SWISSI INSTITUTE FOR AI

Generic Agnostic AI and Distributed Ledger Enterprise System for Scalable Domain Adaptation

Architecture and Methodology for Vertical-Specific AI Deployment from a Unified Core Framework

Walter Kurz[1], Michel Malara[1], Velimir Dedić[1]

August 2025

Abstract

The objective of this study is to define a compliance-first, conceptually generalisable architecture for a multi-agent artificial intelligence platform integrated with distributed ledger technology, designed to be domain-, deployment-, and vendor-agnostic. It addresses a persistent shortcoming in current AI deployments, where compliance is often treated as a secondary concern, applied retroactively through prompt engineering rather than embedded within the foundational design. The proposed model encodes regulatory, governance, and ESG requirements into an objective-under-constraints framework, ensuring that all specialised agents operate within legally admissible and verifiably auditable parameters prior to any domain-specific implementation. A DAG-based verification layer is incorporated to enable scalable, low-latency, and cost-efficient operation while preserving evidentiary integrity. The analysis evaluates the feasibility of this conceptual model to support sustainable, rapid-deployment vertical applications without inducing vendor lock-in, preserving operational neutrality, and ensuring environmental accountability. The findings suggest that integrating compliance, ESG metrics, and agent specialisation at the architectural level provides a transferable foundation for cross-domain AI–DLT infrastructures.

Keywords: Compliance-first AI, Multi-agent systems, Distributed ledger technology, Directed acyclic graph, Regulation by design, ESG integration, Domain-agnostic architecture, Deployment-agnostic architecture, Vendor-agnostic architecture, Objective-under-constraints

1 Introduction

Contemporary consumer-facing artificial intelligence deployments are dominated by large language models (LLMs) and related generative architectures, typically accessed through proprietary API endpoints.[1, 2] In such systems, the generated output itself constitutes the primary product.[1]

Internal optimisation targets coherence, syntactic precision, and surface-level factuality, achieved through prompt engineering, instruction tuning, retrieval-augmented generation (RAG) pipelines, and post-processing moderation.[3, 4, 5, 6, 7] These mechanisms are effective in improving immediate relevance and reliability, particularly where the value of the system is measured by the quality of a single response.[6, 5]

Such output-centric architectures exhibit structural limitations in enterprise-scale, multi-agent systems operating in regulated domains.[8]

In consumer deployments, compliance, ethical alignment, and environmental, social, and governance (ESG) considerations are typically applied as non-binding guidelines or post hoc filters rather than embedded constraints.[8]

Provider control is exercised primarily through model training choices and prompt manipulation, which provides no systemic assurance that outputs, or the intermediate reasoning steps producing them, comply with procedural, legal, or evidentiary requirements.[9, 8]

From the provider's perspective, this optimisation objective can be formalised as the constrained problem in Equation (1), where the aim is to maximise perceived output quality within bounded constraints on compliance, factual accuracy, user-perceived helpfulness, stylistic acceptability, safety, and latency.

$$
\begin{aligned}
\max_{y \in \mathcal{Y}} \quad & Q_{\text{user}}(y, x, \theta) && && (1)\\
\text{s.t.} \quad & C_{\text{policy}}(y) = 1 && \text{(compliance with policies and applicable laws)}\\
& C_{\text{factual}}(y, \mathcal{K}) \geq \tau_f && \text{(factual correctness relative to knowledge base } \mathcal{K})\\
& C_{\text{context}}(y, x) \geq \tau_c && \text{(alignment with inferred user intent)}\\
& C_{\text{style}}(y) \geq \tau_s && \text{(stylistic acceptability and engagement quality)}\\
& C_{\text{safety}}(y) \geq \tau_\sigma && \text{(safety and risk minimisation)}\\
& T(y) \leq \tau_t && \text{(latency limit for output delivery)}
\end{aligned}
$$

Here, $y \in \mathcal{Y}$ denotes the output text sequence generated by the model in response to the input prompt x, with θ representing the model parameters. The scalar utility function Q_{user} estimates the perceived output quality from the end-user's perspective. The constraint C_{policy} enforces compliance with provider policies and applicable legal frameworks. Factual correctness is measured by C_{factual} relative to a knowledge base $\mathcal{K}$, bounded below by a threshold τ_f. Contextual alignment with inferred user intent is expressed as C_{context} with threshold τ_c, while C_{style} ensures stylistic and rhetorical acceptability above threshold τ_s. Safety and reputational risk minimisation are modelled through C_{safety} with threshold τ_σ, and $T(y)$ denotes the total generation latency, constrained by τ_t.

Enterprise-grade AI, by contrast, functions as a delegated agent on behalf of an organisation, whether corporate, governmental, medical, or otherwise, and must operate within a normative framework.[10, 11] In such contexts, accuracy of output is necessary but insufficient for admissibility.[9] All actions and decisions must be bounded by a formally defined *objective-under-constraints* environment in which compliance, ESG adherence, and procedural validity are structural guarantees rather than aspirational properties.[8, 9] This mirrors the legal and contractual obligations of human actors, whose conduct is constrained by performance requirements as well as enforceable compliance obligations.[8]

Within this frame, generative models, including LLMs, are best understood as modular components within a larger multi-agent architecture, rather than as the defining element of the system.[11] The proposed approach comprises two co-dependent layers: a modular, role-specific multi-agent system capable of both persistent and ad hoc committee formation for continuous and event-driven tasks; and a distributed ledger infrastructure, incorporating a directed acyclic graph (DAG) verification layer to deliver low-latency, tamper-evident auditability.[10, 12, 13] This design is deliberately deployment- and vendor-agnostic, enabling heterogeneous AI modules, generative or otherwise, to be integrated without compromising compliance or scalability.[8, 12]

2 Related Work and Research Gap

2.1 Multi-Agent Architectures in AI Systems

Multi-agent systems (MAS) have a long history in distributed problem solving, coordination, and planning, with established mechanisms for task allocation and cooperation such as the Contract Net Protocol and cooperative planning pipelines.[14, 15] Classical accounts describe autonomous, locally rational agents that coordinate through negotiation, commitments, or shared plans to achieve global objectives in heterogeneous environ-

ments.[16, 11] This tradition forms the basis of many contemporary frameworks, which adopt the decomposition of complex objectives into role-specific competencies, local decision making, and explicit protocols for inter-agent communication.[16, 15]

Recent advances in large language models (LLMs) have prompted renewed interest in agentic architectures that use LLMs as planning, communication, and tool-use primitives. These support conversational teams and orchestration layers, as shown by frameworks such as AutoGen, which enable conversation-programmed teams of specialised agents, and by role-playing and workflow-based approaches (e.g., CAMEL, MetaGPT), which demonstrate how division of labour and inter-agent critique can improve performance on open-ended tasks.[17, 18, 19] Methodologically related patterns like ReAct integrate chain-of-thought reasoning with action selection, reinforcing tool-using agent designs.[20]

Beyond research prototypes, developer toolchains provide agent modules for domain-specific pipelines. For example, LangChain-based agents have been applied in data- and tool-integrating systems such as bilingual real-estate advisory assistants, illustrating the accessibility, but also the variability, of engineering practices in current agent stacks.[21]

Most of these systems are assessed primarily on functional metrics such as task success, sample efficiency, or benchmark scores, with embedded compliance properties rarely enforced at the orchestration layer. This gap is reflected in the creation of dedicated safety and trustworthiness benchmarks for LLM agents, which document frequent policy violations and limited risk-awareness even when task outcomes meet performance targets.[22, 23] Surveys of LLM-based MAS and of LLM security and privacy similarly characterise the field as performance-driven, with governance and assurance positioned as peripheral rather than architectural concerns.[24, 25] These findings motivate the design of architectures in which compliance, procedural validity, and auditability are encoded as binding constraints on inter-agent coordination, rather than implemented through external monitoring after deployment.

2.2 Compliance-by-Design in AI

Compliance-by-design, also referred to as regulation-by-design, is embedded in European data protection law through the principle of *data protection by design and by default* (DPbDD), which requires ex ante technical and organisational measures by controllers (Article 25 GDPR) and is elaborated in the European Data Protection Board's Guidelines 4/2019.[26, 27] Contemporary AI governance frameworks extend this preventive approach: the EU Artificial Intelligence Act adopts a risk-based regime in which classification and obligations depend on risk level (notably Article 6 and Annex III for high-risk applications), while the NIST AI Risk Management Framework (AI RMF 1.0) structures governance, mapping, measurement, and management functions across the AI lifecycle.[9, 8]

For high-risk AI systems, the AI Act sets both design-time and run-time requirements that operationalise compliance-by-design. These include a documented, lifecycle *risk management system* (Article 9); *data and data governance* controls for training, validation, and testing datasets (Article 10); *technical documentation* demonstrating conformity (Article 11); *logging* capabilities for traceability and supervision (Article 12); *transparency* obligations toward deployers (Article 13); *human oversight* mechanisms (Article 14); and thresholds for *accuracy, robustness, and cybersecurity* (Article 15).[9] Conformity is reinforced through *conformity assessment* procedures (Article 43), post-market *monitoring and incident reporting* (Articles 72–73), and registration in an EU database for traceability (Articles 49, 71). Together, these measures embed auditability into orchestration rather than treating it as an external control.[9] Complementary standards such as ISO/IEC 42001:2023 (AI management systems) and ISO/IEC 23894:2023 (AI risk management guidance) codify organisational processes that support such embedding at scale.[28, 29]

In practice many implementations remain model-centric: compliance mechanisms are attached to single-model pipelines via pre-deployment documentation and post hoc audits, rather than enforced as binding constraints across system orchestration.[30, 31] Reporting artefacts such as *model cards* and *datasheets for datasets* improve transparency but do not, on their own, guarantee procedural validity under operational conditions or cross-component accountability in complex deployments.[32, 33] Distributed, multi-agent ecosystems amplify

these gaps because compliance must propagate consistently across inter-agent communication, task delegation, and decision rights. Current surveys and risk frameworks describe assurance in such contexts as an unresolved socio-technical challenge, reinforcing the need to integrate governance and oversight into orchestration logic rather than relying solely on downstream monitoring.[25, 8]

2.3 Distributed Ledger Technology for Verification

Distributed ledger technology (DLT) is widely used to secure data integrity, provenance, and auditability by maintaining append-only, tamper-evident records in adversarial settings.[12, 34] These guarantees are realised through hash-chaining and authenticated data structures, such as Merkle trees, which enable efficient inclusion and append-only proofs, as demonstrated in transparency-log systems for public auditing.[35, 36, 37] Enterprise-grade, permissioned platforms such as Hyperledger Fabric extend these primitives with identity management, endorsement policies, and configurable consensus, providing fine-grained access control and verifiable transaction histories.[38]

Conventional blockchains impose a linear order of blocks, simplifying verification but constraining throughput and latency.[12, 34] Directed acyclic graph (DAG) ledgers generalise this structure to blockDAGs or transaction DAGs, enabling higher concurrency and, in practice, improved throughput–latency trade-offs.[39, 40, 41] Peer-reviewed analyses of DAG-based designs—including IOTA's Tangle, Hashgraph, and GHOSTDAG/PHANTOM—formalise consensus and irreversibility properties while identifying performance regimes relevant to near-real-time verification.[39, 40, 41]

In AI contexts, DLT has been applied to safeguard dataset lineage, access logs, and model or update provenance, as well as to provide auditable coordination in federated learning (FL) settings.[42, 43, 44] These applications create tamper-evident trails for coarse-grained artefacts, such as data snapshots, gradient or model updates, and access events. They do not, however, address the fine-grained, low-latency verification of inter-agent messages and decisions required for distributed multi-agent orchestration. Recent work on verifiable ledger databases, such as GlassDB, shows how transparency-log abstractions can deliver efficient inclusion and append-only proofs within transactional workloads, supporting low-latency, tamper-evident verification layers beneath higher-level applications.[45]

The literature positions DLT as a robust substrate for evidentiary integrity and for public or consortium-scale auditability, with DAG-based ledgers particularly suited to high event-rate verification. What remains underexplored is the explicit instrumentation of agent-to-agent coordination and delegation decisions with these verifiability guarantees at the orchestration layer of multi-agent AI systems, extending beyond static data and FL update logging.[43, 42]

2.4 ESG Integration in Digital Architectures

ESG requirements in digital infrastructures have been institutionalised primarily through disclosure regimes and reporting standards rather than through runtime control of systems. In the EU, the Corporate Sustainability Reporting Directive (CSRD) mandates comprehensive sustainability disclosures, operationalised via the European Sustainability Reporting Standards (ESRS), which specify topic-level reporting (e.g., ESRS E1 on climate) and alignment with established greenhouse gas (GHG) accounting practices.[46, 47, 48] In parallel, AI governance proposals emphasise sustainability as a design value, yet the technical literature consistently reports that energy and carbon metrics are underreported or appended post hoc, rather than embedded as binding operational constraints.[49, 50, 51]

For AI workloads, environmental impacts vary by model class, training versus inference phase, facility efficiency, hardware choice, geography, and time-varying grid carbon intensity.[52, 53] Empirical studies on frontier-scale models, such as BLOOM 176B, show that adopting a full life-cycle perspective materially alters footprint estimates compared to training-only views, underscoring the importance of inventory boundaries for any architecture-level constraint.[54] These findings strengthen calls to treat emissions and efficiency as first-class performance criteria rather than externalities assessed after deployment.[49]

Standardised metrics already translate facility- and product-level sustainability attributes into machine-readable constraints. For data centres, the ISO/IEC 30134 series defines key performance indicators such as Power Usage Effectiveness (PUE), Energy Reuse Factor (ERF), and Carbon Usage Effectiveness (CUE), enabling comparison and thresholding of energy efficiency, energy reuse, and operational CO_2 intensity during use-phase operations.[55, 56, 57] At product and organisational levels, ISO 14067 specifies principles for quantifying a product carbon footprint, while the Greenhouse Gas Protocol's Corporate Standard governs Scope 1–3 accounting, providing a bridge between operational telemetry and corporate reporting baselines.[58, 59] ESRS E1 requires disclosure of Scope 1 and Scope 2 emissions and material Scope 3 categories, with guidance to draw on the GHG Protocol (including Scope 2 Guidance) and to include upstream services such as cloud computing and data centre usage under Scope 3, linking digital procurement to reportable inventories.[48]

In most AI systems, environmental metrics are recorded for transparency or for post-deployment optimisation (e.g., shifting regions or times to lower marginal grid intensity) rather than enforced as ex ante constraints at the orchestration layer.[53, 51] A compliance-first multi-agent architecture can reverse this pattern by: (i) parameterising tasks with explicit environmental budgets, such as caps on energy use or CUE-derived kg CO_2e/kWh; (ii) binding scheduler decisions to standards-aligned telemetry, including PUE, ERF, CUE, and ISO 14067-compliant product footprints; and (iii) recording allocations and outcomes for auditability and ESRS-aligned reporting. In such a model, ESG becomes a set of machine-enforceable constraints shaping planning, placement, and execution pathways in real time.[49, 57, 48]

2.5 Identified Research Gap

Across the surveyed strands of multi-agent orchestration, compliance-by-design, ledger-based verification, and ESG governance, the literature shows substantial progress in isolation. What is missing is an integrated reference architecture that (i) encodes regulatory and assurance obligations as binding orchestration constraints for multi-agent coordination, (ii) instruments agent-to-agent interactions with tamper-evident, low-latency verification, and (iii) operationalises ESG targets as run-time budget constraints rather than retrospective disclosures.

First, contemporary LLM-based multi-agent systems and orchestration frameworks emphasise task performance, division of labour, and communication protocols, while governance and assurance remain secondary concerns or external monitoring functions.[24, 25] Surveys of LLM security and privacy report persistent gaps in policy adherence, traceability, and cross-component accountability in agentic settings, indicating the absence of orchestration-layer guarantees.[25]

Second, compliance frameworks codify obligations such as lifecycle risk management, logging, technical documentation, and human oversight, yet they do not specify how such requirements can be expressed as machine-enforceable constraints governing inter-agent delegation, voting, and escalation within MAS.[9, 8] This leaves a design space between policy intent and system realisation, where orchestration semantics rarely bind compliance artefacts, including risk registers, logs, and conformity documentation, to the decision rights and protocols of interacting agents.[8]

Third, distributed ledgers provide tamper-evident auditability, DAG-based designs improve throughput and latency characteristics, and verifiable ledgers offer efficient inclusion proofs. Existing AI integrations, however, focus on dataset or model provenance and federated learning update trails rather than on fine-grained verification of agent-to-agent messages and decisions in real time.[39, 40, 41, 45, 43] As a result, evidentiary integrity is typically attached to coarse events, such as data access or model versioning, rather than to the internal coordination steps that determine enterprise actions.[43, 45]

Finally, ESG standards and reporting regimes define metrics and disclosures, such as CUE for data centres and ESRS E1 for Scope 1–3, yet they are rarely embedded into schedulers, planners, or resource allocators as enforceable budgets or thresholds that shape agent behaviour at run time.[57, 48] The prevailing pattern is post-deployment measurement and reporting, not ex ante constraint satisfaction within the operational fabric of AI systems.[48]

The gap is the absence of an enterprise-grade, domain- and vendor-agnostic orchestration layer for multi-agent AI that encodes compliance obligations and ESG constraints as structural properties, and couples these with

DAG-backed, tamper-evident verification of inter-agent coordination to achieve admissible, auditable operations at scale.

3 Objective and Contribution

The objective of this work is to define and operationalise a compliance-first architecture for enterprise multi-agent AI in which regulatory, governance, and ESG obligations function as binding orchestration primitives. The architecture is domain-, deployment-, and vendor-agnostic, positioning generative models as interchangeable components within a governed multi-agent fabric. A role-specific agent layer supports both persistent and ad hoc committees, while a verification-and-governance layer enforces policies, captures evidence, and integrates a directed acyclic graph (DAG) verification substrate for tamper-evident, low-latency auditability. ESG requirements are embedded as run-time budgets, with scheduling and placement decisions bound to standards-aligned telemetry and recorded for auditability. The design preserves operational neutrality across cloud, on-premise, and edge deployments, and incorporates assurance mechanisms for human oversight, incident response, and procedural validity.

The contributions of this work are, first, a reference architecture and orchestration semantics that bind decision rights, escalation paths, and separation-of-duties to machine-enforceable constraints. Second, a DAG-based verification layer that records inter-agent coordination at message level with inclusion proofs and evidentiary retention policies. Third, an ESG-aware scheduler that treats environmental targets as binding optimisation constraints, exposing trade-offs between utility, latency, and resource impact. Fourth, a deployment blueprint and threat model addressing trust zones, policy evasion, and evidence manipulation. Finally, the paper outlines cross-domain applicability through healthcare triage, public procurement, and incident management scenarios, and identifies open research directions in formal policy verification, cross-organisation committee incentives, and performance–auditability trade-offs at scale.

4 Research

Organisations can be broadly classified into two categories: profit-oriented and mission-oriented. Profit-oriented entities focus on maximising financial performance, while mission-oriented entities, such as governmental bodies, healthcare providers, educational institutions, and non-governmental organisations, prioritise service delivery, public welfare, or other non-financial mandates. Despite these differences, both operate under bounded resources and are subject to regulatory, governance, and operational constraints. This organisational taxonomy provides a useful lens for framing optimisation problems in enterprise AI design.

In its most general form, the operational objective of an organisation can be expressed as the joint maximisation of utility and viability, where viability is understood as the sustained capacity of the organisation to operate, adapt, and meet its objectives under changing internal and external conditions, as shown in Equation (2):

$$\begin{aligned} \max_{\pi \in \Pi} \quad & [\alpha\, U_{\text{org}}(\pi) + \beta\, V_{\text{org}}(\pi)] && (2) \\ \text{s.t.} \quad & \mathcal{C}_{\text{reg}}(\pi) = 1 && \text{(regulatory compliance)} \\ & \mathcal{C}_{\text{gov}}(\pi) = 1 && \text{(governance compliance)} \\ & \mathcal{C}_{\text{res}}(\pi) \le R_{\max} && \text{(resource budget)} \end{aligned}$$

Here, $\pi \in \Pi$ is an operational policy, U_{org} and V_{org} denote organisational utility and viability, and $\alpha, \beta \ge 0$ are weighting parameters. A binary constraint $\mathcal{C}_x(\pi) = 1$ indicates that the requirement must be fully satisfied.

When applied to profit-oriented organisations, this formulation specialises to Equation (3):

$$\begin{aligned} \max_{\pi \in \Pi} \quad & [\alpha\, U_{\text{profit}}(\pi) + \beta\, V_{\text{profit}}(\pi)] \\ \text{s.t.} \quad & \mathcal{C}_{\text{reg}}(\pi) = 1 \\ & \mathcal{C}_{\text{gov}}(\pi) = 1 \\ & \mathcal{C}_{\text{res}}(\pi) \leq R_{\max} \end{aligned} \tag{3}$$

Here, U_{profit} may include profit, revenue growth, return on equity, or market share, while V_{profit} reflects the ability to sustain financial and operational performance over time.

For mission-oriented organisations, the same structure yields Equation (4):

$$\begin{aligned} \max_{\pi \in \Pi} \quad & [\alpha\, U_{\text{mission}}(\pi) + \beta\, V_{\text{mission}}(\pi)] \\ \text{s.t.} \quad & \mathcal{C}_{\text{reg}}(\pi) = 1 \\ & \mathcal{C}_{\text{gov}}(\pi) = 1 \\ & \mathcal{C}_{\text{res}}(\pi) \leq R_{\max} \end{aligned} \tag{4}$$

Here, $\pi \in \Pi$ denotes an operational policy, U_{mission} and V_{mission} capture mission-related utility and viability, and $\alpha, \beta \geq 0$ are weighting parameters reflecting their relative importance. U_{mission} may include service coverage, quality, accessibility, or stakeholder satisfaction, while V_{mission} measures the organisation's capacity to maintain or improve service delivery under changing conditions. As before, a binary constraint $\mathcal{C}_x(\pi) = 1$ indicates that the requirement must be fully satisfied, while inequality constraints bound performance metrics within acceptable thresholds.

This distinction matters because the underlying objectives shape both the optimisation target and the permissible coordination and competition patterns among agents. Profit-oriented entities often operate in competitive markets where strategic advantage depends on outperforming or displacing rivals, aligning with adversarial or non-cooperative game-theoretic models. Mission-oriented organisations, by contrast, frequently operate in settings where cooperation and information sharing improve collective outcomes, such as public health, safety, or environmental protection. In these domains, cooperative game theory and equilibrium concepts such as the Nash equilibrium can support joint optimisation rather than zero-sum competition. In profit-driven AI systems, the practical relevance of such equilibria is diminished, as design incentives typically favour unilateral maximisation over collaborative stability. Within this taxonomy, enterprise AI orchestration in regulated multi-agent environments can be seen as a domain-specific instance of the general organisational formulations in Equations (2)–(4). Regardless of whether the organisation is profit- or mission-oriented, the orchestration problem differs fundamentally from the consumer-facing formulation in Equation (1). Rather than optimising the quality of a single model output, the aim is to maximise organisational utility while meeting a set of normative, procedural, evidentiary, ESG, and safety constraints that bind all agents in the system. This enterprise-specific objective is formalised in Equation (5):

$$\begin{aligned} \max_{\pi \in \Pi} \quad & U_{\text{org}}(\pi, G, \Theta) && \\ \text{s.t.} \quad & \mathcal{C}_{\text{reg}}(\pi) = 1 && \text{(regulatory compliance)} \\ & \mathcal{C}_{\text{proc}}(\pi) = 1 && \text{(procedural validity)} \\ & \mathcal{C}_{\text{esg}}(\pi) \geq \tau_{\text{esg}} && \text{(ESG performance thresholds)} \\ & \mathcal{C}_{\text{evid}}(\pi) = 1 && \text{(evidentiary completeness)} \\ & \mathcal{C}_{\text{safety}}(\pi) \geq \tau_{\sigma} && \text{(safety and risk minimisation)} \\ & L(\pi) \leq \tau_{l} && \text{(latency bound for execution)} \\ & \mathcal{C}_{\text{other}}(\pi) = 1 && \text{(additional enterprise policies and operational constraints)} \end{aligned} \tag{5}$$

Here, $\pi \in \Pi$ denotes an orchestration policy mapping goals G to coordinated agent actions, with Θ representing system-wide parameters and configuration. The function U_{org} captures organisational utility, including task success, resource efficiency, and mission alignment. Each constraint $\mathcal{C}_x(\pi)$ returns 1 if and only if it is satisfied, meaning that equations of the form $\mathcal{C}_x(\pi) = 1$ represent requirements that must be met for the policy to be admissible. $\mathcal{C}_{\text{reg}}$ enforces legal and regulatory compliance; $\mathcal{C}_{\text{proc}}$ ensures procedural validity, including separation-of-duties and escalation rules; $\mathcal{C}_{\text{esg}}$ measures ESG performance against a minimum threshold τ_{esg}; $\mathcal{C}_{\text{evid}}$ guarantees evidentiary completeness and auditability; $\mathcal{C}_{\text{safety}}$ enforces operational safety and risk minimisation with threshold τ_σ; $L(\pi)$ is the end-to-end execution latency, bounded by τ_l; and $\mathcal{C}_{\text{other}}$ covers any additional enterprise policies, strategic objectives, or operational constraints. The organisational taxonomy outlined above defines the admissible objectives and constraints and also influences the strategic interactions embedded in the orchestration layer. Real-world organisations rarely operate in a state of pure competition or pure cooperation. Mission-oriented hospitals may compete for funding, talent, or innovation leadership, while profit-oriented enterprises may cooperate in areas such as industry safety standards, shared infrastructure, or joint lobbying.

To capture this variability, we introduce a tunable *cooperation–competition continuum* as a first-class orchestration parameter, formalised in Equation (6). Let $\gamma \in [0, 1]$ denote the degree of cooperative strategic alignment across the multi-agent system:

$$U^{\gamma}_{\text{org}}(\pi) = \gamma \cdot U_{\text{coop}}(\pi) + (1 - \gamma) \cdot U_{\text{comp}}(\pi) \tag{6}$$

Here, $U_{\text{coop}}(\pi)$ measures the utility from cooperative behaviours such as resource sharing, mutual aid, and joint optimisation; $U_{\text{comp}}(\pi)$ measures the utility from competitive behaviours such as market share capture, strategic positioning, or adversarial advantage. The cooperation parameter γ may be set as part of organisational policy or adjusted dynamically in response to environmental conditions, market signals, or strategic triggers.

This formulation supports orchestration policies ranging from purely competitive ($\gamma = 0$) to purely cooperative ($\gamma = 1$), with intermediate values representing blended strategies. For example, a public administration without a competitive mandate may operate with γ close to 1, while a mission-oriented hospital could set $\gamma \approx 0.7$ to reflect cooperative public health coordination alongside selective competition for scarce resources. A profit-oriented enterprise might operate near $\gamma = 0$ but temporarily increase it in contexts where cooperation improves expected payoffs. Embedding γ at the orchestration level makes the cooperative–competitive balance an explicit, governed parameter rather than an emergent property of ad hoc agent design. It also enables scenario planning in which the system can simulate or execute strategic shifts. For instance, an organisation could operate with $\gamma = 0.8$ in a stable supply chain, but reduce it to $\gamma = 0.5$ when market volatility calls for stronger competitive positioning. This capability provides strategic agility and preserves an auditable record of the decision logic behind changes in inter-agent interaction patterns.

Building on the cooperation–competition continuum, we extend the framework to allow γ to vary with context rather than remain fixed across all scenarios. Organisations may define a baseline profile, for example $\gamma = 0.3$ to reflect a 30% cooperative and 70% competitive stance, while retaining the ability to override this default when circumstances change. In a stable market, the baseline may be sufficient. During a critical competitive challenge the system could shift to $\gamma = 0.0$ for a fully competitive posture. In a joint-response emergency the system could move to $\gamma = 1.0$ to maximise coordination and shared resource use.

The adaptive parameter is expressed as a context-driven function in Equation (7), which extends the formulation in Equation (6):

$$\gamma_t = f(\mathcal{S}_t, \mathcal{P}, \Theta) \tag{7}$$

Here, γ_t is the cooperation parameter at time t, $\mathcal{S}_t$ describes the prevailing situational context, $\mathcal{P}$ is the set of organisational policies, and Θ represents system-level parameters and thresholds. The mapping $f(\cdot)$ determines the operative γ_t from observed conditions and policy rules, updating its value as the context evolves.

This adaptive formulation keeps multi-agent reasoning and decision outputs aligned with situational demands, echoing patterns seen in social and biological systems. In nature, cooperation is often extended to in-group members, while competition or defensive actions are directed toward external threats. Protecting

community members is a cooperative imperative, whereas repelling an intruder may require aggressive and competitive tactics. In organisational terms, applying a collaborative strategy to a hostile market entrant risks undermining viability, while treating a trusted supply chain partner as an adversary can erode long-term utility. The proposed approach enables this flexibility without presupposing a single correct posture or enforcing a fixed set of cooperative or competitive behaviours. It leaves the choice to the orchestration logic and the governing policies, which can interpret the operational context in real time and select the most suitable balance. This preserves adaptability as a design property, allowing the system to respond to complex or ambiguous situations without being constrained by rigid strategic assumptions.

4.1 Orchestration Architecture with Context-Adaptive Strategic Parameters

We extend the enterprise orchestration formulation in Equation (5) by embedding the cooperation–competition continuum from Equation (6) together with its context-driven adaptation from Equation (7). The resulting objective is given in Equation (8), where the cooperation weight γ_t varies over time in response to observed conditions:

$$\begin{aligned} \max_{\pi \in \Pi} \quad & U_{\text{org}}^{\gamma_t}(\pi, G, \Theta) \\ \text{s.t.} \quad & \mathcal{C}_{\text{reg}}(\pi) = 1,\ \mathcal{C}_{\text{proc}}(\pi) = 1,\ \mathcal{C}_{\text{esg}}(\pi) \geq \tau_{\text{esg}},\ \mathcal{C}_{\text{evid}}(\pi) = 1, \\ & \mathcal{C}_{\text{safety}}(\pi) \geq \tau_\sigma,\ L(\pi) \leq \tau_l,\ \mathcal{C}_{\text{other}}(\pi) = 1 \\ \text{where} \quad & U_{\text{org}}^{\gamma_t}(\pi, G, \Theta) = \gamma_t \cdot U_{\text{coop}}(\pi, G, \Theta) + (1 - \gamma_t) \cdot U_{\text{comp}}(\pi, G, \Theta) \end{aligned} \tag{8}$$

Here, $\gamma_t \in [0, 1]$ is obtained from the prevailing situational context $\mathcal{S}_t$ through the mapping $f(\mathcal{S}_t, \mathcal{P}, \Theta)$, where $\mathcal{P}$ denotes the organisation's governing policies and Θ contains system-level parameters and thresholds. Values of γ_t close to one direct orchestration towards cooperative modes, with emphasis on resource pooling, mutual aid, and joint optimisation. Values near zero shift the balance toward competitive modes, prioritising market capture, strategic positioning, or adversarial advantage. Intermediate values produce blended regimes in which cooperation and competition are weighted according to the current context.

The mapping $f(\cdot)$ supports strategic shifts in response to market signals, environmental changes, or operational events. A dedicated *Context Monitor* ingests telemetry, market indicators, ESG metrics, and incident data to update γ_t. Each update is checked by the *Policy Store* and *Rule Engine*, transformed into enforceable constraints by the *Constraint Compiler*, and passed through the *Compliance Verification* stage. Pre-execution outputs may be reviewed with human oversight before passing the *Execution Gate*.

Chain-of-Thought scopes, variable states, and decision records are notarised in the *DAG Verification Layer* alongside event logs and inclusion proofs. This ensures that policy compliance, reproducibility, and auditability are preserved. The *Audit Layer* offers retrieval and simulation capabilities through a dedicated query API. The layered orchestration architecture in Figure 1 integrates these elements.

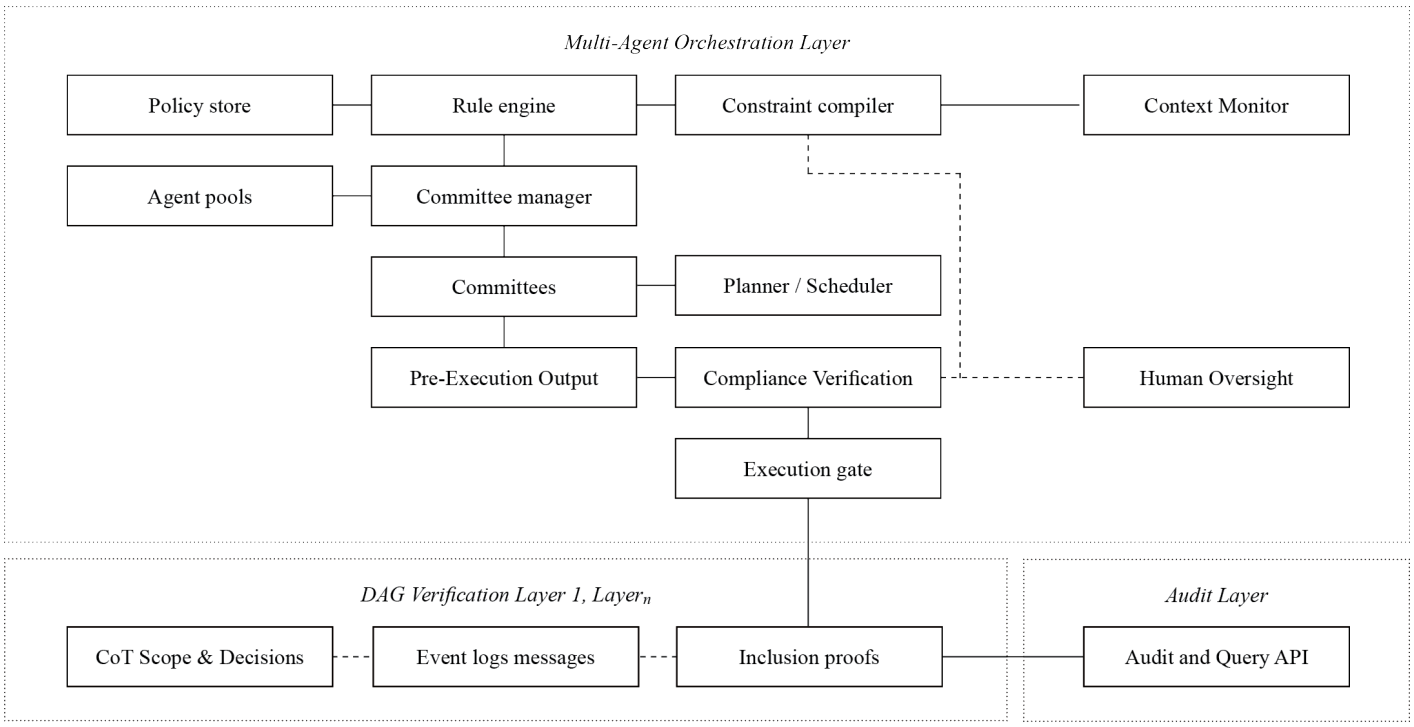

Figure 1: Layered orchestration architecture integrating the context-adaptive cooperation–competition parameter γ_t. The Policy and Governance components compute γ_t via $f(\mathcal{S}_t, \mathcal{P}, \Theta)$ and compile constraints $\{\mathcal{C}_x\}$. The Orchestration Layer plans and verifies outputs, while the DAG Verification Layer notarises Chain-of-Thought scopes, decisions, and constraint outcomes. The Audit Layer enables forensic and simulation queries.

While existing multi-agent orchestration frameworks tend to address coordination, optimisation, or compliance as separate concerns, few combine them in a single architecture that treats the cooperation–competition balance as a governed, context-adaptive parameter. The design in Figure 1 embeds γ_t directly into the decision loop, applies compliance checks before execution, and records all strategic changes with verifiable provenance. This integration closes two persistent gaps in the literature. The first is the absence of a mechanism for adjusting inter-agent posture along a continuum from fully cooperative to fully competitive in response to contextual triggers. The second is the lack of a transparent link between those triggers, the orchestration decisions they influence, and their eventual operational outcomes. Addressing both allows enterprise-scale multi-agent systems to remain strategically agile while demonstrably aligned with long-term viability objectives.

A Directed Acyclic Graph (DAG) is adopted as the notarisation substrate to avoid the latency bottlenecks inherent in block-generation intervals of conventional blockchains. For audit logs that are technical rather than financial, the confirmation delays and transaction fees typical of Layer 1 financial ledgers are poorly aligned with operational requirements. In the proposed model, events are minted into a Layer 2 DAG with negligible or zero transaction costs, allowing high-frequency, low-latency notarisation of decisions, constraint checks, and contextual triggers. Where minting costs are non-zero, a game-theoretic incentive design can be applied to favour comprehensive notarisation while retaining the ability to omit non-critical events. Let $\mathcal{E}$ be the set of events, $c(e)$ the minting cost of event $e \in \mathcal{E}$, and $v(e)$ its expected verification or audit value. The resulting notarisation preference function is given in Equation (9):

$$\max_{\mathcal{E}' \subseteq \mathcal{E}} \quad \sum_{e \in \mathcal{E}'} \big[v(e) - \lambda \cdot c(e) \big] \tag{9}$$

Here, λ captures the operator's sensitivity to cost. By calibrating λ, the system can favour high-value, low-cost events while discouraging the omission of critical actions. Under realistic cost conditions, comprehensive notarisation can be made the dominant operational choice. The Layer 2 DAG can periodically anchor its

state via Merkle tree commitments to a Layer 1 ledger, which may be either a general-purpose blockchain or a specialised ledger dedicated to AI system activity logs. In both cases, the economic model of the underlying ledger influences the completeness of notarisation. Within this framework, the DAG functions as a verifiable record of system activity and a primary source for forensic analysis, simulation replay, and regulatory audits.

In a dedicated DAG system, the vectorisation of smart contracts and stored hashes is proposed to enable AI agents to parse and reason over notarised content with greater efficiency and security. Representing contract logic and state proofs as vectorised structures allows faster retrieval and interpretation while ensuring compatibility with embedding-based reasoning systems. This can reduce parsing overhead and lower the risk of errors arising from ambiguous or inconsistent contract encoding. The feasibility of this approach, along with its performance trade-offs and security implications, must be validated through rigorous testing in realistic deployment environments. We regard this as a priority direction for future research, with the potential to strengthen the integration between notarisation layers and AI orchestration systems.

4.2 Human Oversight as a Governance Primitive

Human oversight is embedded in the orchestration framework as a governance primitive, spanning the full operational lifecycle from policy definition to post-execution audit. In the proposed architecture, human actors interact with the system at points chosen for their alignment with governance objectives and risk profiles.

At the *policy and strategy* level, human decision-makers establish baseline rules, operational constraints, and strategic priorities, including the cooperation–competition balance parameter γ. While γ may be dynamically adjusted by the system in response to situational data $\mathcal{S}_t$, human operators retain the authority to refine or override these values when contextual insight exceeds the agents' situational model. In volatile markets or emergent public health crises, for instance, a human may recalibrate the cooperative–competitive weighting on the basis of information not yet incorporated into the system state.

At the *pre-execution* stage, humans participate in compliance validation for high-impact or high-risk actions. This may involve reviewing an *output draft* from the orchestration layer, with compliance agents—human or AI—verifying adherence to regulatory, procedural, and ESG constraints before execution. Human reviewers act as an adaptive safeguard, addressing edge cases, interpretative ambiguities, or ethical considerations that fall outside the scope of formalised constraints.

At the *audit and review* stage, human oversight complements AI-based auditing. Post-execution logs notarised in the DAG layer are accessible for forensic review, simulation replay, and regulatory reporting. High-volume routine verification may be handled algorithmically, whereas exceptional cases, disputed actions, or policy updates can trigger human-led audits. This dual-mode design combines the speed and coverage of automated verification with the interpretability and accountability of human review.

Integrating human oversight at multiple stages allows the system to incorporate human intuition, ethical reasoning, and domain expertise without diminishing the autonomy and scalability of the multi-agent fabric. Oversight frequency and depth can be adapted through scenario policies, increasing human involvement during periods of regulatory uncertainty or heightened reputational risk. By formalising oversight within the orchestration semantics rather than treating it as an ad hoc intervention, the architecture ensures that human judgement remains a governed, auditable, and strategically aligned element of system operation.

5 Discussion and Theoretical Implications

The proposed architecture integrates a multi-agent orchestration framework with a context-adaptive cooperation–competition continuum, notarisation through a Layer 2 DAG, and governed human oversight. While these components have each been examined in prior work, their combination into a unified, policy-driven enterprise AI system represents a shift in both theoretical framing and operational design.

From a theoretical perspective, the cooperation–competition parameter γ_t moves beyond static game-theoretic formulations toward a dynamic, context-aware representation. This formulation recognises that organisational

strategy is not fixed along the competitive–cooperative axis, but evolves in response to market dynamics, environmental triggers, and internal strategic priorities. Embedding γ_t as a tunable orchestration parameter brings the model closer to observed organisational behaviour in both profit-oriented and mission-oriented contexts. The continuum also creates a formal bridge between strategic management theory and multi-agent system design, supporting scenario modelling in which competitive advantage and cooperative stability are jointly optimised.

The use of a Layer 2 DAG as the notarisation substrate introduces additional theoretical considerations. It reframes distributed ledger technology in AI governance from a financial transaction ledger to a high-frequency, low-latency record of decision processes and constraint evaluations. This shift positions the ledger as a strategic memory that enables replay, forensic analysis, and policy refinement without the performance costs of Layer 1 blockchains. The game-theoretic cost–value function for notarisation in Equation (9) formalises the trade-off between completeness and operational expenditure, allowing preferred behaviours to be embedded directly into the economic logic of the logging mechanism.

Embedding human oversight across multiple orchestration stages also carries theoretical significance. By defining human interaction points at policy formulation, pre-execution validation, and post-execution audit, the architecture treats oversight as a structural governance mechanism rather than a reactive safeguard. This ensures that human intervention is both traceable and integrated with operational constraints, supporting a hybrid governance model in which ethical reasoning and domain expertise complement algorithmic scale and consistency.

Taken together, the architecture operationalises concepts that have often been addressed separately in AI research: strategic adaptability, distributed trust infrastructure, and governed human–AI collaboration. The theoretical contribution lies in positioning these elements as interdependent components of an enterprise-grade, multi-agent AI system. This perspective offers new avenues for research on adaptive game-theoretic strategies, ledger-augmented governance mechanisms, and empirical studies of human oversight in complex socio-technical environments.

6 Limitations and Future Research

The proposed architecture offers a coherent framework for enterprise-grade multi-agent AI orchestration, yet several limitations remain that merit further investigation.

The cooperation–competition continuum parameter γ_t is theoretically well-founded but has not been empirically validated across heterogeneous organisational settings. Robust calibration of γ_t will require longitudinal studies and controlled simulation experiments to test its stability, responsiveness, and resilience to strategic manipulation. The behavioural effects of abrupt γ_t adjustments in high-stakes domains such as emergency healthcare or volatile financial markets remain largely unexplored.

The Layer 2 DAG notarisation mechanism addresses latency and transaction cost constraints, but introduces its own operational trade-offs. The cost–value formulation in Equation (9) presumes accurate estimation of $v(e)$ and a stable cost-sensitivity parameter λ. In practice, both are likely to vary with shifting priorities, governance changes, and external pressures. Future work should examine adaptive strategies for tuning λ, as well as the implications of partial notarisation on the completeness and reliability of forensic records. The proposal to vectorise smart contracts and hashed proofs for AI-native access also requires targeted benchmarking to assess retrieval speed, parsing accuracy, and security performance under realistic deployment loads.

Human oversight, although structurally embedded at multiple orchestration stages, has not yet been assessed in terms of cognitive demand and process latency. Without careful design, oversight points could become operational bottlenecks when the complexity or frequency of agent actions exceeds feasible review capacity. A promising direction is the development of hybrid audit workflows that combine AI-assisted triage with selective human review, preserving both scalability and accountability.

Interoperability with existing enterprise systems and regulatory environments is another open challenge. Cross-layer anchoring strategies, jurisdiction-specific compliance rules, and sector-calibrated ESG thresholds will likely require domain-specific tailoring. These factors are critical for adoption in regulated industries and for ensuring cross-border operational validity.

References

1. Stanford HAI. AI Index Report 2025: Chapter 4 Economy. Stanford Institute for Human-Centered AI. 2025. Available from: https://hai.stanford.edu/assets/files/hai_ai-index-report-2025_chapter4_final.pdf
2. OpenAI. Models and API overview. Product documentation; accessed 2025-08-11. 2025. Available from: https://platform.openai.com/docs/models
3. Schulhoff S et al. A systematic survey of prompt engineering techniques. arXiv preprint. 2024. DOI: 10.48550/arXiv.2406.06608. Available from: https://arxiv.org/abs/2406.06608
4. Longpre S et al. Designing data and methods for effective instruction tuning. arXiv preprint. 2023. DOI: 10.48550/arXiv.2301.13688. Available from: https://arxiv.org/abs/2301.13688
5. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CL, Mishkin P, Zhang C, Agarwal S, Slama K, Ray A, Schulman J, Hilton J, Kelton F, Miller L, Simens M, Askell A, Welinder P, Christiano P, Leike J, and Lowe R. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems 35*. NeurIPS 2022; arXiv:2203.02155. 2022. DOI: 10.48550/arXiv.2203.02155. Available from: https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf
6. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih W, Rocktäschel T, Riedel S, and Kiela D. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems 33*. NeurIPS 2020; arXiv:2005.11401. 2020. DOI: 10.48550/arXiv.2005.11401. Available from: https://proceedings.neurips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf
7. OpenAI. Moderation guide and models. Product documentation; accessed 2025-08-11. 2024. Available from: https://platform.openai.com/docs/guides/moderation
8. National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0). Tech. rep. NIST, 2023. DOI: 10.6028/NIST.AI.100-1. Available from: https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf
9. European Parliament and Council of the European Union. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal L, 12 July 2024. 2024. Available from: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng
10. Chopra AK, Torre L van der, Verhagen H, and Villata S, eds. Handbook of normative multiagent systems. Open-access PDF. College Publications, 2018. Available from: https://www.collegepublications.co.uk/downloads/handbooks00004.pdf
11. Wooldridge M. An introduction to multiagent systems. 2nd ed. Missing DOI. Chichester, UK: John Wiley & Sons, 2009. Available from: https://www.wiley.com/en-us/An%2BIntroduction%2Bto%2BMultiAgent%2BSystems%2C%2B2nd%2BEdition-p-9780470519462
12. Yaga D, Mell P, Roby N, and Scarfone K. Blockchain technology overview. 2018. DOI: 10.6028/NIST.IR.8202. Available from: https://doi.org/10.6028/NIST.IR.8202
13. Baird L. The Swirlds hashgraph consensus algorithm: Fair, fast, Byzantine fault tolerance. Tech. rep. Technical report SWIRLDS-TR-2016-01. Swirlds Inc., 2016. Available from: https://www.swirlds.com/downloads/SWIRLDS-TR-2016-01.pdf
14. Smith RG. The contract net protocol: High-level communication and control in a distributed problem solver. IEEE Transactions on Computers 1980; C-29:1104–13. DOI: 10.1109/TC.1980.1675516. Available from: https://www.eecs.ucf.edu/~lboloni/Teaching/EEL6788_2008/papers/The_Contract_Net_Protocol_Dec-1980.pdf

15. Torreño A, Onaindia E, Komenda A, and Štolba M. Cooperative multi-agent planning: A survey. ACM Computing Surveys 2017; 50:84:1–84:32. DOI: 10.1145/3128584. Available from: https://dl.acm.org/doi/10.1145/3128584

16. Jennings NR, Sycara K, and Wooldridge M. A roadmap of agent research and development. Autonomous Agents and Multi-Agent Systems 1998; 1:7–38. DOI: 10.1023/A:1010090405266. Available from: https://link.springer.com/article/10.1023/A:1010090405266

17. Wu Q, Bansal G, Zhang J, Wu Y, Li B, Zhu E, Jiang L, Zhang X, Zhang S, Liu J, Awadallah AH, White RW, Burger D, and Wang C. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. arXiv 2023. Preprint. DOI: 10.48550/arXiv.2308.08155. eprint: 2308.08155. Available from: https://arxiv.org/abs/2308.08155

18. Li G, Hammoud HAAK, Itani H, Khizbullin D, and Ghanem B. CAMEL: Communicative agents for "mind" exploration of large language model society. arXiv 2023. Preprint. DOI: 10.48550/arXiv.2303.17760. eprint: 2303.17760. Available from: https://arxiv.org/abs/2303.17760

19. Hong S et al. MetaGPT: Meta programming for a multi-agent collaborative framework. arXiv 2023. Preprint. DOI: 10.48550/arXiv.2308.00352. eprint: 2308.00352. Available from: https://arxiv.org/abs/2308.00352

20. Yao S, Zhao J, Yu D, Du N, Shafran I, Narasimhan K, and Cao Y. ReAct: Synergizing reasoning and acting in language models. *Proceedings of the Eleventh International Conference on Learning Representations (ICLR)*. Missing DOI. 2023. Available from: https://openreview.net/forum?id=WE_vluYUL-X

21. Haurum KR, Ma R, Wen L, and Wen L. Real estate with AI: An agent based on LangChain. Procedia Computer Science 2024; 242:1082–8. DOI: 10.1016/j.procs.2024.08.199. Available from: https://www.sciencedirect.com/science/article/pii/S1877050924019185

22. Yuan T et al. R-Judge: Benchmarking safety risk awareness for LLM agents. *Findings of EMNLP 2024*. 2024. DOI: 10.18653/v1/2024.findings-emnlp.79. Available from: https://aclanthology.org/2024.findings-emnlp.79/

23. Levy I, Wiesel B, Marreed S, Oved A, Yaeli A, and Shlomov S. ST-WebAgentBench: A benchmark for evaluating safety and trustworthiness in web agents. arXiv 2024. Preprint. DOI: 10.48550/arXiv.2410.06703. eprint: 2410.06703. Available from: https://arxiv.org/abs/2410.06703

24. Li X, Wang S, Zeng S, Wu Y, and Yang Y. A survey on LLM-based multi-agent systems: Workflow, infrastructure, and challenges. Vicinagearth 2024; 1. DOI: 10.1007/s44336-024-00009-2. Available from: https://link.springer.com/article/10.1007/s44336-024-00009-2

25. Yao Y, Duan J, Xu K, Cai Y, Sun Z, and Zhang Y. A survey on large language model (LLM) security and privacy: the good, the bad, and the ugly. High-Confidence Computing 2024; 4:100211. DOI: 10.1016/j.hcc.2024.100211. Available from: https://doi.org/10.1016/j.hcc.2024.100211

26. European Parliament and Council of the European Union. Regulation (EU) 2016/679 (General Data Protection Regulation). Official Journal L 119, 4 May 2016; Article 25 DPbDD. 2016. Available from: https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng

27. European Data Protection Board. Guidelines 4/2019 on Article 25 Data protection by design and by default (Version 2.0). Adopted 20 October 2020; final version. 2020. Available from: https://www.edpb.europa.eu/sites/default/files/files/file1/edpb_guidelines_201904_dataprotection_by_design_and_by_default_v2.0_en.pdf

28. International Organization for Standardization and International Electrotechnical Commission. ISO/IEC 42001:2023 — Information technology — Artificial intelligence — Management system. AI management systems; Missing DOI. 2023. Available from: https://www.iso.org/standard/42001

29. International Organization for Standardization and International Electrotechnical Commission. ISO/IEC 23894:2023 — Artificial intelligence — Risk management. Guidance on AI risk management; Missing DOI. 2023. Available from: https://www.iso.org/standard/77304.html

30. Raji ID, Smart A, White RN, Mitchell M, Gebru T, Hutchinson B, Smith-Loud J, Theron D, and Barnes P. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, 2020 :33–44. DOI: 10.1145/3351095.3372873. Available from: https://dl.acm.org/doi/10.1145/3351095.3372873

31. Costanza-Chock S, Harvey E, Raji ID, Czernuszenko M, and Buolamwini J. Who audits the auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, 2022 :1571–83. DOI: 10.1145/3531146.3533213. Available from: https://dl.acm.org/doi/10.1145/3531146.3533213

32. Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, Spitzer E, Raji ID, and Gebru T. Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 2019 :220–9. DOI: 10.1145/3287560.3287596. Available from: https://dl.acm.org/doi/10.1145/3287560.3287596

33. Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, III HD, and Crawford K. Datasheets for datasets. Communications of the ACM 2021; 64:86–92. DOI: 10.1145/3458723. Available from: https://dl.acm.org/doi/10.1145/3458723

34. Nakamoto S. Bitcoin: a peer-to-peer electronic cash system. White paper; Missing DOI. 2008. Available from: https://bitcoin.org/bitcoin.pdf

35. Laurie B, Messeri E, and Stradling R. Certificate Transparency Version 2.0. RFC 9162. 2021. DOI: 10.17487/RFC9162. Available from: https://doi.org/10.17487/RFC9162

36. Crosby SA and Wallach DS. Efficient data structures for tamper-evident logging. *Proceedings of the 18th USENIX Security Symposium*. Missing DOI. 2009 :317–34. Available from: https://static.usenix.org/event/sec09/tech/full_papers/crosby.pdf

37. Tomescu A, Bhupatiraju V, Papadopoulos D, Papamanthou C, Triandopoulos N, and Devadas S. Transparency logs via append-only authenticated dictionaries. *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2019 :1299–316. DOI: 10.1145/3319535.3345652. Available from: https://dl.acm.org/doi/10.1145/3319535.3345652

38. Androulaki E, Barger A, Bortnikov V, Cachin C, Christidis K, Caro AD, Enyeart D, Ferris C, Laventman G, Manevich Y, Muralidharan S, Murthy C, Nguyen B, Sethi M, Singh G, Smith K, Sorniotti A, Stathakopoulou C, Vukolić M, Cocco SW, and Yellick J. Hyperledger Fabric: a distributed operating system for permissioned blockchains. *Proceedings of EuroSys '18*. ACM, 2018. DOI: 10.1145/3190508.3190538. Available from: https://dl.acm.org/doi/10.1145/3190508.3190538

39. Silvano WF and Marcelino R. IOTA Tangle: a cryptocurrency to communicate Internet-of-Things data. Future Generation Computer Systems 2020; 112:307–19. DOI: 10.1016/j.future.2020.05.047. Available from: https://doi.org/10.1016/j.future.2020.05.047

40. Baird L. The Swirlds hashgraph consensus algorithm: fair, fast, Byzantine fault tolerance. Tech. rep. SWIRLDS-TR-2016-01. Swirlds Inc., 2016. Available from: https://www.swirlds.com/downloads/SWIRLDS-TR-2016-01.pdf

41. Sompolinsky Y, Wyborski S, and Zohar A. PHANTOM GHOSTDAG: a scalable generalization of Nakamoto consensus. *Proceedings of the 3rd ACM Conference on Advances in Financial Technologies (AFT 2021)*. ACM, 2021 :57–70. DOI: 10.1145/3479722.3480990. Available from: https://doi.org/10.1145/3479722.3480990

42. Kuo T, Kuo H, Kuo C, Ohno-Machado L, and Pao HK. Privacy-preserving model learning on a blockchain network-of-networks. Journal of the American Medical Informatics Association 2020; 27:343–54. DOI: 10.1093/jamia/ocz214. Available from: https://doi.org/10.1093/jamia/ocz214

43. Qu Y, Uddin MP, Gan C, and Xiang Y. Blockchain-enabled federated learning: a survey. ACM Computing Surveys 2022; 55. DOI: 10.1145/3524104. Available from: https://dl.acm.org/doi/10.1145/3524104

44. Özdayi MS, Kantarcioglu M, and Malin B. Leveraging blockchain for immutable logging and querying across multiple sites. 2020. eprint: arXiv:2001.08529. Available from: https://arxiv.org/abs/2001.08529

45. Yue C, Dinh TTA, Xie Z, Zhang M, Chen G, Ooi BC, and Xiao X. GlassDB: an efficient verifiable ledger database system through transparency. Proceedings of the VLDB Endowment 2023; 16:1359–71. DOI: 10.14778/3583140.3583152. Available from: https://www.vldb.org/pvldb/vol16/p1359-ooi.pdf

46. European Parliament and Council. Directive (EU) 2022/2464 of 14 December 2022 on Corporate Sustainability Reporting (CSRD). Official Journal of the European Union. CELEX: 32022L2464. 2022. Available from: https://eur-lex.europa.eu/eli/dir/2022/2464/oj

47. European Commission. Commission Delegated Regulation (EU) 2023/2772 of 31 July 2023 supplementing Directive 2013/34/EU as regards European Sustainability Reporting Standards. Official Journal of the European Union. CELEX: 32023R2772. 2023. Available from: https://eur-lex.europa.eu/eli/reg_del/2023/2772/oj

48. EFRAG. ESRS E1 Climate Change (Delegated Act Annex). Tech. rep. Published in OJ via Commission Delegated Regulation (EU) 2023/2772. European Financial Reporting Advisory Group, 2023. Available from: https://www.efrag.org/Assets/Download?assetUrl=%2Fsites%2Fwebpublishing%2FSiteAssets%2FESRS%2520E1%2520Delegated-act-2023-5303-annex-1_en.pdf

49. Schwartz R, Dodge J, Smith NA, and Etzioni O. Green AI. Communications of the ACM 2020; 63:54–63. DOI: 10.1145/3381831. Available from: https://cacm.acm.org/research/green-ai/

50. Strubell E, Ganesh A, and McCallum A. Energy and Policy Considerations for Deep Learning in NLP. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2019 :3645–50. DOI: 10.18653/v1/P19-1355. Available from: https://aclanthology.org/P19-1355/

51. Henderson P, Hu J, Romoff J, Brunskill E, Jurafsky D, and Pineau J. Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning. Journal of Machine Learning Research 2020; 21. Missing DOI:1–43. Available from: https://jmlr.org/papers/volume21/20-312/20-312.pdf

52. Patterson D, Gonzalez J, Le Q, Liang C, Munguia L, Rothchild D, So D, Texier M, and Dean J. Carbon Emissions and Large Neural Network Training. arXiv 2021. eprint: 2104.10350. Available from: https://arxiv.org/abs/2104.10350

53. Dodge J, Prewitt T, Tachet des Combes R, Odmark E, Schwartz R, Strubell E, Luccioni AS, Smith NA, DeCario N, and Buchanan W. Measuring the Carbon Intensity of AI in Cloud Instances. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. Association for Computing Machinery, 2022 :1877–94. DOI: 10.1145/3531146.3533234. Available from: https://dl.acm.org/doi/10.1145/3531146.3533234

54. Luccioni AS, Viguier S, Ligozat A, and Jernite Y. Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model. Journal of Machine Learning Research 2023; 24. Missing DOI:1–53. Available from: https://www.jmlr.org/papers/volume24/23-0069/23-0069.pdf

55. ISO/IEC. Information Technology — Data Centres — Key Performance Indicators — Part 2: Power Usage Effectiveness (PUE). ISO/IEC 30134-2:2016; Missing DOI. 2016. Available from: https://www.iso.org/standard/63451.html

56. ISO/IEC. Information Technology — Data Centres — Key Performance Indicators — Part 6: Energy Reuse Factor (ERF). ISO/IEC 30134-6:2021; Missing DOI. 2021. Available from: https://www.iso.org/standard/71717.html

57. ISO/IEC. Information Technology — Data Centres — Key Performance Indicators — Part 8: Carbon Usage Effectiveness (CUE). ISO/IEC 30134-8:2022; Missing DOI. 2022. Available from: https://www.iso.org/standard/77691.html

58. ISO. Greenhouse Gases — Carbon Footprint of Products — Requirements and Guidelines for Quantification. ISO 14067:2018; Missing DOI. 2018. Available from: https://www.iso.org/standard/71206.html

59. World Resources Institute and World Business Council for Sustainable Development. The GHG Protocol: Corporate Accounting and Reporting Standard. Includes Scope 2 Guidance (2015); Missing DOI. 2015. Available from: https://ghgprotocol.org/corporate-standard