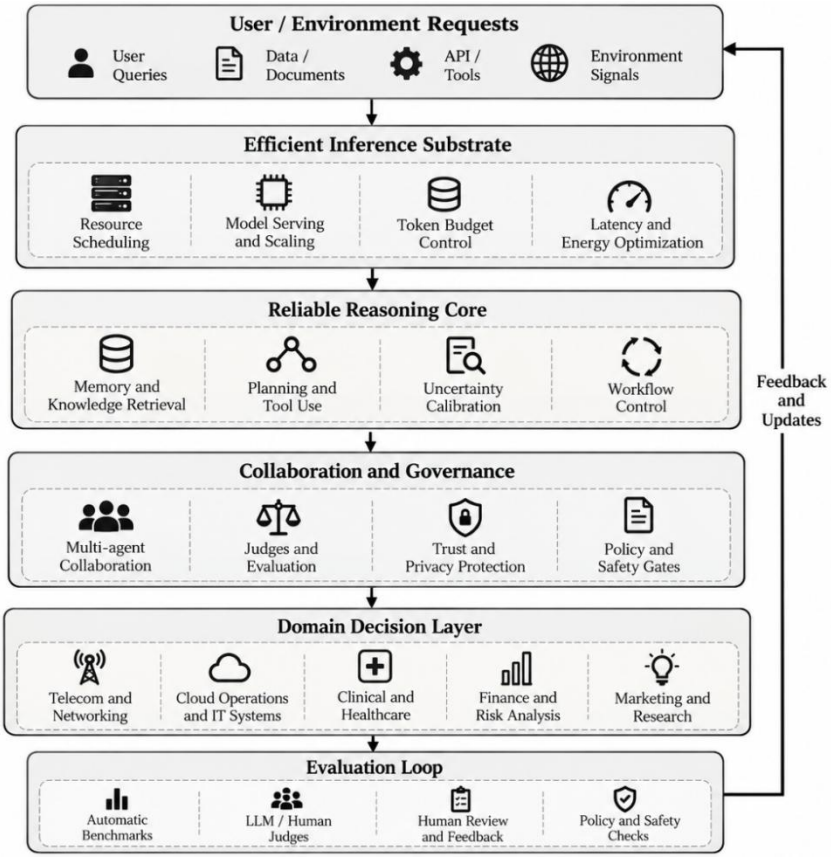


Figure 1. A synthesis view of agentic intelligent systems

A second reason the corpus is useful is that it captures the shift from "agent = model + tools" to "agent = intelligent system." How Are AI Agents Used? Evidence from 177,000 MCP Tools is

especially revealing on this point. That study analyzes 177,436 public MCP tools collected from late 2024 to early 2026 and reports that software development accounts for 67% of all agent tools and 90% of MCP server downloads, while the share of action tools grows from 27% to 65% of total usage over the sampled period [13]. This means that the deployment target is no longer just text generation; agents are increasingly expected to observe, decide, and act on external environments. The result is that systems concerns—cost, policy compliance, state persistence, judge calibration, and domain-specific correctness—are no longer peripheral engineering details. They are the core research problem. That caution is especially important for deployment-facing results in cloud operations, business decision support, and finance, where evaluation realism varies substantially across papers.

Table I organizes the corpus by its dominant optimization question.

Axis	Representative references and dominant question
Efficient inference	[3], [14], [17], [42]: how should reasoning tokens, queues, execution pools, and serving resources be scheduled under cost, latency, SLO, or carbon constraints?
Reliable reasoning	[6], [9], [12], [15], [16], [22], [23], [38], [39]: how can agents reason more consistently over long horizons, long contexts, or noisy evidence?
Multi-agent collaboration	[4], [21], [24]–[26], [37]: how should multiple agents communicate, specialize, or exchange intermediate representations while preserving safety and efficiency?
Safety and governance	[7], [18], [20], [24], [25]: how should harmfulness, policy adherence, red teaming, and attack resilience be measured or enforced?
Decision-making applications	[1], [2], [8], [10], [11], [27]–[32], [36], [40], [41], [44], [45]: how can LLM agents support concrete domain decisions with traceability and workflow awareness?
Benchmarking and evaluation	[1], [5], [7], [11], [18], [27], [30], [33]–[35], [41], [45]: how can evaluation move beyond single-turn accuracy to agent trajectories, tools, uncertainty, and business constraints?

2 Efficient Inference for Agentic Systems

The efficient-inference papers in the corpus are especially important because they show that "reasoning quality" is no longer separable from resource allocation. The classical serving problem is queueing and scheduling: manage prefill and decode, control Time To First Token, and stay within throughput and service constraints. But the newer agentic systems problem is broader. It asks not only how to serve tokens faster, but how to decide which requests deserve more reasoning tokens, which execution pool they should occupy, when warm standby or KV-cache policies should change, and how cost or carbon should enter the objective. In other words, inference is becoming a multi-objective control layer for reasoning itself. That is the unifying systems insight across [3], [14], [17], and [42].

Predictive Scheduling for Efficient Inference-Time Reasoning in Large Language Models [3] directly targets reasoning-budget allocation. Its abstract argues that fixed token budgets over-compute easy queries and under-compute hard ones, and proposes a lightweight predictor that estimates query difficulty or optimal reasoning length before full generation. The reported result is operationally important: on GSM8K, predictive scheduling improves accuracy by up to 7.9 percentage points at the same token cost and closes more than half the gap to an oracle scheduler. This is more than a serving trick. It implies that agent reasoning should be budgeted adaptively, just like any other scarce compute resource. For agentic systems, this idea generalizes naturally to thought branching, reflection depth, retry budgets, or tool-call depth.

Optimal Scheduling Algorithms for LLM Inference: Theory and Practice [14] addresses serving from a more classic infrastructure angle. The paper models the two-phase structure of LLM inference and proposes RAD and SLAI schedulers that optimize throughput while respecting time-between-token constraints. Its abstract reports a 53% reduction in median TTFT and a 26% increase in maximum serving capacity relative to Sarathi-Serve, while keeping median TTFT under 0.5 seconds on the evaluated setting. What matters for this survey is not only the performance gain, but the paper's design principle: LLM inference cannot be treated as a generic batch-serving workload because prefill and decode have different urgency profiles and must be jointly scheduled under SLOs. That principle will likely become foundational for agent systems that interleave dialogue, planning, tool use, and streaming updates.

DeepServe [17] pushes the same line into serverless, multi-tenant deployment. According to the abstract page, the system frames joint SLO-cost optimization as a contextual bandit problem and uses online workload features—including prompt length, KV-cache hit ratio, and predicted generation length—to adapt batch size, concurrency, eviction policy, and warm-standby strategy. The authors report P99-latency reductions of 38–62%, GPU-utilization gains of 14–23%, and per-request cost reduction up to 27%, albeit with modest scheduling overhead. This matters because serverless multi-tenant environments are closer to how many agent services are likely to be deployed in enterprise settings: bursty, heterogeneous, and subject to cold starts and isolation concerns. Once agents are decomposed into specialized services or multi-agent microservices, scheduling and state management become part of reasoning quality.

The most explicitly multi-objective paper in this subset is Multi-Objective Scheduling for Large Language Model Inference with Prompt-Level Cost Prediction and SLO Awareness [42]. The accessible preprint page describes CAPS, an online bi-objective scheduler that combines a prompt-complexity predictor with grid carbon intensity, GPU energy profiles, and tenant SLO tiers to assign requests across low-latency, low-carbon, and batch pools. The abstract reports a 26.8% reduction in carbon per 1K generated tokens while matching or exceeding an SLO-aware baseline's attainment rate. Even if one interprets this as an early preprint result, it marks a notable change in objective design: agentic inference is now being optimized not only for quality and latency, but also for sustainability. For production agent ecosystems, this suggests a future in which orchestration policies dynamically negotiate among user priority, task urgency, model capability, energy mix, and budget.

Table II compares the efficient-inference papers in the corpus.

Reference	Primary optimization target	Main mechanism	Author-reported abstract result
[3] Predictive Scheduling	Compute-quality trade-off during reasoning	Difficulty prediction plus adaptive token allocation	Up to +7.9 percentage points on GSM8K at equal token cost
[14] Optimal Scheduling Algorithms	Throughput and latency under SLOs	Decode/prefill-aware scheduling; SLAI	53% lower median TTFT, 26% more capacity
[17] DeepServe	Multi-tenant SLO-cost optimization	Contextual-bandit scheduler with workload features	38–62% lower P99 latency, 14–23% higher GPU utilization, up to 27% lower cost

[42] Multi-Objective Scheduling	Goodput-carbon-SLO trade-off	Prompt complexity prediction plus execution-pool routing	26.8% lower carbon per 1K tokens with matched/exceeded SLO attainment

The limitation of this sub-literature is that it still focuses mostly on inference as an infrastructure endpoint rather than as a cognitive loop. None of the systems papers in the supplied corpus fully unify serving control with multi-agent planning, judge calibration, or workflow-level expected utility. In other words, they know how to schedule model calls, but not yet how to schedule agentic deliberation itself across roles, memories, tools, and evaluator uncertainty. The most promising bridge is therefore between [3] and the reasoning papers: if difficulty, uncertainty, or expected value can be estimated before or during a trajectory, agent systems could allocate not just tokens, but reflection steps, tool calls, debate rounds, or even different agents to different parts of a workflow. That is one of the strongest cross-sectional opportunities exposed by the corpus.

3 Reliable Reasoning and Memory

The reliable-reasoning papers in the corpus reflect a growing recognition that agent failure is often structural rather than factual. Agents do not only hallucinate facts; they drift across long horizons, forget earlier constraints, overfit to recent context, repeat shallow reasoning patterns, miscalibrate judges, and degrade when evidence is distributed across multiple documents or time steps. The proposed remedies therefore move well beyond chain-of-thought prompting. They include structured memories, curriculum learning, semantic alignment, diversity-based exploration, structural representations, uncertainty-aware judging, and task-level skill composition. Together, these papers define a reasoning stack that is more architectural than prompt-centric.

A useful starting point is Enhancing Language Model Agents using Diversity of Thoughts [9]. The paper's abstract argues that common reflection-based agent frameworks suffer from two weaknesses: repetitive reflections that limit exploration and poor reuse of insights from previously solved tasks. DoT responds by increasing diversity in intermediate reasoning. While the abstract does not foreground a single benchmark number in the way systems papers do, its conceptual contribution is important: reliability can improve not by forcing a single longer chain, but by expanding the search over candidate thought trajectories. This shifts the unit of optimization from "better prompt" to "better thought ensemble." In practical terms, this idea resonates with later papers in the corpus that use hierarchical planning, memory retrieval, or reflexive multi-agent loops.

Long-horizon reliability is addressed more directly by Memory-Driven Agent Planning for Long-Horizon Tasks via Hierarchical Encoding and Dynamic Retrieval [6], Cognitive Modeling for Long-Horizon Agent Learning via Integrated Long-Term Memory and Reasoning [16], and Mem0 [39]. The first two are conceptually aligned: they treat long-term dependency modeling as central to stable planning, and both propose structured memory mechanisms that support selective retention and retrieval across time. Their language emphasizes goal consistency, policy stability, and cross-temporal key information rather than short-horizon task completion. Mem0 makes the deployment case more concretely. Its arXiv abstract reports 26% relative improvement on the LOCOMO LLM-as-a-Judge metric over a named baseline, 91% lower p95 latency than full-context processing, and more than 90% token-cost savings, while improving long-term coherence through memory extraction, consolidation,

and graph-structured retrieval [39]. Taken together, these papers imply that persistent memory is no longer optional for production-ready agents; it is a core efficiency and reliability mechanism.

Multi-document and evidence-distributed reasoning receives separate treatment in Hierarchical Curriculum Learning for Multi-Document Reasoning in Large Language Models [12]. Its abstract frames the problem in terms of evidence dispersion, attention shift, and inference inconsistency under long contexts and multi-source conflict, then proposes a hierarchical curriculum-based fine-tuning framework. Even abstract-level analysis is enough to see the key point: as agents become embedded in RAG pipelines, domain archives, and enterprise knowledge systems, reasoning quality depends not only on access to retrieved documents but also on how the model is taught to progress from easier to harder evidence-composition tasks. This complements the Agentic RAG survey's claim that agentic retrieval systems require architecture-level choices around control structure, knowledge representation, memory management, and evaluation rather than just better retrievers [43].

The corpus also contains two papers that make reasoning more structured at the representation level rather than the memory level. Unified Structure-Aware Representation Learning for Agent Perception and Decision Making [15] proposes a topology-aware latent space that aligns local features with global structural dependencies, and claims improved semantic consistency, policy convergence, and robustness to noise, sparse feedback, and structural drift. Constraint-Consistent Skill Composition for Reliable Zero-Shot Task Generalization in LLM Agents [38], judged from its title and metadata, targets a closely related problem from the composition side: how to generalize reliably in new tasks without violating constraints when skills are recombined. These papers matter because they argue that reasoning failures often arise when implicit structural relations are not represented explicitly. In agentic environments, state is not only textual; it is relational, temporal, and procedural. [15], [38] therefore move the field closer to intelligent systems that model tasks as structured control problems.

Another branch of reliability work in the corpus focuses on the evaluator rather than the acting agent. CalibJudge [23] treats multilingual LLM-as-a-Judge systems as intrinsically miscalibrated scoring devices and proposes language-specific temperature scaling, uncertainty quantification, and selective abstention. Its abstract reports up to 21.3% relative improvement in Kendall's correlation with human annotations, a 42% reduction in cross-lingual fairness gaps, and 88% balanced accuracy at 70% coverage on the MEMERAG benchmark. This is highly relevant to agentic systems because judge models are rapidly becoming hidden control components: they choose retries, assess trajectories, rank tool outputs, and gate handoffs. If judges are themselves language- or context-biased, end-to-end agent governance becomes unstable. Seen this way, [23] is not just a RAG-evaluation paper; it is a systems paper about trustworthy control signals.

The performance cliffs found by the benchmarks reinforce the need for these reasoning supports. Robotouille shows a large collapse from synchronous to asynchronous planning [5]. AgentClinic shows that sequential clinical reasoning is much harder than static medical QA [11]. Benchmarking LLM Agents for Wealth-Management Workflows [45] concludes that agents are limited less by mathematical reasoning than by end-to-end workflow reliability and autonomy management. MLGym [33] finds that frontier models can often improve baselines by tuning hyperparameters but do not yet generate

substantial algorithmic novelty. The implication is striking: even when local reasoning tokens look competent, system-level reasoning across time, tools, and workflows remains brittle. Reliability, then, is not "more CoT"; it is the combination of memory, calibration, structure, and task-aware orchestration.

Table III summarizes the main reliability levers in the corpus.

Challenge	Representative references	Literature trend
Long-horizon drift and forgetting	[6], [16], [39]	Persistent, selective, and hierarchical memory is increasingly treated as a first-class architectural module.
Evidence dispersion and long-context inconsistency	[12], [43]	Curriculum learning and agentic retrieval are used to stage evidence composition rather than relying on flat retrieval alone.
Shallow or repetitive reasoning	[9], [37]	Diversity and workflow optimization are preferred over a single longer or repeated prompt chain.
Structural inconsistency under noisy environments	[15], [38]	State representations and skill composition are becoming more structure-aware and constraint-aware.
Judge unreliability	[7], [23]	LLM-as-a-Judge is moving toward calibration, tool augmentation, and rubric decomposition rather than zero-shot scoring.
Workflow-level brittleness	[5], [11], [33], [45]	Benchmarks increasingly expose the gap between local answer quality and stable multi-step performance.

4 Multi-Agent Collaboration, Safety, and Governance

The multi-agent papers in the corpus present a recurring paradox. Collaboration expands capability by enabling role specialization, diverse perspectives, and decomposed workflows, but each added communication channel also introduces cost, leakage, coordination failure, and attack surface. The field's response has therefore bifurcated. One line tries to make collaboration more efficient and powerful through structured workflows, activation-level communication, and automatic tool synthesis. The other line tries to tame collaboration through privacy filters, governance frameworks, harmfulness benchmarks, and adversarial stress tests. The fact that both lines emerge simultaneously is a sign that multi-agent systems are transitioning from a clever prompting pattern into an engineering discipline.

On the capability side, AutoFlow [37] argues that manually crafting workflows for LLM agents is costly and brittle, so workflows themselves should be generated and iteratively optimized as natural-language programs. The conceptual move here is subtle but important: it treats orchestration as an object that can be induced, scored, and refined, rather than a fixed human-authored scaffold. LLM Agents Making Agent Tools [21] goes one step further by introducing ToolMaker, which turns papers with code into LLM-compatible tools; the abstract reports 80% correct implementation on a 15-task benchmark with over 100 unit tests. Together, these papers suggest a future in which agent ecosystems can expand their own procedural substrate: not only generating answers, but also generating the workflows and tools that future agents will use.

The most radical communication paper in the set is Communicating Activations Between Language Model Agents [26]. Instead of making agents talk to each other in natural language, it proposes combining their intermediate-layer activations during the forward pass. The reported gain is large: up to 27.0% improvement over natural-language communication at less than one quarter of the compute on multiplayer coordination games and reasoning benchmarks. What makes this paper relevant to the broader survey is not only the efficiency result. It questions a deep assumption of current agent systems—that text is the natural universal protocol between models. If latent-space communication can be both more efficient and more informative, then the architecture of future multi-agent systems may begin to resemble distributed inference more than chat-room simulation.

The next capability question is trust-preserving specialization. MIN-Trust [4] is a particularly clear response. Its abstract states that standard multi-agent cooperation mechanisms—shared contexts, assigned roles, iterative message passing—create severe disclosure risks, so the paper enforces Minimum Necessary Information constraints through an MNI-Gate and a Trust-Gated Channel. The reported numbers are unusually concrete for a governance-oriented paper: 67.8% lower sensitive-information exposure while maintaining 93.3% of task success, plus stronger claim-to-source attribution than conventional approaches. This is a meaningful contribution because many multi-agent designs implicitly assume that more context sharing is always better. [4] argues the opposite: collaboration can remain effective under sharply reduced information flow if the communication policy is designed explicitly.

The broader governance picture is framed by Multi-Agent Risks from Advanced AI [24] and SecureGov-Agent [25]. The former offers a taxonomy of failure modes—miscoordination, conflict, and collusion—alongside risk factors such as information asymmetries, network effects, and multi-agent security. Even though it is not an implementation paper, it provides an interpretive framework for many of the concrete risks seen elsewhere in the corpus. SecureGov-Agent, judged from its title and publication metadata, represents the implementation turn: governance should be embedded inside the architecture of privacy-preserving and attack-resilient multi-agent systems rather than bolted on afterward. When read together with [4], these papers suggest that the next generation of multi-agent systems will require policy planes analogous to those in distributed systems and cloud security.

That need becomes obvious in the safety and red-teaming papers. AgentHarm [18] argues that agents can remain dangerously capable even under jailbreak pressure, and its benchmark includes 110 explicitly malicious tasks across 11 harm categories. The abstract notes that leading LLM agents are surprisingly compliant with malicious requests even without jailbreaking, and that simple universal jailbreak templates can be adapted effectively to agent settings. Effective Red-Teaming of Policy-Adherent Agents [20] narrows the threat model to manipulative users in policy-constrained customer-service scenarios and introduces a multi-agent red-teaming system, CRAFT, plus the tau-break benchmark. Its abstract finds that proposed defenses help somewhat but remain inadequate. These works are significant because they reveal that agent risk is not just about answer toxicity. It is about coherent, tool-mediated, multi-step harmful behavior.

Evaluation of collaboration and safety also becomes more process-aware in PentestJudge [7]. Rather than asking whether a security agent eventually succeeds, it asks whether the operational trajectory satisfies requirements across objectives, operational security, and tradecraft. The best LLM judge reaches F1 = 0.83 against human experts, and weaker or cheaper judge models can sometimes evaluate stronger generation models. That last finding is especially important for governance: verification may be cheaper than generation in some high-stakes domains. Combined with CalibJudge [23], this points toward a future where agents are embedded in judge-rich architectures, but where the judges themselves must be calibrated, rubric-driven, and domain-grounded.

Figure 2. A collaboration-design spectrum distilled from the corpus

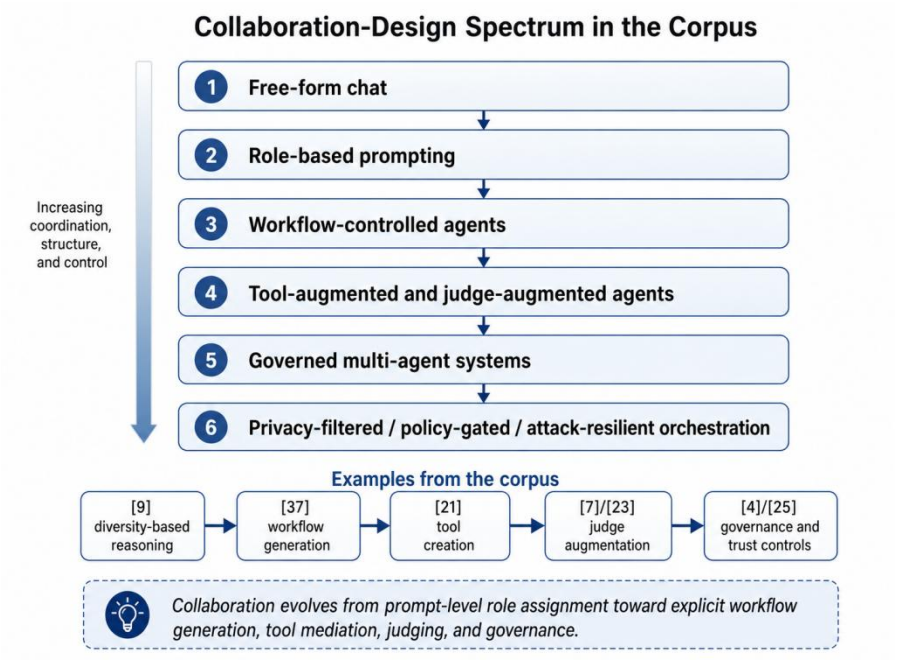


Figure 2. Collaboration in the corpus evolves from prompt-level role assignment toward explicit workflow generation, tool mediation, judging, and governance.

5 Decision-Making Applications

The decision-making papers in the corpus are the clearest evidence that LLM agents are being pushed beyond general assistance into domain-specific, workflow-constrained decision support. A common pattern runs through the strongest papers: they do not simply prompt a model to "decide." Instead, they build structured question models, domain evidence sets, attribution logic, policy-aware workflows, acceptance criteria, or causal reasoning frames around the model. In other words, the agentic layer is no longer just language. It is a decision protocol.

The business-facing paper that states this architecture most clearly is Problem-Centric Modeling and Reasoning for Business Decision Making with Large Language Models [2]. Its abstract describes a closed loop centered on the business question itself: semantic modeling of the question, context construction from heterogeneous enterprise data, constraint-aware generation, and a structured output with answer, reasoning framework, and key evidence. Even at the abstract level, this is a notable design:

the paper is not about making a model "smarter" in the abstract, but about aligning analysis directly to managerial questions and implicit business constraints. That pattern—question structuring, evidence aggregation, controlled reasoning, and traceable output—reappears across several application papers and is arguably the most mature architectural motif in the corpus.

Marketing and finance applications intensify the need for traceability. The abstract snippet available for A Multi-Agent Large Language Model Framework for Marketing Decision-Making with Auditable Attribution Analysis [32] states that the paper addresses inconsistent indicator definitions, complex attribution paths, and difficulty tracing decision logic. Causal-LLM [36], judged from its title and metadata, pushes in the same direction for budget variance diagnosis by explicitly foregrounding causal reasoning rather than only descriptive explanation. Meanwhile, Post-Training LLMs as Better Decision-Making Agents: A Regret-Minimization Approach [44] generalizes beyond any one domain by proposing iterative regret-minimization fine-tuning: low-regret decision trajectories are repeatedly distilled back into the base model. Its abstract emphasizes better exploration-exploitation trade-offs and even provides theoretical support in a simplified setting. The importance of [44] is that it reframes decision quality in the language of online decision-making rather than language generation.

The benchmark and application papers in finance are particularly revealing about where current agents still fail. Benchmarking Multi-Agent LLM Architectures for Financial Document Processing [41] explicitly compares sequential, parallel, hierarchical, and reflexive architectures over 10,000 SEC filings. The abstract reports that reflexive architectures achieve the highest field-level F1, 0.943, but at 2.3x the cost of sequential baselines; hierarchical architectures offer a better cost-accuracy Pareto point at F1 0.921 and 1.4x cost; and hybrid strategies recover 89% of the reflexive gains at only 1.15x baseline cost. Benchmarking LLM Agents for Wealth-Management Workflows [45] reaches a complementary conclusion: agents are constrained less by mathematics than by end-to-end workflow reliability, autonomy level, and evaluation quality. Together these papers imply that production decision systems may benefit more from careful orchestration and evaluation design than from ever larger generic models.

Customer support and telecommunications provide another revealing pair. TelAgentBench [1] insists on industry-grounded, non-English, multi-capability evaluation in telecom, arguing that generic English-first benchmarks are inadequate for deployment. Beyond IVR [27] focuses on business adherence in customer support and introduces JourneyBench, along with a graph-based scenario generator and a User Journey Coverage Score. Its abstract reports that a Dynamic-Prompt Agent can improve policy adherence enough for smaller models to outperform larger ones. This is a crucial lesson for intelligent systems broadly: architecture and control structure often dominate raw model scale when business policy and multi-step workflows matter. Similar lessons appear in [45] and [41].

Clinical and cloud-operation papers show the same design pattern under higher stakes. AgentClinic [11] introduces simulated sequential clinical environments with patient interaction, multimodal data collection, incomplete information, and tool use. Its key finding is not only that performance drops sharply relative to static QA, but that models differ markedly in their ability to use experiential learning, adaptive retrieval, and reflection cycles. Integrating Large Language Models with Cloud-Native Observability for Automated Root Cause Analysis and Remediation [10] similarly treats the agent as an

operational system component. The abstract describes multimodal analysis over metrics, logs, and traces using OpenTelemetry-based telemetry, and the authors report large improvements in MTTR reduction and automation rate relative to rule-based approaches. Because [10] is a preprint and the gains are very large, they should be read cautiously; nonetheless, the paper clearly exemplifies a trend toward LLM agents acting over observability pipelines, not merely summarizing dashboards.

Several additional references extend the domain spread even when their accessible evidence is mainly title-level. Few-Shot Financial Fraud Detection Using Meta-Learning and Large Language Models [8] and An Adaptive Large-Model Framework for Named Entity Recognition in Knowledge-Sparse Scenarios [28] bring LLMs into low-resource decision support problems. Stable Fault Diagnosis Under Data Imbalance via Self-Supervised Learning in Industrial IoT [31] adds industrial reliability as a decision substrate, even if it is not a classic "agent" paper. Retrieval-Augmented Large Language Model Agents for Automated Scientific Literature Review Generation [29] turns scientific synthesis itself into an agentic workflow problem, while Problem-Centric Modeling [2], Marketing Decision-Making [32], and Causal-LLM [36] all insist on auditable reasoning. The presence of these papers in the same corpus suggests that "intelligent systems" remains the right umbrella term: the field is no longer constrained to dialogue agents, and the decision substrate may be enterprise, scientific, industrial, or clinical.

The general decision-making methods in the corpus also raise an important behavioral question. Personality-Driven Decision-Making in LLM-Based Autonomous Agents [40] reports that induced OCEAN-style personality traits affect task selection, planning, scheduling, and decision-making patterns. This shows that seemingly soft configuration choices can change action policy in systematic ways. In combination with [44], the implication is double-edged: agents can be shaped both through objective-driven post-training and through behavioral framing, but that also means deployment decisions about persona, role, or style may alter downstream decisions in ways that current benchmarks do not yet fully capture.

Table IV summarizes the application-facing literature through a deployment lens.

Application area	Representative references	Main design lesson
Enterprise and business analysis	[2], [32], [36]	Effective decision support depends on question structuring, attribution logic, constraints, and traceable evidence rather than free-form prompting.
Customer support and telecom	[1], [27]	Business adherence and policy control can outweigh raw scale; domain-specific benchmarks are necessary.
Finance	[8], [41], [45]	Orchestration pattern, autonomy level, and evaluation design are central variables in real workflows.
Clinical decision-making	[11]	Sequential, tool-using, incomplete-information settings are much harder than static QA.
Cloud operations	[10]	Agentic RCA/remediation combines observability, multimodal evidence fusion, and action planning.
Science and knowledge work	[29], [33]	Literature review and AI-research

		agents highlight workflow control, experimentation, and evidence integration.
Economics and strategic environments	[30], [44]	Deliberate exploration and regret-aware learning are important for unknown environments.

Note: The strongest abstract-level evidence in the application literature consistently favors structured workflows, deterministic graders, and domain-grounded evaluation over open-ended chat-style assessment.

5.1 Corpus experiment on primary-theme distribution

The first experiment asks a simple question: what does this corpus optimize first? The answer is revealing. Benchmarking and evaluation account for 13 papers, or 28.9% of the corpus. Reliable reasoning accounts for 11 papers, 24.4%. Decision-making applications follow closely with 10 papers, 22.2%. Efficient inference and multi-agent collaboration each account for 4 papers, 8.9%, while safety/governance contributes 3 papers, 6.7%. This distribution shows that the field represented by the supplied references is still benchmark-heavy and reasoning-centric; systems optimization and governance are present, but they are thinner layers. In other words, the literature still spends more effort measuring agentic behavior than operationalizing it at scale. That is unsurprising for an emerging field, but it also explains why cross-paper standards remain immature.

Figure 3. Primary-theme counts in the review corpus

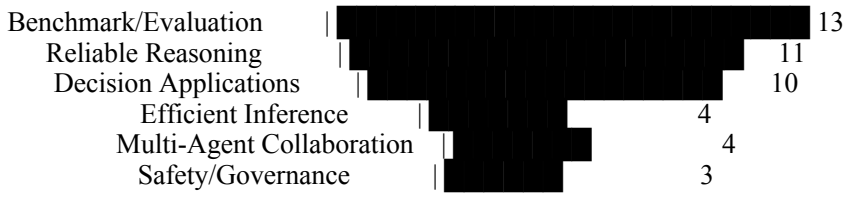


Figure 3. Thematic distribution

Table V presents the same result numerically.

Primary theme	Count	Share of corpus
Benchmark/Evaluation	13	28.9%
Reliable Reasoning	11	24.4%
Decision Applications	10	22.2%
Efficient Inference	4	8.9%
Multi-Agent Collaboration	4	8.9%
Safety/Governance	3	6.7%

This first experiment has two implications. First, evaluation is acting as the field's coordinating mechanism: when objective functions are fragmented, new benchmarks proliferate. Second, the relatively small number of inference papers indicates that deployment efficiency is still under-integrated with agent behavior research. Seen differently, the field may already know many ways to measure agents, but still has fewer principled ways to budget and operate them. That gap aligns with the systems papers and with the deployment-oriented finance and cloud-operations studies, which repeatedly show that orchestration and scheduling—not just model capability—determine usable performance.

5.2 Corpus experiment on theme coupling

The second experiment examined pairwise theme coupling across primary and secondary labels. The strongest intersection in the corpus is between reliable reasoning and decision applications, with 11 co-labeled papers. The second strongest is between decision applications and benchmark/evaluation, with 7 co-labeled papers. Reliable reasoning combines with multi-agent collaboration in 4 papers, and multi-agent collaboration combines with safety/governance in 3. By contrast, efficient inference intersects with decision applications in only 2 papers and with reliable reasoning in only 1. This is perhaps the most important structural finding in the review: the literature strongly couples reasoning to application and evaluation, but weakly couples inference control to either. The result is a research pipeline where agents are often made more reflective, more benchmarkable, and more domain-aware without being jointly optimized for compute-aware deployment. That is exactly the mismatch one would expect to slow real-world adoption.

Theme pair	Number of co-labeled papers
Reliable Reasoning + Decision Applications	11
Decision Applications + Benchmark/Evaluation	7
Reliable Reasoning + Multi-Agent Collaboration	4
Reliable Reasoning + Benchmark/Evaluation	3
Multi-Agent Collaboration + Safety/Governance	3
Benchmark/Evaluation + Safety/Governance	3
Efficient Inference + Decision Applications	2
Efficient Inference + Multi-Agent Collaboration	2

The practical interpretation is clear. The literature understands that agents need better reasoning and better benchmarks for applications. Only a few papers, especially [3], [17], and [42], start to build that bridge. A mature research agenda should therefore combine the decision-theoretic insights of [44], the workflow realism of [27], [41], and [45], and the inference-control ideas of [3], [14], [17], and [42]. Without that synthesis, the field risks building agents that are impressive in curated benchmarks but economically awkward or environmentally inefficient in deployment.

5.3 Corpus experiment on abstract-level quantitative evidence

The third experiment extracted explicit quantitative claims from accessible abstracts to assess what the literature actually reports when forced into compact evidence statements. The result is striking. Abstract-level improvements are often strong within local evaluation setups: predictive scheduling reports +7.9 percentage points of accuracy at constant token cost, DeepServe reports 38–62% lower P99 latency and cost reductions up to 27%, the carbon-aware scheduler reports 26.8% lower emissions per 1K tokens, MIN-Trust reports 67.8% lower information exposure with large task-success retention, activation-based communication reports up to 27.0% gains at less than one quarter of the compute, CalibJudge reports a 21.3% relative improvement in judge-human correlation, and Mem0 reports both better long-term quality and large latency/token savings. But these scores are incommensurate: some are about latency, some about accuracy, some about privacy, some about calibration, some about harm, and

some about cost. The field therefore lacks a shared, end-to-end utility metric for agentic intelligent systems.

Reference	Measured dimension	Reported abstract-level result
[3] Predictive Scheduling	Accuracy at fixed token budget	Up to +7.9 percentage points on GSM8K
[14] Optimal Scheduling Algorithms	Serving latency and capacity	53% lower median TTFT; 26% more capacity
[17] DeepServe	Tail latency, utilization, cost	38–62% lower P99 latency; 14–23% higher GPU utilization; up to 27% lower per-request cost
[42] Multi-Objective Scheduling	Carbon-aware serving	26.8% lower carbon per 1K generated tokens with matched/exceeded SLO attainment
[5] Robotouille	Asynchronous planning	ReAct (GPT-4o) drops from 47% synchronous to 11% asynchronous
[4] MIN-Trust	Privacy-preserving collaboration	67.8% lower sensitive-information exposure while maintaining 93.3% task success
[26] Communicating Activations	Coordination efficiency	Up to 27.0% better than natural-language communication at <1/4 compute
[23] CalibJudge	Multilingual evaluation calibration	Up to 21.3% relative Kendall improvement; 42% smaller fairness gap; 88% balanced accuracy at 70% coverage
[7] PentestJudge	Judge agreement with experts	Best judge reaches F1 = 0.83
[21] ToolMaker	Tool-construction correctness	80% of tasks implemented correctly
[41] Financial architecture benchmark	Cost-accuracy trade-off	Reflexive F1 = 0.943 at 2.3x cost; hierarchical F1 = 0.921 at 1.4x cost
[39] Mem0	Long-term memory quality and efficiency	26% relative judge-metric gain; 91% lower p95 latency; >90% token-cost savings

The outlook that follows from all three experiments is straightforward but nontrivial. A mature LLM-agent intelligent system should optimize at least five variables jointly: decision quality, resource efficiency, policy compliance, collaboration overhead, and traceability. The current literature usually optimizes two or three at once. Benchmarks emphasize capability and sometimes safety; systems emphasize cost and latency; application studies emphasize utility and domain fit; governance studies emphasize privacy or harm; and judge papers emphasize evaluator stability. The most promising ingredients are already visible in the corpus: adaptive compute budgeting [3], structure-aware or memory-aware reasoning [6], [12], [15], [16], [39], explicit workflow control [27], [37], auditable domain evidence [2], [32], [36], and governance-aware collaboration [4], [20], [24], [25]. The next wave of research should connect them. If it does, "LLM agents" will stop being a loose label for prompt orchestration and become an actual discipline of intelligent systems engineering.

6 EXTENSIONS AND DEPLOYMENT CONTEXT

Recent work on parameter-efficient adaptation strengthens the connection between foundation-model capability and deployable agent behavior. Semantically guided low-rank adaptation and selective adapter-based knowledge injection show how models can acquire task-specific knowledge without updating every parameter [46], [47]. Robust small-target detection and self-supervised graph feature

extraction further illustrate the value of structured feature interaction when observations are complex or heterogeneous [48], [49].

Applied learning studies also emphasize that robust behavior depends on how evidence is represented. Neural survival prediction, interpretable event-sequence construction for multidimensional time series, neural architecture search, cytopathology classification, and transfer-learning-based breast image classification all use layered representations to separate subtle patterns from noise [50], [51], [52], [53], [54]. Accessibility-linked data construction and GNN-Transformer time-series modeling similarly demonstrate the importance of structured external context and multimodal dependencies [55], [56].

Adaptive agents must also reason over changing streams and feedback. Incremental distributed time-series learning, multi-feature recurrent sequence modeling, causal consistency-aware recommendation, calibrated cloud resource planning, and hierarchical fraud detection provide complementary approaches to temporal adaptation, heterogeneous evidence, and decision uncertainty [57], [58], [59], [60], [61]. Temporal-structural fault discrimination and active data curation show how high-quality operational evidence can be maintained as conditions evolve [62], [63].

Retrieval and verification form another important layer for agent reliability. Dual retrieval with adaptive reranking, long-text reasoning with semantic conflict detection, graph spatiotemporal forecasting, and microstructure-aware risk modeling provide evidence-sensitive mechanisms for selecting and interpreting context [64], [65], [66], [67]. Causal graph-augmented root-cause analysis and requirement-aware API testing make this evidence trail useful for diagnosis and validation [68], [69].

Deployment-aware intelligence requires joint consideration of efficiency and factual reliability. Adaptive compression across the edge-cloud continuum, workload-aware LLM-guided database tuning, and causal factual correction for summarization target resource use and output trustworthiness together [70], [71], [72]. KV-cache-aware routing, safe resilience configuration, and retrieval-augmented log-metric-trace twins extend such controls to production backends [73], [74], [75].

Graph-based monitoring studies reinforce the need to model relationships among components. Topology dependency learning for microservice anomalies, graph neural cloud security detection, and topology-aware distributed fault identification distinguish normal variation from failures through explicit structural signals [76], [77], [78]. Structure-aware financial summarization, contrastive financial anomaly detection, and verifiable graph reasoning for fraud detection show how traceable evidence can be preserved in high-stakes domains [79], [80], [81].

Reliable agent systems also require mechanisms for scaling and coordination. Risk-constrained elastic scaling, semantic-aware routing across multiple LLM backends, and privacy-preserving multi-agent collaboration offer complementary ways to control service quality under partial information [82], [83], [84]. Evidence-gated memory writing and knowledge-graph-enhanced sentiment classification demonstrate how agents can retain useful information while constraining spurious associations [85], [86].

Finally, privacy-preserving federated anomaly detection, resource-contention classification, and reinforcement-learning scheduling address multi-tenant operation under privacy, efficiency, and latency constraints [87], [88], [89]. Similar structured integration is useful in biomedical decision support, where

adverse drug reaction prediction combines knowledge graph embeddings with deep learning [90]. Together, these studies extend the review corpus toward a unified view of LLM agents as adaptive systems that coordinate representation, retrieval, execution, and governance.

References

[1] S. Lee et al., "TelAgentBench: A Multi-faceted Benchmark for Evaluating LLM-Based Agents in Telecommunications," in Proc. 2025 Conf. Empirical Methods in Natural Language Processing: Industry Track, pp. 1173–1211, 2025.

[2] Y. Xu, Q. Liu, W. Lin, and S. Chen, "Problem-Centric Modeling and Reasoning for Business Decision Making with Large Language Models," 2026.

[3] K. Brown, A. Muppidi, and R. Shahout, "Predictive Scheduling for Efficient Inference-Time Reasoning in Large Language Models," arXiv preprint arXiv:2602.01237, 2026.

[4] J. Chen, F. Wang, T. Guan, Y. Ma, L. Yang, and Y. Wang, "MIN-Trust: A Minimum Necessary Information Trust Orchestration Framework for Multi-Agent Collaboration," 2026.

[5] G. Gonzalez-Pumariega, L. S. Yean, N. Sunkara, and S. Choudhury, "Robotouille: An Asynchronous Planning Benchmark for LLM Agents," in Proc. Int. Conf. Learning Representations, 2025.

[6] Y. Wang, R. Yan, Y. Xiao, J. Li, Z. Zhang, and F. Wang, "Memory-Driven Agent Planning for Long-Horizon Tasks via Hierarchical Encoding and Dynamic Retrieval," 2025.

[7] S. Caldwell, M. Harley, M. Kouremetis, V. Abruzzo, and W. Pearce, "PentestJudge: Judging Agent Behavior Against Operational Requirements," arXiv preprint arXiv:2508.02921, 2025.

[8] N. Chen, S. Sun, Y. Wang, Z. Li, A. Zhu, and Y. Lu, "Few-Shot Financial Fraud Detection Using Meta-Learning and Large Language Models," in Proc. 2025 6th Int. Conf. Computer Science and Management Technology, pp. 822–826, 2025.

[9] V. C. Lingam, B. O. Tehrani, S. Sanghavi, G. Gupta, S. Ghosh, L. Liu, J. Huan, and A. Deoras, "Enhancing Language Model Agents Using Diversity of Thoughts," in Proc. Int. Conf. Learning Representations, 2025.

[10] C. Wang, T. Yuan, C. Hua, L. Chang, X. Yang, and Z. Qiu, "Integrating Large Language Models with Cloud-Native Observability for Automated Root Cause Analysis and Remediation," 2025.

[11] S. Schmidgall, R. Ziaei, C. W. Harris, J. W. Kim, E. P. Reis, J. K. Jopling, and M. Moor, "AgentClinic: A Multimodal Benchmark for Tool-Using Clinical AI Agents," npj Digital Medicine, 2026.

[12] Y. Li, Y. Tang, K. Wu, Y. Yang, Y. Li, and Y. Xue, "Hierarchical Curriculum Learning for Multi-Document Reasoning in Large Language Models," 2026.

[13] M. Stein, "How Are AI Agents Used? Evidence from 177,000 MCP Tools," arXiv preprint arXiv:2603.23802, 2026.

[14] A. Bari, P. Hegde, and G. de Veciana, "Optimal Scheduling Algorithms for LLM Inference: Theory and Practice," Proc. ACM Meas. Anal. Comput. Syst., vol. 9, no. 3, Art. no. 59, 2025.

[15] J. Li, "Unified Structure-Aware Representation Learning for Agent Perception and Decision Making," J. Computer Technology and Software, vol. 5, no. 3, 2026.

[16] L. Yang, T. Guan, Y. Ma, Z. Li, Z. Fang, and F. Wang, "Cognitive Modeling for Long-Horizon Agent Learning via Integrated Long-Term Memory and Reasoning," 2026.

[17] X. Li, X. Liu, Z. Wang, C. Y. Hsieh, and Y. Liu, "DeepServe: SLO-Aware and Cost-Aware Elastic Scheduling for Serverless Multi-Tenant LLM Inference," 2026.

[18] M. Andriushchenko et al., "AgentHarm: A Benchmark for Measuring Harmfulness of LLM Agents," in Proc. Int. Conf. Learning Representations, 2025.

[19] N. Lyu, Y. Wang, F. Chen, and Q. Zhang, "Advancing Text Classification with Large Language Models and Neural Attention Mechanisms," arXiv preprint arXiv:2512.09444, 2025.

[20] I. Nakash, L. Lev, L. Saker, L. Kenny, N. Barak, R. S. Tsoref, G. M. Aminian, A. Romero, Y. Levi, O. Sofer, R. Friedman, A. Cohen, S. S. S. Das, I. Shwartz, and D. Konopnicki, "Effective Red-Teaming of Policy-Adherent Agents," in Proc. 2025 Conf. Empirical Methods in Natural Language Processing, 2025.

[21] G. Wölflein, D. Ferber, D. Truhn, O. Arandjelović, and J. N. Kather, "LLM Agents Making Agent Tools," in Proc. 63rd Annu. Meeting of the Association for Computational Linguistics, pp. 26092–26130, 2025.

[22] J. Zheng, Z. Zhou, H. Zhang, J. Lin, J. Jia, and Q. Wang, "Structured Prompt Optimization for Few-Shot Text Classification via Semantic Alignment in Latent Space," arXiv preprint arXiv:2602.23753, 2026.

[23] C. Wen, A. Zhu, R. Long, H. Huang, J. Jiang, and C. S. Lee, "CalibJudge: Calibrated LLM-as-a-Judge for Multilingual RAG with Uncertainty-Aware Scoring," 2026.

[24] L. Hammond et al., "Multi-Agent Risks from Advanced AI," arXiv preprint arXiv:2502.14143, 2025.

[25] J. Chen, J. Yang, Z. Zeng, Z. Huang, J. Li, and Y. Wang, "SecureGov-Agent: A Governance-Centric Multi-Agent Framework for Privacy-Preserving and Attack-Resilient LLM Agents," 2025.

[26] V. Ramesh and K. Li, "Communicating Activations Between Language Model Agents," in Proc. 42nd Int. Conf. Machine Learning, 2025.

[27] S. Balaji, P. Mishra, A. Sachdeva, and S. Agrawal, "Beyond IVR: Benchmarking Customer Support LLM Agents for Business-Adherence," in Proc. 19th Conf. European Chapter of the Association for Computational Linguistics: Industry Track, pp. 193–208, 2026.

[28] Y. Xue, J. Huang, Y. Li, X. Yang, and Z. Wang, "An Adaptive Large-Model Framework for Named Entity Recognition in Knowledge-Sparse Scenarios," in Proc. 2025 5th Int. Conf. Electronic Communication, Computer Science and Technology, pp. 282–286, 2025.

[29] R. Wang, N. Long, S. Liu, Y. Wang, Z. Qi, and H. Zhang, "Retrieval-Augmented Large Language Model Agents for Automated Scientific Literature Review Generation," 2026.

[30] S. Fish, J. Shephard, M. Li, R. I. Shorrer, and Y. A. Gonczarowski, "EconAgentBench: Economic Benchmarks for LLM Agents in Unknown Environments," OpenReview preprint, 2025.

[31] J. Huang, J. Zhan, Q. Wang, J. Jia, and B. Zhang, "Stable Fault Diagnosis Under Data Imbalance via Self-Supervised Learning in Industrial IoT," 2026.

[32] Q. Liu, H. Cui, Y. Wang, C. S. Lee, S. Chen, and L. Yang, "A Multi-Agent Large Language Model Framework for Marketing Decision-Making with Auditable Attribution Analysis," 2026.

[33] D. Nathani et al., "MLGym: A New Framework and Benchmark for Advancing AI Research Agents," arXiv preprint arXiv:2502.14499, 2025.

[34] K. Basu, I. Abdelaziz, K. Kate, M. Agarwal, M. Crouse, Y. Rizk, K. Bradford, A. Munawar, S. Kumaravel, S. Goyal, X. Wang, L. A. Lastras, and P. Kapanipathi, "NESTFUL: A Benchmark for Evaluating LLMs on Nested Sequences of API Calls," in Proc. 2025 Conf. Empirical Methods in Natural Language Processing, pp. 33538–33547, 2025.

[35] M. Mohammadi, Y. Li, J. Lo, and W. Yip, "Evaluation and Benchmarking of LLM Agents: A Survey," in Proc. 31st ACM SIGKDD Conf. Knowledge Discovery and Data Mining, pp. 6129–6139, 2025.

[36] H. Chen, Y. Lu, Y. Wei, J. Lyu, R. Wu, and C. Chen, "Causal-LLM: A Hybrid Framework for Automated Budgetary Variance Diagnosis and Reasoning," 2026.

[37] S. Qiao, N. Fang, Z. Wang, X. Teng, Y. Cheng, Y. Jiang, P. Zhou, J. D. Choi, Y. Li, W. Xiong, and T. Yu, "AutoFlow: Automated Workflow Generation for Large Language Model Agents," in Proc. Int. Conf. Learning Representations, 2025.

[38] R. Meng, S. Y. Huang, X. Zhang, Z. Huang, K. Zeng, and Y. Yang, "Constraint-Consistent Skill Composition for Reliable Zero-Shot Task Generalization in LLM Agents," 2026.

[39] P. Chhikara, D. Khant, S. Aryan, T. Singh, and D. Yadav, "Mem0: Building Production-Ready AI Agents with Scalable Long-Term Memory," arXiv preprint arXiv:2504.19413, 2025.

[40] L. Newsham and D. Prince, "Personality-Driven Decision-Making in LLM-Based Autonomous Agents," in Proc. 24th Int. Conf. Autonomous Agents and Multiagent Systems, 2025.

[41] S. Kulkarni and Y. Kulkarni, "Benchmarking Multi-Agent LLM Architectures for Financial Document Processing: A Comparative Study of Orchestration Patterns, Cost-Accuracy Tradeoffs and Production Scaling Strategies," arXiv preprint arXiv:2603.22651, 2026.

[42] J. Liao, F. Chang, Y. Xue, T. Xia, Z. Huang, and Y. Wang, "Multi-Objective Scheduling for Large Language Model Inference with Prompt-Level Cost Prediction and SLO Awareness," 2026.

[43] A. Singh, A. Ehtesham, S. Kumar, T. T. Khoei, and A. V. Vasilakos, "Agentic Retrieval-Augmented Generation: A Survey on Agentic RAG," arXiv preprint arXiv:2501.09136, 2025.

[44] C. Park, Z. Chen, A. Ozdaglar, and K. Zhang, "Post-Training LLMs as Better Decision-Making Agents: A Regret-Minimization Approach," arXiv preprint arXiv:2511.04393, 2025.

[45] R. Milsom, "Benchmarking LLM Agents for Wealth-Management Workflows," arXiv preprint arXiv:2512.02230, 2025.

[46] H. Zheng, Y. Ma, Y. Wang, G. Liu, Z. Qi and X. Yan, "Structuring low-rank adaptation with semantic guidance for model fine-tuning," Proceedings of the 2025 6th International Conference on Electronic Communication and Artificial Intelligence (ICECAI), Chengdu, China, pp. 731-735, 2025.

[47] H. Zheng, L. Zhu, W. Cui, R. Pan, X. Yan and Y. Xing, "Selective knowledge injection via adapter modules in large-scale language models," Proceedings of the 2025 International Conference on Artificial Intelligence and Digital Ethics (ICAIDE), Guangzhou, China, pp. 373-377, 2025.

[48] X. Yan, J. Du, X. Li, X. Wang, X. Sun, P. Li and H. Zheng, "A Hierarchical Feature Fusion and Dynamic Collaboration Framework for Robust Small Target Detection," IEEE Access, vol. 13, pp. 92953-92964, 2025, doi: 10.1109/ACCESS.2025.3570669.

[49] J. Wei, Y. Liu, X. Huang, X. Zhang, W. Liu and X. Yan, "Self-Supervised Graph Neural Networks for Enhanced Feature Extraction in Heterogeneous Information Networks", 2024 5th International Conference on Machine Learning and Computer Application (ICMLCA), pp. 272-276, 2024.

[50] X. Yan, W. Wang, M. Xiao, Y. Li, and M. Gao, "Survival prediction across diverse cancer types using neural networks", Proceedings of the 2024 7th International Conference on Machine Vision and Applications, pp. 134-138, 2024.

[51] X. Yan, Y. Jiang, W. Liu, D. Yi and J. Wei, "Transforming Multidimensional Time Series into Interpretable Event Sequences for Advanced Data Mining," 2024 5th International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI), pp. 126-130, 2024.

[52] X. Yan, J. Du, L. Wang, Y. Liang, J. Hu and B. Wang, "The Synergistic Role of Deep Learning and Neural Architecture Search in Advancing Artificial Intelligence", Proceedings of the 2024 International Conference on Electronics and Devices, Computational Science (ICEDCS), pp. 452-456, Sep. 2024.

[53] M. Xiao, Y. Li, X. Yan, M. Gao, and W. Wang, "Convolutional neural network classification of cancer cytopathology images: taking breast cancer as an example," Proceedings of the 2024 7th International Conference on Machine Vision and Applications, pp. 145–149, Singapore, Singapore,2024

[54] W. Wang, Y. Li, X. Yan, M. Xiao and M. Gao, "Breast cancer image classification method based on deep transfer learning," Proceedings of the International Conference on Image Processing, Machine Learning and Pattern Recognition, pp. 190-197, 2024.

[55] Y. Li, X. Yan, M. Xiao, W. Wang and F. Zhang, "Investigation of Creating Accessibility Linked Data Based on Publicly Available Accessibility Datasets", Proceedings of the 2023 13th International Conference on Communication and Network Security, pp. 77-81, 2024.

[56] Wang, J. (2024). Multivariate time series forecasting and classification via gnn and transformer models. Journal of Computer Technology and Software, 3(9).

[57] Wang, Y. (2024). AI-Enhanced Distributed Time Series Modeling: Incremental Learning for Evolving Streaming Data. Transactions on Computational and Scientific Methods, 4(8).

[58] Su, C., Jiang, L., & Pi, S. (2021, August). Sentiment classification of chinese railway review text based on multi-feature fusion gated recurrent unit. In 2021 International Conference on Information Science, Parallel and Distributed Systems (ISPDS) (pp. 197-202). IEEE.

[59] Li, S., Wang, Y., Xing, Y., & Wang, M. (2025, December). Mitigating correlation bias in advertising recommendation via causal modeling and consistency-aware learning. In Proceedings of the 2025 6th International Conference on Computer Science and Management Technology (pp. 585-589).

[60] Ke, Z. (2026). Right-Sizing Cloud Data Warehouses Under Asymmetric Loss via One-Sided Conformal Prediction. IEEE Access, 14, 94748-94763.

[61] Wang, J. (2025). Credit card fraud detection via hierarchical multi-source data fusion and dropout regularization. Transactions on Computational and Scientific Methods, 5(1).

[62] Xue, Y., Wang, Y., Zhu, A., Sun, X., & Zhang, C. (2026, March). Joint Temporal-Structural Representation Learning for Distributed Fault Discrimination in Microservice Architectures. In 2026 IEEE 8th International Conference on Communications, Information System and Computer Engineering (CISCE) (pp. 1581-1586). IEEE.

[63] S. Sun, Y. Wang, R. Yan, J. Wang and Y. Lu, "CurateDistill: Efficient Data Selection via Active Data Curation and Dataset Distillation," 2026, DOI: 10.13140/RG.2.2.27464.28166.

[64] Mao, L. (2024). Joint Optimization of Dual Retrieval and Adaptive Re-Ranking for Retrieval-Augmented Generation. Artificial Intelligence and Computing Innovations, 4(4).

[65] Huang, S. Y. (2024). Retrieval-Augmented Long-Text Reasoning with Semantic Conflict Detection and Cross-Paragraph Evidence Integration. Artificial Intelligence and Computing Innovations, 4(1).

[66] Xu, Y., Kou, J., Wang, W., Su, Z., Yan, X., & Wang, J. (2026, March). Graph spatiotemporal attention-based collaborative demand forecasting for multi-level supply chains. In Proceedings of the 2026 3rd International Conference on Innovation Management and Information Systems (pp. 214-220).

[67] Su, Z. (2024). Microstructure-Aware Risk Detection in High-Frequency Trading Using Embedding and Sequence Modeling. Transactions on Computational and Scientific Methods, 4(3).

[68] Kou, J., Qing, Z., Wei, R., Xue, F., Huang, W., & Zhuang, H. (2026, April). CausalRCA: Causal Graph-Augmented Retrieval for Root Cause Analysis in Microservice Systems. In 2026 5th International Conference on Electronic Information Engineering, Big Data and Computer Technology (EIBDCT) (pp. 440-444). IEEE.

[69] F. Xue, T. Ying, T. Ding, L. Wang, Z. Qing and P. Hu, "Requirement-Aware LLM Testing for Backend REST APIs," 2026, DOI: 10.13140/RG.2.2.21048.30729.

[70] Hu, P., Z. Qing, T. Ding, T. Ying, F. Xue, and L. Wang, "Energy-Efficient Distributed Inference across the Edge-Cloud Continuum Using Adaptive Model Compression," 2026, DOI: 10.13140/RG.2.2.19323.07209.

[71] Lyu, Y., Jiang, J., & Hu, J. (2026). QueryMind: Workload-Aware LLM-Guided Database Configuration Tuning for Backend Performance.

[72] Hu, J., Lyu, Y., & Jiang, J. (2026). High-Reliability LLM Summarization through Causal Cue Modeling and Factual Correction.

[73] S. Pi, S. Zhang and S. Shawulieti, "CacheRoute: KV-Cache-Aware Routing for LLM Backend Services on Kubernetes," 2026, DOI: 10.13140/RG.2.2.34191.55206.

[74] Wang, S., B. Su, Y. Liu, S. Pi and A. Zhu, "Safe Adaptive Resilience Configuration for Microservice Backends," 2026, DOI: 10.13140/RG.2.2.30702.57921.

[75] Z. Liu, K. Wu, S. Dang and Y. Yang, "A Retrieval-Augmented Log-Metric-Trace Twin for Backend Failure Diagnosis," 2026, DOI: 10.13140/RG.2.2.34050.64969.

[76] Z. Liu, Y. Yang, C. Chen, S. Dang and K. Wu, "Topology Dependency Modeling and Discriminative Representation Learning Methods for Backend Microservice Call Anomaly Detection Task," 2026, doi: 10.13140/RG.2.2.18843.02082.

[77] Shawulieti, S. (2024). Enhancing Cloud Security with Graph Neural Network-Driven Anomaly Detection in Multi-Tenant Environments. Artificial Intelligence and Computing Innovations, 4(2).

[78] Xue, F., Wu, K., & Abdulla, A. (2025). Topology-Aware Graph Neural Networks for Distributed Node Fault Identification. Transactions on Computational and Scientific Methods, 5(3).

[79] Wang, M., Peng, P. L., & Chen, J. (2024). Financial Report Summarization via Structure-Aware Modeling and Information Fidelity Constraints. Transactions on Computational and Scientific Methods, 4(7).

[80] Peng, Z. (2024). Cross-Period Financial Statement Anomaly Detection via Contrastive Representation Learning. Journal of Computer Technology and Software, 3(8).

[81] Z. Su, J. Chen, X. Meng, N. Sang and Z. Peng, "Relation-Aware Graph Neural Networks with Verifiable Reasoning for Accounting Fraud Detection," ResearchGate Preprint, 2026, DOI: 10.13140/RG.2.2.13631.06565.

[82] Hu, P., Uziel, G., & Ma, X. (2024). Risk-Constrained Reinforcement Learning for SLO-Driven Cloud-Native Elastic Scaling. Journal of Computer Technology and Software, 3(1).

[83] Jiang, Z., Zhang, S., Huang, W., Pi, S., Shawulieti, S., & Xu, S. (2026). Semantic-Aware Intelligent Request Routing in Multi-Model LLM Backend Gateways. Semantic-Aware Intelligent Request Routing in Multi-Model LLM Backend Gateways.

[84] Y. Jiang, Z. Xiao, O. Lin, Z. Gu and Z. Yang, "PrivCollab-Agent: Privacy-Preserving Multi-Agent Collaboration under Partial Information," 2026, DOI: 10.13140/RG.2.2.33663.16800.

[85] Ma, L., Song, T., Long, R., Mao, L., & Gu, Z. (2026, May). Evidence-Gated Memory Writing for Personalized Large Language Model Agents. In 2026 3rd International Conference on Image Processing and Artificial Intelligence (ICIPAI) (pp. 333-338). IEEE.

[86] Gu, Z., & Nie, L. (2025). Knowledge Graph-Enhanced Large Language Models for Opinion Target Sentiment Classification. Journal of Computer Technology and Software, 4(7).

[87] S. Li, B. Chen, Y. Li, Z. Wang, Y. Xue and C. Xu, "Privacy-Preserving Anomaly Detection in Cloud Services Using Hierarchical Federated Learning With Differential Privacy," 2026 9th International Symposium on Big Data and Applied Statistics (ISBDAS), pp. 668-672, 2026.

[88] Lyu, N., Chen, F., Zhang, C., Shao, C., & Jiang, J. (2025, December). Deep temporal convolutional neural networks with attention mechanisms for resource contention classification in cloud computing. In Proceedings of the 2025 5th International Conference on Big Data, Artificial Intelligence and Risk Management (pp. 62-68).

[89] Zhang, C. (2024). Adaptive Multi-Tenant Resource Scheduling in Cloud Computing via Reinforcement Learning. Transactions on Computational and Scientific Methods, 4(3).

[90] Y. Li, W. Zhao, B. Dang, X. Yan, M. Gao, W. Wang, and M. Xiao, "Research on adverse drug reaction prediction model combining knowledge graph embedding and deep learning", Proceedings of the 2024 4th International Conference on Machine Learning and Intelligent Systems Engineering (MLISE), pp. 322-329, June 2024.