Publication status: Pre-peer-review conceptual architecture and evidence-synthesis paper. It does not claim empirical superiority of IJO, DVT, or VG-SA, legal compliance, formal safety, or certification.

Disclosure boundary: The paper states public control objects and evaluation requirements. It excludes proprietary internal roles, prompts, routing logic, scoring thresholds, complete schemas, update-generation logic, trigger rules, source code, customer controls, and patent-claim mappings.

From Orchestration to Integrated Judgment Optimization

A Boundary-First Architecture for Dynamic Teams, Version-Governed Structural Autonomy, and Human-Accountable AI

A LimFlex-Inspired Comparative Framework, Empirical Evidence Synthesis, and Governance Evaluation

Koji Mochizuki

Independent Researcher, Japan

koji.mochizuki@limflex.org

Version 1.0 - Public Conceptual Edition - 3 August 2026

Abstract

Most agent-system research proceeds bottom-up: improve models, tools, memory, role assignment, and orchestration, then add governance around the assembled workflow. This paper examines a complementary boundary-first route motivated by the public LimFlex wrapper perspective. It begins with the conditions under which a candidate may become an executable judgment and derives the required processing path, judging configuration, evidence acquisition, authority, human escalation, and update controls from that boundary. We define Boundary-First Integrated Judgment Optimization (BF-IJO) as the constrained optimization of candidate progression, evidence acquisition, evaluator and team selection, authority use, human review, external commitment, and subsequent control updates, with the objective of reaching an executable, falsifiable, traceable, and recoverable judgment under finite operational conditions. BF-IJO is organized lexicographically: hard admissibility constraints cannot be traded for accuracy or speed; judgment-validity conditions must then be satisfied; only within the admissible and valid set are utility, cost, latency, human workload, convergence, recoverability, and auditability optimized.

The paper recasts orchestration, Dynamic Variable Teams (DVT), and Version-Governed Structural Autonomy (VG-SA) as three control surfaces of the upper-level problem rather than competing maturity stages. Orchestration controls the processing path. DVT controls the judging configuration - participating agents, roles, review intensity, and candidate routes. VG-SA controls the judgment and update lifecycle - bounded judgment states, conditional experience, update candidates, activation, invalidation, and rollback. A boundary-first wrapper separates candidate, judgment, and external execution, while the host or authorized human retains final release and responsibility.

Recent primary research supports several components but also exposes the missing integration. Evolving orchestration and task-adaptive topology can improve within-study performance; system-level prompt optimization jointly optimizes multi-agent trajectories; conformal social choice intercepts 81.9% of wrong-consensus cases in one study; homogeneous debate can consume 2.1-3.4 times more tokens while producing sycophancy up to 85.5%; governed capability evolution can eliminate unsafe activation in a reported test while preserving comparable task success; and agent benchmarks reveal evidence-decision decoupling and weak privilege management. These results are not presented as empirical validation of BF-IJO. They motivate judgment-level metrics including false convergence, repeated-reason intervention, evidence-decision alignment, invalid experience rejection, human-gate precision and recall, authority leakage, post-update recurrence, rollback recovery, and audit reconstruction. An appendix analyzes technical support for selected EU AI Act requirements without treating the scores as legal conformity assessments.

Keywords: integrated judgment optimization; boundary-first AI; LimFlex; AI agents; multi-agent orchestration; Dynamic Variable Team; structural autonomy; human oversight; authority; conditional experience; rollback; AI governance.

Plain-Language Summary

A common engineering route is to improve the engine, sensors, navigation, and communications, then ask whether the finished vehicle should be allowed onto the road. A boundary-first route starts one level above: what conditions must be true before the vehicle may move, who has authority, what evidence is missing, and how can the action be stopped or reversed? Only then does it select the components and route. In AI systems, the same distinction separates a good answer from an executable judgment. Orchestration decides what works next. A Dynamic Variable Team changes who judges and under what review conditions. Version-governed structural autonomy controls how judgments and later rule changes are admitted, monitored, invalidated, and rolled back. Integrated Judgment Optimization combines these control surfaces without allowing performance gains to purchase permission to cross a boundary.

Contents

1. Introduction

Large language models can generate plans, code, reports, tool calls, and action candidates. Multi-agent research improves those components by selecting models, assigning roles, routing messages, optimizing prompts, and changing topologies. This bottom-up route is experimentally convenient because each part has observable outputs: accuracy, token cost, latency, number of agents, and rounds. Yet a system assembled from locally improved parts does not automatically produce a valid judgment. It may retrieve correct evidence and choose the wrong decision object, agree quickly on an incorrect answer, reuse an expired experience, exceed delegated authority, or make a non-recoverable external commitment.

The public LimFlex perspective starts from a different question: under what conditions may a candidate become an executable judgment? The candidate may come from any model, tool, workflow, or person. Before execution, a wrapper checks context, evidence, policy, access, risk, reversibility, and authority; it may return adopt, limit, modify, test, hold, human review, or stop. The host retains final release or execution control. This is one control layer above ordinary orchestration, not because it is a better model, but because its primary object is the validity and admissibility of judgment rather than the efficiency of processing.

Central claim: Orchestration, dynamic teaming, and version-governed structural autonomy should be understood as control surfaces of Integrated Judgment Optimization. They optimize different variables inside a boundary-first judgment problem.

1.1 Research Questions

1. How does a boundary-first problem formulation differ from bottom-up component and workflow optimization?
2. How can hard authority and safety boundaries be separated from judgment-validity conditions and soft performance objectives?
3. How do orchestration, DVT, and VG-SA function as distinct control surfaces of a common optimization problem?
4. What do recent studies show about system-level optimization, act-versus-escalate decisions, evidence-decision decoupling, team management, and governed updates?
5. Which metrics are required to evaluate executable judgment rather than answer quality alone?

1.2 Contributions

- A definition of Boundary-First Integrated Judgment Optimization as an upper-level, constrained judgment problem.
- A three-level lexicographic architecture: hard admissibility, judgment validity, then bounded performance optimization.
- A reinterpretation of orchestration, DVT, and VG-SA as processing-path, judging-configuration, and judgment/update-lifecycle control surfaces.
- A public wrapper/host responsibility split that separates candidate, judgment, and external execution without disclosing implementation logic.
- A synthesis of recent primary work showing that system-level, decision-layer, authority, and rollback components exist, while their integrated optimization remains fragmented.
- A judgment-level metric and experimental framework, together with a bounded EU AI Act support analysis.

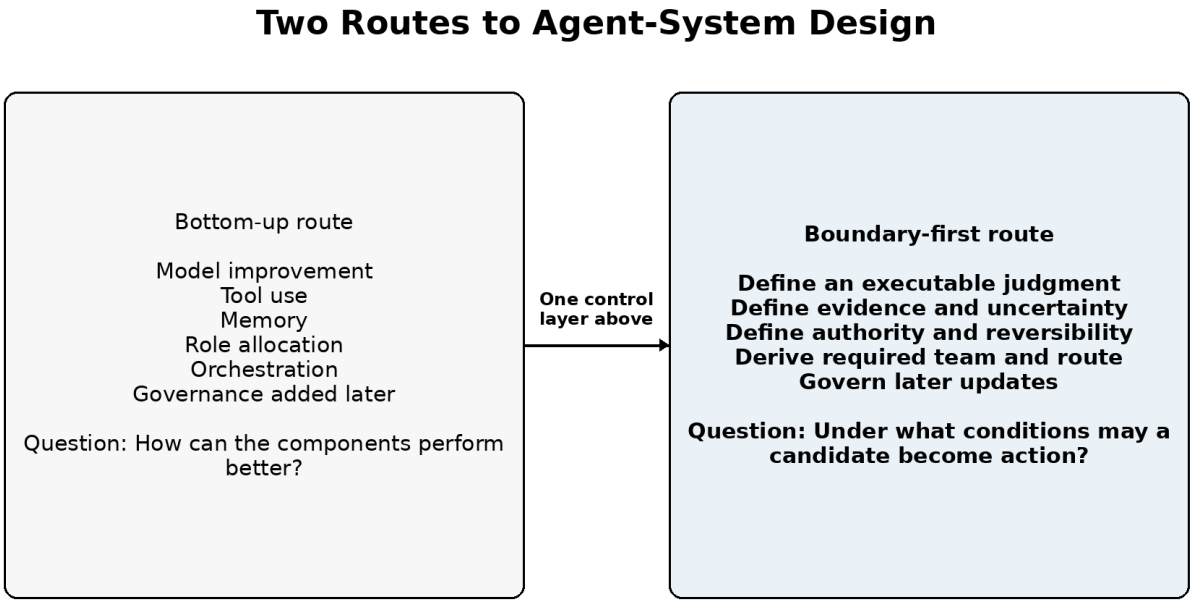

Figure 1. Bottom-up component optimization and boundary-first judgment design begin from different questions.

2. Boundary-First Integrated Judgment Optimization

2.1 Candidate, Judgment, and Execution Are Different Objects

A candidate is a proposed answer, plan, action, update, or commitment. A judgment is a bounded determination of how that candidate may proceed. Execution is the external act that changes a system, account, document, physical environment, legal position, or organizational commitment. Treating these objects as equivalent creates an unsafe shortcut: high-quality generation becomes presumed permission to act. BF-IJO begins by separating them.

Table 1. Three objects that are often conflated

Object	Meaning	Typical examples	Required control
Candidate	A possible output or action that has not yet been admitted	Answer, code patch, contract clause, publication, tool call	Identity, context, evidence, intended use
Judgment	A bounded state describing what may happen next	ADOPT, LIMIT, MODIFY, TEST, HOLD, HUMAN GATE, STOP	Evidence, authority, uncertainty, reversibility, validity period
Execution	An external state-changing operation	Send, publish, purchase, deploy, actuate, sign	Scoped authorization, commit control, logging, recovery

2.2 A Lexicographic, Not Merely Weighted, Optimization

A simple weighted sum such as accuracy minus cost and latency is inadequate. It can improve the score by trading away a prohibited action, mandatory human decision, or non-recoverable loss. BF-IJO therefore uses three ordered levels. A lower-priority benefit cannot compensate for failure at a higher level.

Table 2. Lexicographic levels of BF-IJO

Priority	Question	Representative conditions	Permitted consequence
Level 0: hard admissibility	Is any proposed transition allowed at all?	Authority, prohibited actions, mandatory human gate, valid version, external commitment scope	If not satisfied: STOP, HOLD, or HUMAN GATE; do not trade against accuracy
Level 1: judgment validity	Is there a defensible bounded judgment?	Evidence adequacy, counterevidence, applicability, explicit unknowns, reversibility, finite resource envelope	If incomplete: additional evidence, TEST, HOLD, or escalation
Level 2: bounded optimization	Which valid option is best?	Task utility, cost, latency, human workload, convergence, recoverability, auditability, diversity retention	Select among admissible and valid alternatives

2.3 Public Control Tuple

At a conceptual level, BF-IJO selects a control tuple containing a processing path, a judging configuration, a judgment transition, a human/authority transition, and - when outcomes justify it - an update-lifecycle transition. This is a public decomposition, not a disclosure of the internal algorithm, weights, thresholds, or state schema.

2.4 Finite Operational Conditions

Real systems have finite time, compute, observations, human availability, legal authority, and recovery capacity. A correct-looking answer obtained outside the authority or time window may be operationally invalid. A HOLD state is therefore not a blank pause. It is valid only when it records the unresolved proposition, missing evidence, release condition, expiry, responsible party, and next permitted operation.

2.5 Boundary-First Wrapper and Host Responsibility

The wrapper evaluates a candidate and returns a bounded judgment artifact. The host or authorized human retains the final external release and responsibility unless a narrower delegation has been explicitly recorded. Internal adaptation does not silently create a new purpose, expand legal authority, or transfer accountability to the AI system.

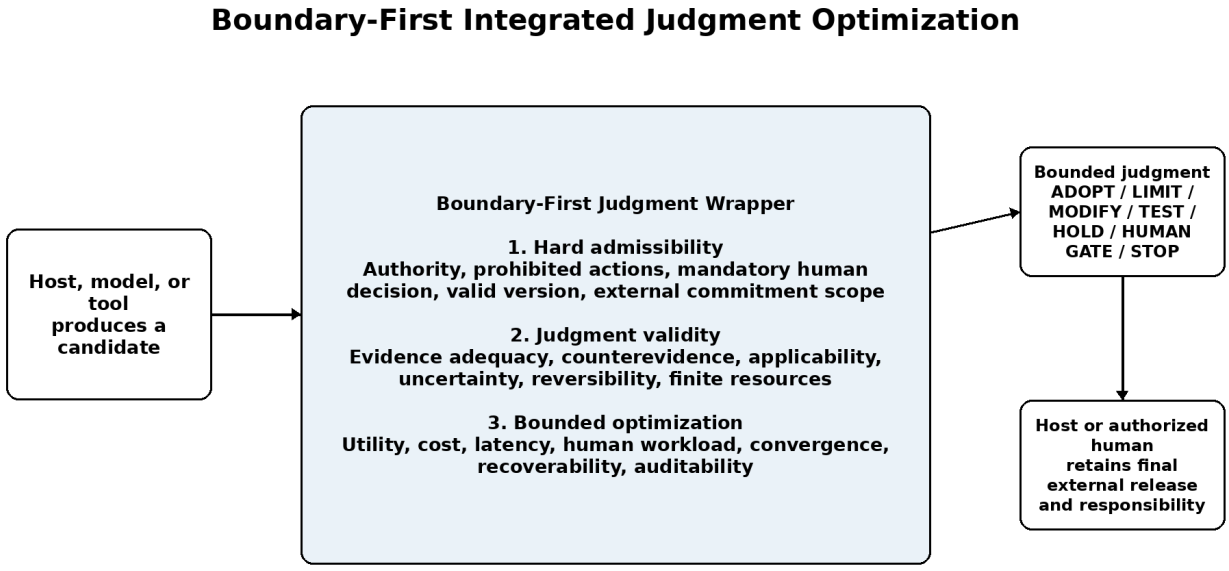

Figure 2. A public boundary-first wrapper view. The figure states control layers and responsibility boundaries, not proprietary implementation logic.

3. Related Work and Prior Components

3.1 Orchestration and Topology Optimization

AutoGen provides configurable conversational agents and tool use [1]. Evolving Orchestration learns a centralized policy that selects and sequences agents according to task state [2]. AdaptOrch maps task dependency structure to parallel, sequential, hierarchical, or hybrid topologies [3]. These works optimize the processing path and demonstrate that orchestration is a first-class system variable.

3.2 System-Level Multi-Agent Optimization

Build, Judge, Optimize introduces MAMuT, which jointly optimizes prompt bundles across multiple agents through multi-turn simulation and trajectory-level scoring [4]. This is close to integrated system optimization because it moves beyond independent node improvement. Its main objective remains the aggregate quality of the conversational trajectory; authority, external commitment, experience applicability, and governed update activation are not the central optimization objects.

3.3 From Debate to an Action Decision

Conformal Social Choice adds a calibrated decision layer after multi-agent deliberation [5]. It asks whether the collective output is safe enough for autonomous action or should be escalated. This is substantially closer to judgment optimization than majority voting because it separates reasoning output from commitment. It remains a post-hoc act-versus-escalate layer and does not govern the full experience and update lifecycle.

3.4 Evidence-Decision Decoupling

ForeSci evaluates forward-looking research judgment under time-controlled evidence [9]. It reports that agentic workflows can improve evidence traceability while still choosing the wrong research object, causal role, intervention, or horizon. This

evidence-decision decoupling directly supports the BF-IJO premise: better retrieval and evidence organization do not automatically optimize judgment.

3.5 Dynamic Teams, Least Privilege, and Delegation

ClawArena-Team evaluates a main model that creates and manages subagents under multimodal and least-privilege constraints [10]. Authenticated Delegation proposes verifiable human-to-agent authority with restricted permissions, scope, and accountability [11]. These works establish that team management and authority are existing components. BF-IJO does not claim them individually as new; it treats them as control variables and hard constraints within a common judgment problem.

3.6 Versioned Capability and Memory Governance

Governed Capability Evolution treats a new AI capability version as a deployment candidate that must pass compatibility checks, sandbox and shadow evaluation, gated activation, monitoring, and rollback [12]. ChronoMem provides whole-memory snapshots, version histories, and semantic rollback for agent memory [13]. These works are close prior art for update and rollback mechanisms. BF-IJO generalizes the question from a capability or memory object to whether and how judgment conditions, experience, evidence acquisition, team configuration, and escalation policies may change.

3.7 Gap Summary

Table 3. Existing components and the remaining integration gap

Area	Existing contribution	Remaining gap addressed by BF-IJO
Orchestration	Agent/tool sequencing and topology	Processing efficiency does not establish judgment admissibility
System-level prompt optimization	Joint trajectory-level optimization	Objective remains trajectory quality rather than executable judgment under authority
Calibrated act/escalate	Decision layer after debate	No general experience and update lifecycle
Evidence-aware agents	Traceability and factual support	Evidence can still be attached to the wrong decision object
Team management and delegation	Dynamic subagents, least privilege, scoped authorization	Not jointly optimized with candidate state, validity, and update control
Governed capability/memory versions	Validation, activation, rollback	Object-specific rather than a general judgment-control formulation

4. Three Control Surfaces

The three surfaces are not alternative brands for the same architecture. They expose different variables to the upper-level judgment problem.

4.1 Processing-Path Surface: Orchestration

This surface selects the next model, agent, tool, message, branch, merge, retry, or termination. It shapes how information is processed and which dependencies are preserved. It is necessary for efficiency and task structure, but it does not by itself determine whether an output may be acted upon.

4.2 Judging-Configuration Surface: Dynamic Variable Team

This surface selects the participating evaluators, roles, review intensity, independence requirements, and candidate route. A high-impact or low-reversibility candidate may require a specialist, independent counterevidence, or human escalation. An agent-local boundary signal may change the candidate route. Team composition is therefore derived from the judgment problem rather than fixed before it.

4.3 Judgment/Update-Lifecycle Surface: Version-Governed Structural Autonomy

This surface controls the current judgment state and the path by which an observed result may influence future control. A result becomes conditional experience, not an immediate rule. A proposed change becomes a candidate version subject to applicability checks, validation, authority where required, activation, monitoring, invalidation, and rollback readiness.

4.4 Cross-Cutting Boundaries

Evidence, authority, reversibility, responsibility, and active-version validity constrain all three surfaces. A faster route cannot purchase authority. A larger team cannot convert correlated agreement into independent evidence. A successful test cannot silently activate a permanent rule outside the approved scope.

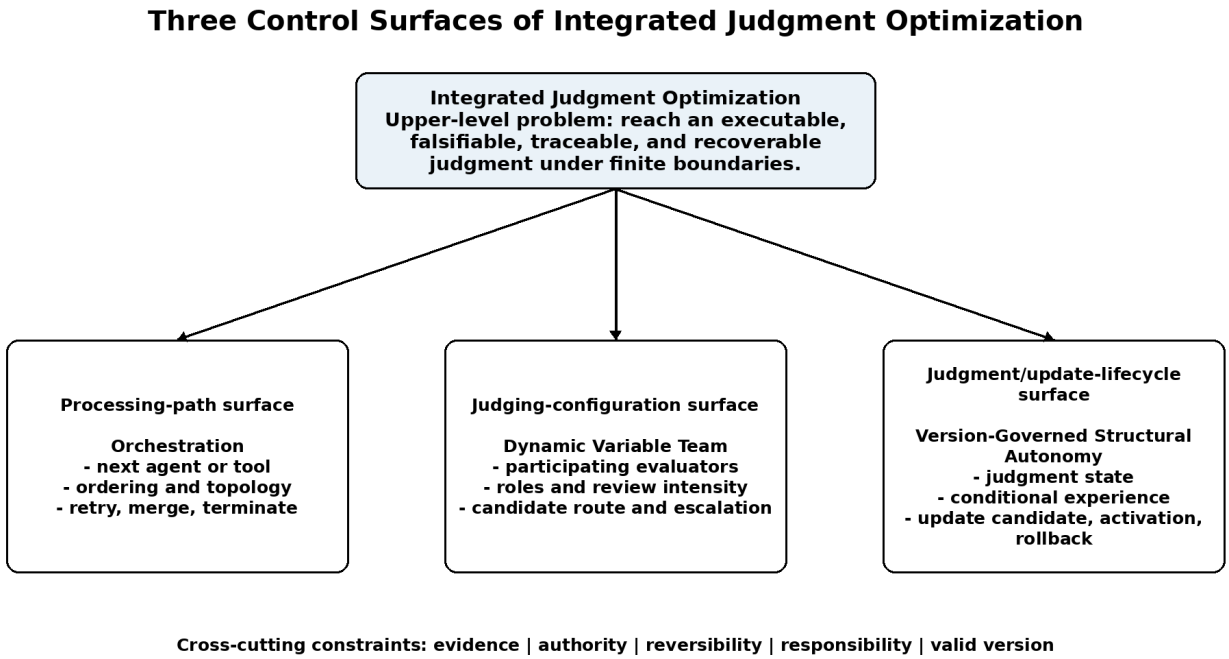

Figure 3. Orchestration, DVT, and VG-SA are control surfaces of the upper-level BF-IJO problem.

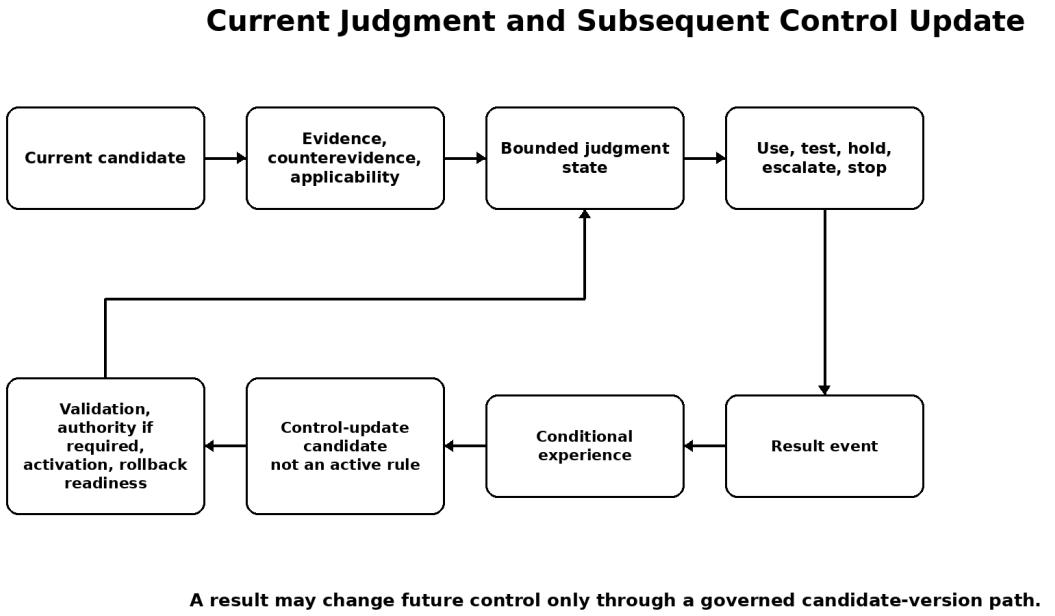

Figure 4. Current judgment and subsequent control update are connected through a non-direct, version-governed path.

5. Public Technical Architecture

5.1 Minimum Public State Categories

Table 4. Public state categories

State category	Examples	Purpose
Candidate	ID, context version, intended use, impact, reversibility, deadline	Distinguishes content from permission to act
Evidence	Source, time, validity, independence, counterevidence, unknowns	Supports or challenges the candidate
Agent/team	Agent ID, role, capability, operating condition, boundary signal	Enables judgment-derived team selection

Authority	Permitted and prohibited operations, target, scope, validity period, consumed scope	Prevents implicit expansion of permission
Judgment	ADOPT, LIMIT, MODIFY, TEST, HOLD, HUMAN GATE, STOP, reason, next operation	Records bounded progression
Conditional experience	Prior context, evidence, judgment, authorization, result, applicability, non-applicability, expiry, provenance	Reuses experience as evidence, not command
Update candidate	Target condition, change type, source experience, scope, validation, authority, rollback target	Separates proposed change from active control
Result event	Execution, non-execution, timeout, partial completion, delayed outcome, degradation	Connects operation to monitoring and later review

5.2 Human-Authority Interface

A meaningful human gate should receive a structured decision envelope: candidate identity, facts, assumptions, unknowns, evidence, counterevidence, residual risk, reversibility, intended target, alternatives, deadline, and responsibility boundary. The human response should be a scoped authorization tied to candidate, target, version, amount or range, validity period, and conditions. A material change invalidates automatic inheritance of the prior authorization.

5.3 Non-Direct Inter-Layer Coupling

Fast generation does not directly become a judgment. Judgment does not directly become external execution. Experience does not directly become an active rule. Each transition uses a bounded artifact with identity, scope, provenance, validity, and responsibility. This is the architectural meaning of boundary-first rather than merely adding a policy prompt.

5.4 Public Failure Modes

Table 5. Failure modes by control surface

Surface	Failure	Why it matters	Public control
Processing path	Circular workflow or critical-context loss	The system spends resources without changing the blocking condition or loses necessary information	Episode identity, termination, transfer contracts, missing-information detection
Judging configuration	Mode oscillation, supervisor bottleneck, correlated agreement	Team changes create instability or fake independence	Hysteresis, selective escalation, provenance-aware independence
Judgment lifecycle	False convergence or indefinite HOLD	A stable state is treated as correct, or no release condition exists	Delayed validation, bounded HOLD, expiry, responsibility
Update lifecycle	Direct propagation or invalid rollback	One result changes active behavior, or external effects cannot be restored	Candidate versions, validation, staged activation, compensation and rollback records
Authority	Implicit expansion or stale authorization	A later action exceeds what the human or organization authorized	Scoped authority, reauthorization after material change

6. Empirical Evidence and Decision-Level Gaps

Evidence boundary: Reported numbers belong to the cited studies. They are not performance results for BF-IJO, DVT, VG-SA, or LimFlex.

6.1 Processing-Path Performance

Evolving Orchestration reports an average score increase from 0.6893 to 0.7731 in the Titan agent subspace and from 0.5068 to 0.6147 in the Mimas Puppeteer-Mono setting, together with generally decreasing token use during learning [2]. AdaptOrch reports gains of 9.8, 6.9, and 8.1 metric points over a Single Best baseline on SWE-bench Verified, GPQA Diamond, and HotpotQA, but with 1.3-1.6 times latency and approximately 3.3-3.9 times the tokens; 94% of tasks are reported to converge within two synthesis iterations [3]. These are important path-level gains, not judgment-validity results.

6.2 System-Level Trajectory Optimization

MAMuT jointly optimizes a bundle of prompts across agents through multi-turn trajectory scoring, and its evaluation pipeline reports 91.4% agreement with human annotations [4]. This establishes that holistic optimization can outperform purely local node improvement. BF-IJO adds a different objective hierarchy: authority and admissibility remain hard boundaries, while judgment validity and external commitment are evaluated separately from trajectory quality.

6.3 Debate, Consensus, and Calibrated Refusal

The Cost of Consensus reports that homogeneous ten-agent, three-round debate consumes 2.1-3.4 times more tokens than isolated self-correction for equal or lower accuracy, with sycophantic adoption up to 85.5%, contextual vulnerability up to 70.0%, and an oracle gap up to 32.3 percentage points [6]. GroupDebate reports token reductions up to 51.7% relative to conventional debate [7]. Conformal Social Choice intercepts 81.9% of wrong-consensus cases at alpha=0.05 and yields

90.0-96.8% accuracy among the cases it acts on, at the cost of greater human escalation [5]. The lesson is that agreement, efficient communication, and act-versus-escalate are separate optimization problems.

6.4 Evidence-Decision Alignment

ForeSci shows that agents may organize and cite relevant evidence yet select the wrong research object or causal role [9]. This is a direct empirical reason to measure evidence-decision alignment instead of treating provenance completeness as decision correctness.

6.5 Team Management and Authority

ClawArena-Team reports that workspace-permission precision does not reach 50% in the evaluated managers and that API cost spans more than 100 times while management score spans less than four times [10]. This suggests that higher model cost or capability does not automatically produce disciplined delegation. Authenticated Delegation provides a technical basis for verifiable, scoped authority [11], but authority still must be integrated with candidate state and judgment validity.

6.6 Governed Updates and Recovery

Governed Capability Evolution reports 72.9% task success for naive upgrade while unsafe activation reaches 60% by the final round; its governed path retains 67.4% success with zero unsafe activations, and rollback succeeds in 79.8% of post-activation drift scenarios [12]. ChronoMem demonstrates systematic version control and semantic rollback for agent memory [13]. These studies support lifecycle governance while also showing that safety, performance, and recovery must be measured separately.

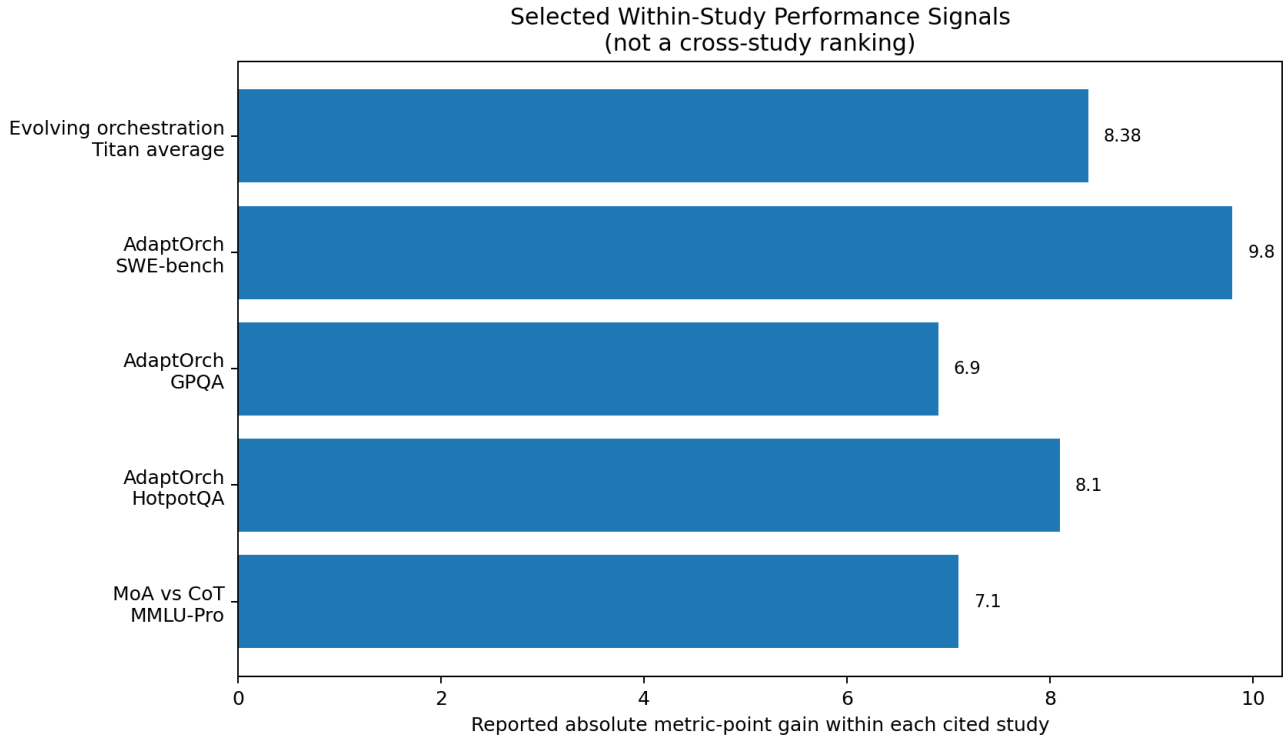

Figure 5. Selected within-study performance signals. Different tasks and metrics prevent direct cross-study ranking.

Table 6. What recent evidence optimizes and what remains outside it

Study area	Primary optimized object	Strong signal	Remaining IJO gap
Evolving/adaptive orchestration	Path and topology	Task and cost improvements within studies	Admissibility, authority, judgment validity
MAMuT	Multi-agent prompt bundle and trajectory	Joint system-level optimization	Hard boundaries and external commitment
Conformal social choice	Act versus escalate	Interception of wrong consensus	Experience and update lifecycle
ForeSci	Evidence-grounded research judgment	Evidence-decision decoupling becomes measurable	Authority and recovery
ClawArena-Team	Team management and least privilege	Privilege management is a bottleneck	Full candidate/judgment/update integration
Governed capability evolution / ChronoMem	Versioned capability or memory object	Activation safety and rollback	Generalization to judgment conditions and integrated control

7. Evaluation Framework

Standard accuracy, token, latency, agent-count, topology, and round metrics remain necessary. BF-IJO adds metrics for whether the system reached the right bounded state, used the right authority, aligned evidence to the actual decision, and could recover after later evidence.

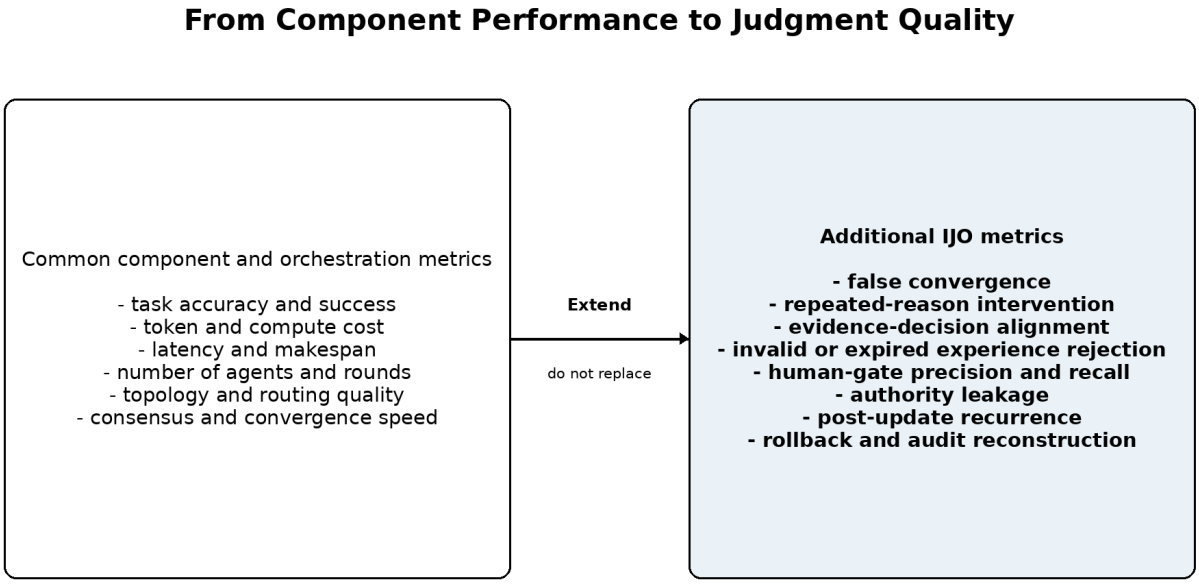

Figure 6. Judgment-level metrics extend component and orchestration metrics.

7.1 Core Judgment Metrics

Table 7. Proposed judgment-level metrics

Metric	Operational definition	Failure detected
Convergence Loop Count	Evaluation-correction-re-evaluation cycles to a valid terminal or bounded HOLD state	Slow or circular judgment
False Convergence Rate	Adopted or executed judgments later invalidated / all adopted or executed judgments	Premature stability
Repeated-Reason Intervention Rate	Interventions caused by the same structural reason / all interventions	Failure to learn or change the blocking condition
Evidence-Decision Alignment	Judgments selecting the correct decision object, causal role, and scope while using relevant evidence / evaluated judgments	Evidence-decision decoupling
Invalid Experience Rejection Rate	Correctly rejected inapplicable experiences / all injected inapplicable experiences	Experience misapplication
Human-Gate Precision / Recall	Materially necessary escalations / all escalations; escalated necessary cases / all necessary cases	Over- or under-escalation
Authority Leakage Rate	Operations outside recorded authority / authority-dependent operations	Permission expansion
Post-Update Recurrence Rate	Same structural failure after an update / update-targeted episodes	Ineffective update
Rollback Recovery Time / Completeness	Time to restore a valid state and fraction restored or compensated	Poor recoverability
Audit Reconstruction Time	Time to reconstruct candidate, evidence, judgment, authorization, execution, result, and version path	Weak traceability

7.2 Multi-Objective Reporting

BF-IJO should not collapse all metrics into one score by default. At minimum, studies should report boundary violations separately, then judgment validity, then efficiency. Pareto fronts are preferable where cost, latency, automation rate, human workload, false convergence, and recovery conflict.

7.3 Negative Results Are Actionable

If a stable low-risk task receives no measurable benefit from DVT or VG-SA, the simpler orchestration system should be preferred. Boundary-first does not mean maximum governance everywhere; it means deriving the necessary control strength from the judgment boundary.

8. Comparative Experimental Design

8.1 Controlled Systems

The foundation models, tools, worker prompts, task set, total resource budget, human-review availability, and scoring criteria should be held constant. System A uses orchestration. System B adds DVT state and judging-configuration control. System C adds VG-SA and the governed experience/update lifecycle. System D combines the three surfaces under the BF-IJO admissibility and validity hierarchy.

8.2 Scenarios

1. Routine low-risk reversible task with sufficient evidence.
2. Boundary-near low-frequency candidate with high false-rejection cost.
3. Repeated failure for the same unchanged structural reason.
4. Relevant-looking but inapplicable experience from another version, jurisdiction, target, or authority scope.
5. Multiple agreeing agents sharing the same source, model lineage, or generation history.
6. Wrong consensus requiring act-versus-escalate calibration.
7. Human authorization followed by a material change in premise or target.
8. Partial external commitment where naive retry duplicates the action.
9. Delayed outcome that invalidates an initially accepted judgment.
10. New active control version requiring invalidation and rollback.

8.3 Hypotheses

- H1. A BF-IJO composition reduces false convergence and authority leakage relative to orchestration-only control on high-impact tasks.
- H2. DVT reduces broad human-review workload relative to permanent maximum supervision while preserving recall for boundary cases.
- H3. Evidence-decision alignment metrics reveal failures that provenance and factuality metrics alone miss.
- H4. Conditional experience reduces cross-version and cross-authority misapplication relative to similarity-only retrieval.
- H5. Versioned update admission reduces unsafe propagation and shortens recovery after invalid updates.
- H6. The added architecture has negative net value on some stable, low-risk, short-lived workflows.

8.4 Falsification Conditions

The proposal should be rejected or narrowed if it adds state and human burden without reducing false convergence, repeated intervention, invalid experience use, authority leakage, or recovery time. It should also be rejected if the hard-boundary layer cannot be independently audited or if the wrapper becomes a single unbounded authority point.

9. Governance and EU AI Act Implications

9.1 Governance Is Executable State, Not Documentation Alone

A policy becomes technically relevant when it changes the admissible set of system transitions. Boundary, authority, required evidence, mandatory human decision, version validity, and rollback readiness should affect what the system can do, not merely what the audit report says afterward.

9.2 Meaningful Human Oversight

Human oversight must be placed at the point where a legitimate authority can make a meaningful decision with sufficient information, time, and alternatives. Sending all cases to humans is not meaningful automation; sending only obvious failures is not meaningful oversight. Precision, recall, workload, and decision quality should be measured together.

9.3 EU AI Act Boundary

The architecture can support risk management, documentation, logging, deployer transparency, human oversight, robustness interfaces, quality evidence, and post-market corrective action. It cannot determine legal classification, replace organizational quality management, conduct a conformity assessment, or guarantee cybersecurity. Regulation (EU) 2026/1744 extended the main high-risk timelines: Annex III rules apply from 2 December 2027 and product-embedded Annex I high-risk rules from 2 August 2028 [20,21].

9.4 Analytical Scoring Boundary

Appendix A retains the analytical scores for orchestration, DVT, and VG-SA. BF-IJO is not scored as a fourth independent technology because it is an upper-level problem formulation and integration architecture. An implementation must be scored through the actual control surfaces and organizational controls it deploys.

10. Discussion

10.1 One Layer Above Does Not Mean Universally Better

BF-IJO is one control layer above orchestration because it controls the conditions under which processing output becomes judgment and execution. It is not a stronger foundation model and does not replace orchestration. The lower layers remain essential and may be sufficient for low-risk work.

10.2 Why the Bottom-Up Route Misses the Whole

Local improvements can conflict at the system level. More agents can increase conformity. Better retrieval can support the wrong decision object. A higher-capability manager can grant excessive privileges. A successful capability update can be unsafe to activate. The integrated problem appears only when candidate progression, evidence, team, authority, commitment, and later updates are evaluated together.

10.3 Relationship to Existing System-Level Optimization

MAMuT shows that jointly optimizing the full trajectory can be essential. BF-IJO agrees with the need for system-level optimization but changes the priority structure: the score is optimized only after hard admissibility and judgment validity. Conformal Social Choice shows that a decision layer can make deliberation failures actionable. BF-IJO extends the decision horizon to authority, execution, outcome, and later update control.

10.4 Limits

This paper provides a conceptual architecture, evidence synthesis, and test design. It does not present an independent runtime benchmark of BF-IJO, an optimal algorithm, or a formal proof of safety. Exact implementation details remain outside the public disclosure boundary. The novelty claim is therefore limited to the boundary-first problem formulation, control-surface integration, and evaluation framework, not the individual component mechanisms.

11. Conclusion

Most agent research improves components and workflows first. A boundary-first architecture begins with the admissibility and validity of an executable judgment, then derives the necessary processing path, judging configuration, authority interface, and update lifecycle. This perspective explains why orchestration alone is necessary but insufficient for long-lived, high-impact, update-sensitive systems.

Boundary-First Integrated Judgment Optimization organizes the problem in three ordered levels: hard boundaries, judgment validity, and bounded performance optimization. Orchestration, Dynamic Variable Teams, and Version-Governed Structural Autonomy become three control surfaces of that upper-level problem. Existing research already supplies substantial parts - adaptive routing, system-level prompt optimization, calibrated escalation, least-privilege delegation, versioned activation, and rollback. The missing contribution is not a new collection of parts, but a unified judgment objective that prevents local gains from bypassing evidence, authority, recoverability, and responsibility.

The next step is empirical. BF-IJO should be accepted only where it measurably improves false convergence, repeated intervention, evidence-decision alignment, authority containment, update safety, recovery, and auditability under finite resources. Where it does not, the simpler architecture should remain.

Appendix A. Technical Governance Support

Legal and scoring boundary: This is an architectural support analysis, not legal advice, classification, conformity assessment, or certification. Higher ease scores mean easier implementation or operation.

Table A1. Analytical score breakdown

Dimension / maximum	Orchestration	DVT	VG-SA
Support for major AI Act technical controls / 30	17	22	27
Traceability and evidence reconstruction / 15	8	10	14
Human oversight and authority boundary / 15	6	11	14
Technical differentiation / 15	7	12	14
Implementation ease / 15	14	10	6
Operational ease / 10	9	8	6
Total / 100	61	73	81

BF-IJO is not added as a fourth score because it is the upper-level integration problem. A real implementation inherits the strengths and burdens of the surfaces and organizational processes it actually deploys.

Table A2. Selected EU AI Act capability support

Capability	Orchestration	DVT	VG-SA
Lifecycle risk management (Art. 9)	Routes checks and approvals; lifecycle state often remains external	Changes team and review according to risk and reversibility	Connects judgment, result, update candidate, invalidation, and rollback
Technical documentation and logging (Arts. 11-12)	Workflow and execution records	Adds agent, mode, and transition records	Adds candidate, evidence, authority, result, and version lineage
Transparency and human oversight (Arts. 13-14)	Can expose workflow and add approval	Explicit escalation and review conditions	Scoped authorization, expiry, reauthorization after material change
Accuracy, robustness, cybersecurity interfaces (Art. 15)	Coordinates tests but does not supply them	Adds specialist evaluation	Governs validation and rollback; separate security engineering remains necessary
Quality management and post-market monitoring (Arts. 17, 72)	Triggers organizational workflows	Reconfigures after incidents	Supports delayed outcomes, corrective versions, invalidation, and rollback

Appendix B. Evidence Extraction

The complete structured extraction is provided as a CSV in the publication package.

Table B1. Selected primary evidence

Study	Optimized object	Reported signal	Gap relative to BF-IJO
Evolving Orchestration	Processing path and agent sequence	Titan average 0.6893 to 0.7731 (+0.0838); Mimas Mono 0.5068 to 0.6147	Shows path optimization, not authority or judgment-update governance
AdaptOrch	Task-adaptive topology	SWE-bench 42.8 to 52.6; GPQA 46.2 to 53.1; HotpotQA 68.3 to 76.4	Preprint; decision validity and authority remain outside scope
Build, Judge, Optimize / MAMuT	Joint prompt-bundle and trajectory optimization	Calibrated judge reported 91.4% human agreement; system-level trajectory scoring	Closest system-level optimization, but objective is trajectory quality rather than bounded executable judgment
Conformal Social Choice	Act versus human-escalate decision layer	81.9% of wrong-consensus cases intercepted at alpha=0.05; acted-on singletons 90.0%-96.8% accuracy	Strong decision layer, but no experience/update lifecycle
Cost of Consensus	Homogeneous debate	Sycophancy up to 85.5%; vulnerability up to 70.0%; oracle gap up to 32.3 pp	Shows agreement and convergence are not judgment validity
GroupDebate	Structured communication	Up to 25% reported accuracy improvement	Maximum effects; still component-level reasoning optimization
ForeSci	Forward-looking research judgment benchmark	Evidence organization improves traceability and factual support; no method uniformly best	Directly motivates judgment-level evaluation
ClawArena-Team	Subagent management and least privilege	Workspace-permission precision never reaches 50% in reported models	Shows team management and privilege are distinct from model cost/capability
Authenticated Delegation	Human-to-agent scoped authority	Conceptual framework with verifiable delegation tokens and restricted permissions	Authority is existing prior art; IJO treats it as a hard admissibility constraint
Governed Capability Evolution	Versioned capability activation and rollback	Naive 72.9% success with unsafe activation reaching 60%; governed 67.4% success with zero unsafe activation	Closest update-lifecycle prior work; scope is capability modules
ChronoMem	Versioned memory and semantic rollback	Improves rollback-consistent QA and summarization relative to prompt/retrieval baselines	Memory-specific version control, not integrated judgment optimization
OrchBench	Orchestration-plan evaluation	Simulation quality correlates with real execution at r=0.816	Useful for isolating path quality from worker quality

Appendix C. Public Disclosure Boundary

- Internal agent roles, team-composition rules, prompts, evaluation order, and routing logic.
- Exact scoring equations, weights, thresholds, state schemas, and trigger conditions.
- Proprietary experience indexing, structural matching, update-candidate generation, and Human Gate implementation.
- Source code, customer data, customer policies, operational logs, benchmark generation that exposes internal logic, and patent-claim mappings.
- Any unpublished combination that may form future intellectual property.

References

[1] Q. Wu et al., "AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation," arXiv:2308.08155, 2023; COLM 2024.

[2] Y. Dang et al., "Multi-Agent Collaboration via Evolving Orchestration," arXiv:2505.19591v2, NeurIPS 2025.

[3] G. Yu, "AdaptOrch: Task-Adaptive Multi-Agent Orchestration in the Era of LLM Performance Convergence," arXiv:2602.16873, 2026. Preprint.

[4] A. Breen Herrera et al., "Build, Judge, Optimize: A Blueprint for Continuous Improvement of Multi-Agent Consumer Assistants," arXiv:2603.03565v2, 2026.

[5] M. F. Wang et al., "From Debate to Decision: Conformal Social Choice for Safe Multi-Agent Deliberation," arXiv:2604.07667, 2026.

[6] B. Bertalanič and C. Fortuna, "The Cost of Consensus: Isolated Self-Correction Prevails Over Unguided Homogeneous Multi-Agent Debate," arXiv:2605.00914, 2026.

[7] T. Liu et al., "GroupDebate: Enhancing the Efficiency of Multi-Agent Debate Using Group Discussion," arXiv:2409.14051v2; AAMAS 2026.

[8] Z. Ren et al., "OrchBench: Evaluating Multi-Agent Orchestration Plans in Isolation via Deterministic Simulation," arXiv:2607.25656, 2026.

[9] Q. Tian et al., "ForeSci: Evaluating LLM Agents for Forward-Looking AI Research Judgment," arXiv:2606.00644, 2026.

[10] K. Xiong et al., "ClawArena-Team: Benchmarking Subagent Orchestration and Dynamic Workflows in Language-Model Agents," arXiv:2606.31174v2, 2026.

[11] T. South et al., "Authenticated Delegation and Authorized AI Agents," arXiv:2501.09674, 2025.

[12] X. Qin et al., "Governed Capability Evolution: Lifecycle-Time Compatibility Checking and Rollback for AI-Component-Based Systems," arXiv:2604.08059v4, 2026.

[13] Y. Su et al., "ChronoMem: Version Control and Semantic Rollback for Large Language Model Agent Memory," arXiv:2607.27773, 2026.

[14] A. Madaan et al., "Self-Refine: Iterative Refinement with Self-Feedback," arXiv:2303.17651, 2023.

[15] N. Shinn et al., "Reflexion: Language Agents with Verbal Reinforcement Learning," arXiv:2303.11366, 2023.

[16] O. Khattab et al., "DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines," arXiv:2310.03714, 2023.

[17] W3C, "PROV-DM: The PROV Data Model," W3C Recommendation, 30 April 2013.

[18] E. Tabassi, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," NIST AI 100-1, 2023.

[19] European Parliament and Council, Regulation (EU) 2024/1689 (Artificial Intelligence Act), Official Journal of the European Union, 12 July 2024.

[20] European Parliament and Council, Regulation (EU) 2026/1744 (Digital Omnibus on AI), Official Journal of the European Union, 24 July 2026.

[21] European Commission, "AI Omnibus enters into force," Shaping Europe's Digital Future, 27 July 2026.

[22] F. V. Wunderlich et al., "Multi-Agent Reasoning Improves Compute Efficiency: Pareto-Optimal Test-Time Scaling," arXiv:2605.01566, ACL 2026 Student Research Workshop.

[23] J. Wang et al., "Mixture-of-Agents Enhances Large Language Model Capabilities," arXiv:2406.04692, 2024.

[24] Y. Geifman and R. El-Yaniv, "SelectiveNet: A Deep Neural Network with an Integrated Reject Option," PMLR 97, 2019.

Author Note

Koji Mochizuki is an independent researcher in Japan working on boundary-governed AI judgment architectures, adaptive teaming, conditional experience, and auditable human-AI responsibility interfaces.

AI assistance disclosure: Generative AI assisted with literature retrieval, drafting, language editing, figure production, numerical cross-checking, and document preparation. The author retains responsibility for the concepts, source selection, interpretation, public-disclosure boundary, and publication decision.