Orchestration as the Locus of Value in Regulated AI Reasoning: Why Frontier Model Commoditization Demands Compound Architectures

Bill Faruki
MindHYVE.ai, Inc.
Newport Beach, California · Islamabad · Nairobi
bill@mindhyve.ai

May 2026

Abstract

The dominant published direction in frontier artificial intelligence research has been to make individual models more capable through scale, sparse expert routing within a single model, and inference-time compute scaling. This paper argues that the locus of value in *regulated AI reasoning* — domains such as clinical medicine, statutory interpretation, financial analysis, and pedagogy where errors carry asymmetric and non-recoverable cost — is migrating in a different direction. As frontier model performance on general benchmarks commoditizes across labs, the value differential is shifting from individual model quality to *orchestration intelligence*: the layer that composes multiple frontier models at runtime under a trained, domain-specialized arbiter. We articulate a five-element taxonomy of compound architectures for regulated reasoning, survey existing implementations of each element across academic and commercial work, and identify a consistent gap: no published or shipped system instantiates all five elements as an integrated architecture. We position the recent reversal of Harvey AI from a custom legal foundation model to a multi-model router as a leading empirical data point that the bet on orchestration is rational, and we engage Self-MoA — the strongest published counterargument to multi-model diversity — directly. We formulate six falsifiable predictions under which the proposed thesis would be refuted, and we identify *verification-without-execution* as the hardest unsolved problem on the path. The paper closes by discussing Eve-Fusion F5, a compound architecture developed at MindHYVE.ai, as one instantiation of the proposed taxonomy rather than as its motivating example.

Keywords: *compound AI systems, multi-agent reasoning, model orchestration, process reward models, regulated artificial intelligence, agentic systems, mixture of agents, sector-specialized models*

1 Introduction

The trajectory of frontier artificial intelligence research over the past five years has been remarkably consistent. Capability gains have come predominantly from three directions: scaling parameters and training compute [1], [2]; introducing sparse expert routing within a single model architecture [3]; and extending inference-time compute through reasoning tokens, chain-of-thought elaboration, and reinforcement learning from verifiable rewards [4], [5]. The dominant architectural assumption underlying all three is that the locus of value sits within a single model — that better, larger, or longer-thinking individual models will solve progressively more of the field's open problems.

This paper argues that the assumption fails to hold in a specific but commercially and societally important class of problems: *regulated reasoning domains*. By this we mean domains in which (i) errors carry asymmetric and frequently non-recoverable cost, (ii) the verification function for a candidate answer cannot be reduced to executable code or

formally checkable proof, and (iii) the consequences of an answer cascade across institutionally distinct decision-making systems with their own evidentiary standards. Clinical medicine, statutory interpretation, financial analysis, and pedagogy share these properties. So do compliance, actuarial science, theological exegesis, and many engineering subdomains.

In these domains, two empirical patterns have emerged that the single-model-scaling paradigm does not account for. First, frontier models from different laboratories — trained on overlapping but not identical corpora, aligned under different objectives, and exhibiting different inductive biases — converge on similar performance ceilings on regulated-reasoning benchmarks while continuing to disagree on individual cases [6]. Second, the highest-profile commercial attempt to overcome this ceiling through vertical-specialized foundation models, Harvey AI's custom legal LLM trained in partnership with OpenAI, was abandoned by Harvey when frontier general-purpose models matched or exceeded its performance on Harvey's own benchmark [7]. Harvey transitioned from a custom-model strategy to a multi-model routing architecture, dispatching legal tasks to whichever frontier model performs best on each task type.

These patterns suggest a different research direction. If models from different labs converge in capability but diverge in their individual failures, the optimal architecture may not be the best single model — it may be a *trained composition* of multiple distinct models, under an arbiter capable of detecting and resolving disagreement in the failure modes that matter most to a given domain. This is the compound-architecture hypothesis. It is not new in the literature [8], [9], but it remains substantially under-explored relative to the resources poured into single-model scaling, and almost entirely unexplored as a deployed architecture in regulated commercial verticals.

The contribution of this paper is fourfold. Section 2 articulates a five-element taxonomy of compound architectures for regulated reasoning, formalizing distinctions that are blurred in the existing literature. Section 3 surveys the landscape of published and shipped systems against the taxonomy, establishing that no system currently instantiates all five elements as an integrated architecture. Section 4 engages the strongest published counterargument — Self-MoA [10] — directly, and addresses four further objections likely to be raised. Section 5 formulates six falsifiable predictions and identifies verification-without-execution as the hardest open problem. Section 6 briefly discusses Eve-Fusion F5, a compound architecture developed at MindHYVE.ai, as one instantiation of the proposed taxonomy.

2 A Taxonomy of Compound Architectures for Regulated Reasoning

We propose that the design space of compound AI architectures for regulated reasoning is usefully decomposed into five elements. Existing systems instantiate one, two, or occasionally three of these elements; we are not aware of any deployed system that instantiates all five. The elements are conceptually separable and can be developed and evaluated independently, although their integration is where we argue the strongest competitive and capability advantages emerge.

2.1 Element 1: Multi-Model Runtime Composition

We define *multi-model runtime composition* as the inference-time invocation of two or more *architecturally distinct* language models — models that differ not only in scale or fine-tuning but in pretraining corpus, base architecture, or laboratory of origin — within a single reasoning trajectory. This is distinct from Mixture-of-Experts (MoE) routing [3], which routes inference through sparsely activated expert subnetworks *within a single jointly trained model*. MoE produces architectural diversity through learned subnetworks of shared parameters; runtime composition produces architectural diversity through independently trained systems.

The distinction matters in regulated reasoning because the dominant failure class — shared training-data artifacts producing correlated hallucinations across models trained on overlapping corpora — is mitigated by independent training in a way that MoE routing within a single model cannot replicate. Two MoE experts within GPT-class models share their training data; two distinct models from different laboratories do not, except to the extent that public web corpora overlap.

2.2 Element 2: Trained Domain-Specialized Orchestration

We define *trained orchestration* as the use of a learned model — not a heuristic, decision tree, or general-purpose LLM acting as a router under prompting — to coordinate the multi-model composition described in Element 1. The orchestrator (i) decomposes the reasoning task into sub-steps, (ii) selects which constituent model executes each sub-step, and (iii) arbitrates when constituent models produce conflicting outputs on a given sub-step.

Process Reward Models (PRMs) [11], [12] provide the theoretical foundation for trained orchestration. A PRM is a learned model that evaluates intermediate reasoning steps rather than only final outputs, and recent work has demonstrated that PRMs as small as 1.7B parameters can approximate the judgment quality of full LLM judges using internal representations rather than generative output [13]. The orchestrator in a compound architecture is functionally a domain-specialized PRM — a small model trained on synthetic domain reasoning data (Element 4) that judges, plans, and routes across multiple larger generators.

2.3 Element 3: Federated Per-Vertical Cognitive Stacks

We define a *federated per-vertical cognitive stack* as a configuration in which each professional vertical receives a dedicated instance of the multi-model composition (Element 1) and the trained orchestrator (Element 2), with separate training data, separate verification frameworks, and separate compliance posture. This is distinct from current enterprise multi-agent frameworks [14], [15], [16], which provide orchestration scaffolding over a shared foundation model and differentiate agents through prompts, tools, and retrieval context.

The federation distinction matters because regulated verticals impose vertical-specific constraints — HIPAA in healthcare, attorney-client privilege in law, FERPA in education, fiduciary duty in finance — that propagate through the entire architecture, not just the surface layer. A federated architecture allows the orchestrator's training data, the constituent model selection, and the verification framework to be tuned per vertical; a shared-foundation architecture forces these to be uniform at the substrate level and addressed only through prompting at the surface.

2.4 Element 4: Synthetic Domain-Reasoning Corpora

We define a *synthetic domain-reasoning corpus* as a large-scale training dataset that encodes structured professional reasoning — clinical causality chains, statutory logic graphs, judicial precedent chains, financial inference trees, pedagogical scaffolding sequences — generated through a process other than scraping naturally occurring text. The exemplar in formal mathematics is AlphaGeometry [17], which generated 100 million synthetic geometry proofs from one billion random diagrams and trained from scratch without human demonstrations, achieving near-gold-medal performance on International Mathematical Olympiad problems.

The AlphaGeometry paradigm has not transferred to clinical, legal, or financial reasoning at any comparable scale. We argue in Section 5 that the obstruction is the verification problem: AlphaGeometry's synthetic proofs are programmatically checkable, but clinical diagnoses, legal arguments, and financial analyses are not. Solving this obstruction is, in our view, the central unsolved technical problem in compound architectures for regulated reasoning.

2.5 Element 5: Cross-Domain Meta-Reasoning

We define *cross-domain meta-reasoning* as a coordination layer above multiple per-vertical cognitive stacks (Element 3), capable of real-time reasoning synthesis when a single decision implicates multiple professional domains. The canonical example: a traumatic brain injury finding has simultaneous implications for legal damages calculation, insurance reserve requirements, long-term care cost projection, and rehabilitative pedagogy planning. Each implication requires reasoning within a distinct vertical; the meta-reasoner integrates them into a coherent recommendation.

This is distinct from what enterprise platforms term *cross-domain* when they refer to integrating HR, Finance, and IT workflows. Those are cross-functional business processes sharing organizational context. Cross-domain meta-reasoning as we define it coordinates reasoning systems with distinct evidentiary standards, distinct failure modes, and distinct regulatory frames — a substantially harder problem and, as we show in Section 3, one with no published or shipped precedent.

3 Survey of the Landscape

This section maps published academic work and shipped commercial systems against the five-element taxonomy of Section 2. We adopt an explicit tiering of evidence: *peer-reviewed* for academic work appearing in refereed venues; *shipped* for commercial products with public technical documentation; *inferred* for architectures deduced from vendor disclosures and analyst reporting; and *opaque* for internal research at frontier laboratories that has not been disclosed. The opaque category is irreducible: we cannot claim what does not exist in the laboratories of OpenAI, Anthropic, Google DeepMind, xAI, or Meta. We can claim only what has not been published or shipped.

3.1 Multi-Model Runtime Composition (Element 1)

The closest peer-reviewed precedent is LLM-Blender [8] (Allen Institute for AI, ACL 2023), which runs multiple LLMs in parallel and uses a trained 0.4B-parameter DeBERTa model (PairRanker) to evaluate outputs pairwise, followed by a separately trained GenFuser that synthesizes the final answer. PairRanker achieved near-GPT-4 correlation with human preferences at a fraction of the parameter count. LLM-Blender remains an academic open-source project; we are not aware of significant commercial deployment.

Together AI's Mixture-of-Agents (MoA) framework [9] (ICLR 2025) demonstrated that layering multiple distinct LLMs — including GPT-4o, Claude, Llama, and Mixtral — achieves 65.1% on AlpacaEval 2.0 against 57.5% for GPT-4 Omni alone. Vanguard subsequently deployed a variant of MoA in production for financial retrieval-augmented generation. The aggregator in MoA, however, is a general-purpose LLM operating under a synthesis prompt rather than a separately trained arbitration model — instantiating Element 1 but not Element 2.

xAI's Grok 4 Heavy (July 2025) and Grok 4.20 Beta (February 2026) are the closest shipped commercial products to Element 1. They run four agents simultaneously — a captain agent for task decomposition and three specialist agents that engage in multi-round internal discussion before converging on a response. Inferred from xAI's available technical documentation: the four agents appear to be instances of the same Grok base model under different role prompts, not architecturally distinct models. If this inference is correct, Grok 4 Heavy is more accurately characterized as multi-instance role-prompted composition rather than architecturally distinct multi-model composition. The arbitration mechanism is embedded in the agents' discussion process rather than implemented as a separately trained reasoning model.

We find **no** published or shipped evidence of multi-model runtime composition with architecturally distinct constituent models under a separately trained domain-specialized orchestrator at any frontier laboratory. OpenAI's consensus@N samples multiple completions from the same model and applies majority voting [18], which is a different paradigm — single-model multi-sample, not multi-model composition. Internal research at frontier laboratories is opaque; this gap may close upon disclosure.

3.2 Trained Domain-Specialized Orchestration (Element 2)

VersaPRM [11] (February 2025) is the closest peer-reviewed precedent. It is a multi-domain Process Reward Model trained on synthetic reasoning data that evaluates step-by-step reasoning from larger models. On MMLU-Pro Law tasks, VersaPRM achieved a 7.9% performance gain over majority voting, substantially exceeding Qwen2.5-Math-PRM's 1.3% gain on comparable mathematical reasoning tasks [12]. INSPECTOR [13] (January 2026) further demonstrated that arbiter models as small as 1.7B parameters can approximate full LLM-judge quality using internal representations.

In commercial systems, the closest analogue is Harvey AI's current multi-model router [7], which dispatches legal tasks to whichever frontier model performs best on each task type. The router is, however, a task *selector*, not an output *arbiter*. The distinction is structural: a selector decides who runs the task; an arbiter evaluates how it was run, identifies disagreement, and produces a synthesized output. Harvey's transition from a custom legal foundation model to a multi-model router validates the principle that orchestration becomes increasingly valuable as frontier models commoditize [7], but does not yet instantiate Element 2.

Thomson Reuters CoCounsel [19] and LexisNexis Protégé [20] operate multi-agent systems with specialist agents for research, planning, and drafting. To our knowledge, neither employs a trained small model to arbitrate between larger model outputs.

3.3 Federated Per-Vertical Cognitive Stacks (Element 3)

Hippocratic AI's Polaris constellation [21] is the only commercial system that approaches the depth of vertical specialization implied by Element 3. Polaris deploys 22+ specialized support models that supervise a primary conversational LLM during clinical interactions. Hippocratic reports 99.38% clinical accuracy across seven million calls and three trillion tokens of medical training data. Polaris uses a many-supervisors-one-generator pattern; Element 3 as we define it admits either this pattern or the one-orchestrator-many-generators pattern of Element 2. Critically, Polaris is single-vertical (healthcare). We are not aware of any federated multi-vertical implementation of comparable depth.

Microsoft AutoGen [14], CrewAI [15], LangGraph [16], and the OpenAI Agents SDK [22] are orchestration scaffolding layers over shared foundation models. Salesforce Agentforce [23] powers all of its agents through a single shared Atlas Reasoning Engine, differentiated only through declarative topics defined in YAML. Palantir AIP [24] differentiates through its Ontology layer — a semantic data model — rather than through dedicated per-vertical reasoning models. IBM watsonx and the Granite model family [25] train foundation models on domain-curated data from five verticals but produce a general-purpose model family rather than dedicated per-vertical reasoning engines.

The federated multi-vertical configuration — Hippocratic-depth specialization across healthcare, legal, education, finance, and additional regulated verticals simultaneously, with separate orchestrators per vertical — has no commercial implementation we have identified.

3.4 Synthetic Domain-Reasoning Corpora (Element 4)

AlphaGeometry [17] is the canonical demonstration that synthetic reasoning data trains effective reasoning systems when the verification function is clean. Microsoft's Phi and Orca series [26], [27] explicitly teach structured reasoning strategies to small models using synthetic data; the Orca 2 paper demonstrated that different tasks benefit from different reasoning strategies, including step-by-step, recall-then-generate, and extract-then-generate approaches. Phi-4 [28] is the most recent member of the lineage and is the model class most plausibly suited to serve as a trained orchestrator under Element 2.

In healthcare, the dominant published direction for synthetic data is privacy-preserving electronic health record replication [29] — statistical replicas that protect patient health information. We have not identified large-scale synthetic clinical causality datasets or bioethics reasoning corpora in the published literature. In legal AI, SDD-LawLLM [30] (2025) used GPT-4 to construct a chain-of-thought legal dataset and fine-tuned Qwen-7B on it. This is, to our knowledge, the only documented attempt of its kind; it is small-scale and narrow in scope. We have not identified large-scale synthetic statutory logic, judicial reasoning, or precedent modeling corpora. In finance, synthetic data activity focuses on tabular augmentation and fraud detection scenarios [31]; we have not identified published activity in structured financial reasoning datasets at scale.

3.5 Cross-Domain Meta-Reasoning (Element 5)

EvenUp's Piai platform [32] is the closest commercial precedent. Piai bridges medical understanding (parsing records, identifying ICD codes) with legal strategy (demand letters, settlement benchmarking) for personal injury cases. EvenUp is, however, a monolithic domain-specific platform with both medical and legal knowledge embedded in a single system. It is not a federated coordination of separate cognitive stacks, and we observe that adding a third domain (actuarial science, financial planning) or repurposing the architecture for a different cross-domain use case (compliance = regulatory + financial + legal) is not straightforward given the monolithic design.

Academically, SciAgents [33] (Advanced Materials, 2025) automates cross-disciplinary scientific hypothesis generation using multiple agents with knowledge graphs, but operates in scientific discovery rather than professional services reasoning. A January 2026 arXiv survey on multi-agent architectures [34] discusses cross-domain

collaboration as a future direction, confirming that real-time cross-domain professional reasoning synthesis remains aspirational in the published literature.

Coordination infrastructure has matured rapidly. Google's Agent-to-Agent (A2A) protocol [35], Anthropic's Model Context Protocol (MCP) [36], and the hub-and-spoke agent patterns adopted across major frameworks provide the communication and coordination primitives. The reasoning layer that understands cross-domain professional implications — that a medical finding changes a legal strategy which affects a financial projection — does not yet exist as a deployed system. The current state of the field is task routing, not reasoning synthesis.

3.6 Summary

Table 1 summarizes the survey. The pattern is consistent across all five elements: partial precedents exist, often impressive ones, but no published or shipped system instantiates the full integrated architecture. The competitive gap is not in any single element — each element has an academic or commercial precedent of some kind — but in the combination.

Element	Closest precedent	Status	Gap to integrated architecture
1: Multi-model composition	LLM-Blender; MoA; Grok 4 Heavy	Peer-reviewed; shipped (single-base)	Architecturally distinct constituents under trained orchestrator
2: Trained orchestration	VersaPRM; INSPECTOR; Harvey router	Peer-reviewed; commercial routing	Domain-specialized arbiter deployed in regulated vertical
3: Federated per-vertical stacks	Hippocratic AI Polaris (1 vertical)	Shipped (single-vertical)	Multi-vertical federation with comparable depth
4: Synthetic reasoning corpora	AlphaGeometry (math); Phi/Orca	Peer-reviewed (math/general)	Large-scale corpora for clinical/legal/financial reasoning
5: Cross-domain meta-reasoning	EvenUp Piai (monolithic, 2 domains)	Shipped (monolithic)	Federated coordination across N professional verticals

Table 1: Summary of the five-element taxonomy against the public landscape of compound AI architectures, May 2026.

4 Counterarguments

A position paper is only useful if it engages the strongest objections. We address five.

4.1 Self-MoA: Diversity Does Not Help

Princeton's Self-MoA paper [10] (February 2025) found that ensembling outputs from a single strong model outperformed mixing different LLMs by 6.6% on AlpacaEval 2.0. If this result generalizes, the diversity premise underlying multi-model composition is wrong, and the path forward is to ensemble samples from the strongest available single model rather than to coordinate multiple distinct models.

We argue that Self-MoA is the correct objection to take seriously, but that it does not refute the compound-architecture thesis for regulated reasoning, for two reasons.

First, AlpacaEval 2.0 is a general instruction-following benchmark. The diversity argument for multi-model composition in regulated reasoning is not that diversity improves general task quality — it is that diversity hedges against *shared training-data failure modes* that do not appear in general benchmarks. Two models trained on overlapping pretraining corpora will hallucinate the same fabricated case citation, miss the same drug-drug interaction, or share the same blind spot in a precedent chain. A trained arbiter that can detect disagreement between architecturally distinct models checks for exactly this class of failure. Self-MoA's finding may generalize to all benchmarks where shared-training-data correlations do not dominate the failure distribution; we conjecture that regulated-vertical benchmarks are not in this class.

Second, Self-MoA tested ensembling without a separately trained arbiter — the synthesis step was a general LLM under a synthesis prompt, comparable to the MoA aggregator. Our taxonomy requires both Element 1 (architecturally distinct constituents) and Element 2 (trained domain-specialized arbitration). Self-MoA's results bear on Element 1 alone.

We acknowledge that the burden of proof shifts onto compound architectures to demonstrate the conjecture empirically in regulated benchmarks. To our knowledge, the comparison has not been published. We identify it in Section 5 as a falsifiable prediction.

4.2 Frontier Laboratories May Be Doing This Internally

OpenAI, Anthropic, Google DeepMind, xAI, and Meta all conduct extensive research that is not publicly disclosed. The thesis advanced in this paper rests on the absence of published and shipped architectures matching the integrated five-element pattern; it does not rest on the absence of internal research. We acknowledge that internal architectures may match or exceed what we describe.

We observe, however, that the commercial incentives of frontier laboratories are structurally aligned with foundation-model API capacity sales, not with orchestration infrastructure that abstracts above their layer. A laboratory that productizes a compound architecture using its competitors' models as constituents accepts a strategic posture that the laboratory's own commercial team will resist. The compound-architecture pattern is more naturally pursued by entities whose business model is not selling foundation-model capacity — by definition, entities outside the frontier-laboratory set. This observation does not foreclose the possibility of internal research; it identifies a reason such research, if it exists, may remain internal rather than shipped.

4.3 Mixture-of-Experts Is Equivalent

MoE architectures route inference through sparsely activated expert subnetworks within a single model [3]. An objector might argue that this is already multi-expert composition and that the additional infrastructure of multi-model runtime composition is unnecessary.

We have addressed this in Section 2.1. MoE is training-time composition: the experts are learned subnetworks of a single model, trained jointly, sharing parameters, embeddings, and pretraining corpus. Multi-model runtime composition is inference-time composition across separately trained systems with non-shared training data and non-shared inductive biases. The diversity in MoE is architectural variation within a learned space; the diversity in runtime composition is independent-witness diversity across separately trained systems. These are structurally different sources of diversity, and they hedge against different failure modes.

4.4 Harvey AI's Reversal Refutes Vertical Specialization

Harvey AI abandoned its custom legal foundation model when frontier models matched or beat its performance. An objector might argue that this proves vertical-specialized foundation models are a dead end, and that orchestration is merely an admission that the underlying models are adequate on their own.

Harvey's reversal proves a narrower claim: a vertical-specialized *foundation model* is difficult to keep ahead of commoditizing frontier models. The compound-architecture thesis advanced in this paper does not propose vertical-specialized foundation models. Element 2 proposes a small trained orchestrator — not a foundation model — that is vertical-specialized in its *coordination of frontier models*, not in its capacity to generate end-user content. Harvey's transition from custom-model to multi-model router validates the underlying claim that orchestration is where commercial value accrues as frontier models commoditize [7]. Harvey landed at the layer above its custom model; the compound-architecture pattern adds a trained arbiter at that same layer.

4.5 The Compounding-Moat Argument Is Rhetorical

A final objection: the claim that the five elements compound into a defensible competitive position is asserted rather than measured. What is the actual cost of replication?

We offer estimated ranges, stated with appropriate uncertainty:

- **Synthetic sector dataset (one vertical, Element 4):** 18 to 30 months for a team of 8–12 including domain experts, machine learning engineers, and verification framework designers. The verification framework is the long pole and the principal research risk.
- **Sector-specialized orchestrator (Element 2):** 3 to 6 months conditional on the synthetic dataset existing and on access to base Phi-class weights and a competent fine-tuning team.
- **Multi-model runtime composition layer (Element 1):** 6 to 12 months as standalone engineering; faster when treated as the natural layer above an existing orchestrator.
- **Per-vertical productization (Element 3):** 6 to 9 months per vertical for surface, integration, compliance posture, and go-to-market.
- **Cross-domain meta-reasoner (Element 5):** 12 to 18 months and requires at least three operational vertical stacks as substrate; it cannot meaningfully precede them.

These are estimates from informal industry conversation and our own engineering experience; we do not claim they are precise. They are intended to render the compounding-moat claim falsifiable. A serious competitor with capital and talent can replicate any single element in under two years. Replicating all five in parallel, with the dependencies between them, is a substantially different organizational commitment. The defensibility is in the integration cost and the time-to-coherence, not in the difficulty of any single element.

5 Falsifiable Predictions

We commit the thesis advanced in this paper to six predictions whose disconfirmation would constitute substantial evidence against the position. These are formulated to be tractable, monitorable, and time-bounded.

1. **P1.** Within 36 months of this paper's publication, a frontier laboratory will publish or ship a compound architecture instantiating Elements 1 and 2 with architecturally distinct constituent models and a separately trained orchestrator, deployed in a regulated vertical with published benchmark performance against single-model baselines. *If P1 fails to occur,* the position that frontier laboratories are commercially mis-incentivized to pursue compound architectures (Section 4.2) is strengthened. *If P1 occurs,* the architectural-whitespace claim narrows substantially and the position must shift to execution differentiation.
2. **P2.** Subsequent peer-reviewed work extending Self-MoA [10] will not generalize its single-model-ensemble advantage to regulated-vertical benchmarks (clinical reasoning, statutory interpretation, financial analysis). *If P2 fails,* Element 1 of the taxonomy is undermined: multi-model composition would no longer be empirically justified over single-model ensembling.
3. **P3.** Hippocratic AI [21] or a comparable vertical-deep system will not federate to a multi-vertical product within 36 months. *If P3 fails,* Element 3 of the taxonomy is no longer architecturally novel and reduces to a comparison of execution.
4. **P4.** Harvey AI or a comparable commercial multi-model router will add a trained arbiter component layered over routing within 24 months. *If P4 occurs,* the thesis of orchestration-as-locus-of-value is strengthened; the architectural pattern advanced in this paper becomes the default in commercial legal AI.
5. **P5.** Within 36 months, a peer-reviewed methodology for synthetic reasoning corpora in clinical, legal, or financial domains at the scale of AlphaGeometry will be published, solving or substantially mitigating the verification-without-execution problem. *If P5 occurs,* Element 4 becomes commoditized infrastructure rather than a defensible research direction. *If P5 fails,* the verification problem is confirmed as the central technical obstruction to compound architectures for regulated reasoning.
6. **P6.** Within 36 months, a reference implementation of A2A [35] or MCP [36] will ship with built-in cross-domain professional reasoning synthesis as a default capability. *If P6 occurs,* Element 5 collapses into protocol infrastructure rather than a defensible architectural layer.

5.1 The Hardest Open Problem: Verification Without Execution

We close this section by identifying what we view as the central unsolved technical problem on the path to compound architectures for regulated reasoning. Reinforcement Learning from Verifiable Rewards (RLVR) [5], [4] — the paradigm underlying DeepSeek-R1, the OpenAI o-series, and most published reasoning improvements — works because mathematical answers can be checked and code can be executed. The reward signal is clean, dense, and automatable.

Clinical diagnoses, legal arguments, and financial analyses cannot be programmatically verified. A clinical reasoning chain may be locally coherent and globally wrong; a legal argument may cite real precedents and synthesize them into a conclusion that does not survive appellate scrutiny; a financial analysis may apply correct formulas to a misframed problem. The reward function in these domains is expert judgment, which is expensive, slow, and subject to inter-rater disagreement.

Generating large-scale synthetic reasoning corpora (Element 4) requires bootstrapping a verification function from expert knowledge that is not encoded in executable form. This is, in our view, the hardest research problem in the field. AlphaGeometry's success demonstrates the methodology when the verification function is clean; transferring the methodology to domains where the verification function is irreducibly judgmental is an open problem that has not been solved in the published literature.

We conjecture that progress on this problem will come from one of three directions: (i) hybrid verification combining executable sub-checks with structured expert review at decision points where executability fails, (ii) inter-model agreement as a noisy but scalable proxy for expert judgment, calibrated against human review, or (iii) domain-specific formalisms that render fragments of the reasoning chain executable even where the global reasoning is not. None of these is established. We commend the problem to the field.

6 An Instantiation: Eve-Fusion F5

We close by briefly describing one instantiation of the proposed five-element taxonomy. Eve-Fusion F5 is a compound reasoning architecture developed at MindHYVE.ai. The naming convention F5 denotes the fifth generation of the architecture; sector instances are denoted Eve-Education F5, Eve-Healthcare F5, Eve-Legal F5, Eve-Theology F5, and so on, indicating the vertical fine-tuning of the orchestrator. The naming does not imply a fixed constituent model count; the constituent set is reconfigurable per release.

The architecture instantiates the five elements as follows. *Element 1 (multi-model runtime composition):* three architecturally distinct frontier models — at the time of writing, Claude Opus 4.7, GPT-5.4, and one additional best-fit model selected per release — are composed at inference time. *Element 2 (trained orchestration):* a Phi-4 reasoner [28] fine-tuned with LoRA [37] adapters on a sector-specific synthetic reasoning corpus (Element 4) plans the reasoning steps and selects the constituent model best suited to each step. *Element 3 (federated per-vertical stacks):* each sector receives its own five-layer stack — infrastructure, orchestrator, compound architecture, Digital Employee surface, and Agentic Operating System layer — with sector-specific compliance posture and constituent model selection. *Element 4 (synthetic reasoning corpora):* a corpus termed Eve-Genesis is generated per vertical and used to fine-tune the orchestrator. *Element 5 (cross-domain meta-reasoning):* a coordination layer above the per-vertical stacks is under active development; we do not yet claim it as fully deployed.

We do not, in this paper, claim performance results for Eve-Fusion F5. The purpose of this section is to demonstrate that the taxonomy of Section 2 is implementable, not to advance a specific instantiation as exemplary. Empirical evaluation of Eve-Fusion F5 against single-model baselines on regulated-vertical benchmarks is the subject of forthcoming work. We acknowledge that the absence of empirical results in this paper limits its claims to the conceptual and the architectural; we believe this limitation is appropriate for a position paper and that empirical work belongs in a separate venue.

7 Related Work

We have woven citations of related work throughout the paper as appropriate to the argument; this section briefly highlights three threads we have engaged less centrally but consider important context.

Agentic AI and the boundary of single-model reasoning. A growing literature argues that genuinely agentic systems require a separation between the reasoning layer and the execution layer, rather than a single model invoking tools through prompted control flow [38], [39]. Our Element 2 takes this position; we observe that the boundary is sharper when the reasoning layer is itself a small trained model rather than a frontier model under prompting.

Compound AI systems. The term *compound AI system* was introduced in the Berkeley AI Research blog post by Zaharia et al. [40] to describe systems built from multiple components rather than from a single monolithic model. The five-element taxonomy of Section 2 can be understood as a specialization of the compound AI systems framing to the regulated-reasoning setting, where the structural constraints we have described impose specific architectural commitments not necessarily required in general compound systems.

Domain-specialized models in regulated industries. An extensive literature explores domain-specialized models in medicine [41], law [42], and finance [43]. The position advanced in this paper is not opposed to vertical specialization; it argues that, in light of frontier model commoditization, the appropriate locus of vertical specialization is the orchestrator, not a vertical-specialized foundation model in the Bloomberg GPT or original Harvey mold.

8 Conclusion

Frontier AI research has bet, on the whole, that the locus of value in AI reasoning sits within increasingly capable single models. This paper has argued that for the class of regulated reasoning domains — clinical medicine, statutory interpretation, financial analysis, and the analogous high-stakes verticals — the bet under-specifies the architectural space. As frontier model capabilities converge across laboratories, the value differential migrates to the orchestration layer that composes them. The compound-architecture pattern, instantiated through the five-element taxonomy we have proposed, is currently substantially under-explored relative to the resources poured into single-model scaling, and is almost entirely unexplored as a deployed architecture in commercial regulated verticals.

We do not claim that the position advanced here is established. We claim that it is sufficiently coherent, sufficiently distinct from the prevailing direction, and sufficiently consequential to warrant the falsifiable predictions of Section 5 and the empirical work that those predictions imply. We commend the verification-without-execution problem to the research community as, in our view, the central technical obstruction on the path.

The position itself is straightforward to state. Better single models will continue to matter. They will not, by themselves, solve regulated reasoning. The architecture that does will be compound, orchestrated by a trained domain-specialized arbiter, federated across verticals, fed by synthetic reasoning corpora whose verification problem has been solved or substantially mitigated, and capable of cross-domain coordination. None of these elements is unprecedented in isolation. Their integration remains the open frontier.

Acknowledgments

The author thanks the engineering, operations, and research staff at MindHYVE.ai whose collective work on the architecture described in Section 6 has shaped the formulation of the taxonomy in Section 2. The author also acknowledges that the abstraction of the taxonomy from the particulars of any one implementation benefited substantially from sustained conversation with colleagues working in clinical, legal, and pedagogical AI deployment. Responsibility for errors and for the positions advanced rests with the author alone.

References

[1] Kaplan, J., McCandlish, S., Henighan, T., et al. (2020). *Scaling Laws for Neural Language Models*. arXiv:2001.08361.

[2] Hoffmann, J., Borgeaud, S., Mensch, A., et al. (2022). *Training Compute-Optimal Large Language Models*. arXiv:2203.15556.

[3] Fedus, W., Zoph, B., & Shazeer, N. (2022). Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *Journal of Machine Learning Research*, 23.

[4] OpenAI (2024). *Learning to Reason with LLMs*. Technical report.

[5] DeepSeek-AI (2025). DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948.

[6] Hendrycks, D., Burns, C., Basart, S., et al. (2021). Measuring Massive Multitask Language Understanding. *ICLR 2021*.

[7] Harvey AI (2025). The Multi-Model Future of Legal AI: Lessons from BigLaw Bench. Harvey AI engineering blog.

[8] Jiang, D., Ren, X., & Lin, B. Y. (2023). LLM-Blender: Ensembling Large Language Models with Pairwise Ranking and Generative Fusion. *ACL 2023*.

[9] Wang, J., Wang, J., Athiwaratkun, B., et al. (2025). Mixture-of-Agents Enhances Large Language Model Capabilities. *ICLR 2025*.

[10] Li, M., Zhao, S., Wang, Q., et al. (2025). Rethinking Mixture-of-Agents: Is Mixing Different Large Language Models Beneficial? arXiv:2502.00674.

[11] Zeng, T., Jin, C., Chen, M., et al. (2025). VersaPRM: Multi-Domain Process Reward Model via Synthetic Reasoning Data. arXiv:2502.06737.

[12] Lightman, H., Kosaraju, V., Burda, Y., et al. (2024). Let's Verify Step by Step. *ICLR 2024*.

[13] Zhou, Y., Lu, X., Wang, S., et al. (2026). INSPECTOR: A Small Model for Step-Level Reasoning Verification via Internal Representations. arXiv:2601.04812.

[14] Wu, Q., Bansal, G., Zhang, J., et al. (2024). AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. *COLM 2024*.

[15] CrewAI Inc. (2025). *CrewAI: Framework for Orchestrating Role-Playing, Autonomous AI Agents*. Technical documentation.

[16] LangChain Inc. (2025). *LangGraph: Building Stateful, Multi-Actor Applications with LLMs*. Technical documentation.

[17] Trinh, T., Wu, Y., Le, Q., He, H., & Luong, T. (2024). Solving Olympiad Geometry without Human Demonstrations. *Nature*, 625, 476–482.

[18] Wang, X., Wei, J., Schuurmans, D., et al. (2023). Self-Consistency Improves Chain of Thought Reasoning in Language Models. *ICLR 2023*.

[19] Thomson Reuters (2024). *CoCounsel: Generative AI for Legal Professionals*. Product technical documentation.

[20] LexisNexis (2025). *Protégé: Personalized AI Assistant for Legal Professionals*. Product technical documentation.

[21] Hippocratic AI (2024). Polaris: A Safety-Focused LLM Constellation Architecture for Healthcare. arXiv:2403.13313.

[22] OpenAI (2025). *Agents SDK and Responses API*. Developer documentation.

[23] Salesforce (2024). *Agentforce and the Atlas Reasoning Engine*. Salesforce engineering blog.

[24] Palantir Technologies (2024). *Palantir AIP: The Ontology Layer*. Technical white paper.

[25] IBM Research (2024). Granite Foundation Models for Business. IBM Research technical report.

[26] Mukherjee, S., Mitra, A., Jawahar, G., et al. (2023). Orca: Progressive Learning from Complex Explanation Traces of GPT-4. arXiv:2306.02707.

[27] Mitra, A., Del Corro, L., Mahajan, S., et al. (2023). Orca 2: Teaching Small Language Models How to Reason. arXiv:2311.11045.

[28] Microsoft Research (2024). Phi-4 Technical Report. arXiv:2412.08905.

[29] Hernandez, A., Kuo, Y. H., Yarbrough, P. M., et al. (2023). Synthetic Data Generation for Electronic Health Records: A Review. *JAMIA*, 30(7).

[30] Liu, X., Zhao, Y., & Chen, P. (2025). SDD-LawLLM: A Synthetic Domain-Driven Approach to Legal Language Model Fine-Tuning. arXiv:2503.07241.

[31] Assefa, S. A., Dervovic, D., Mahfouz, M., et al. (2020). Generating Synthetic Data in Finance: Opportunities, Challenges and Pitfalls. *ICAIF 2020*.

[32] EvenUp Inc. (2025). *Piai: AI for Personal Injury*. Product technical documentation.

[33] Ghafarollahi, A. & Buehler, M. J. (2025). SciAgents: Automating Scientific Discovery through Multi-Agent Intelligent Graph Reasoning. *Advanced Materials*, 37(8).

[34] Tran, K., Vatsalan, D., Hassan, M. M., et al. (2026). Multi-Agent LLM Systems: A Survey of Architectures, Applications, and Open Challenges. arXiv:2601.09712.

[35] Google (2025). *Agent-to-Agent Protocol (A2A) Specification*. Technical specification, version 1.0.

[36] Anthropic (2024). *Model Context Protocol Specification*. Technical specification.

[37] Hu, E. J., Shen, Y., Wallis, P., et al. (2022). LoRA: Low-Rank Adaptation of Large Language Models. *ICLR 2022*.

[38] Faruki, B. (2026, May 22). The Five-Layer Stack of Agentic AI. LinkedIn engineering post.

[39] Schick, T., Dwivedi-Yu, J., Dessì, R., et al. (2024). Toolformer: Language Models Can Teach Themselves to Use Tools. *NeurIPS 2023*.

[40] Zaharia, M., Khattab, O., Chen, L., et al. (2024). The Shift from Models to Compound AI Systems. *Berkeley Artificial Intelligence Research Blog*, February 2024.

[41] Singhal, K., Tu, T., Gottweis, J., et al. (2025). Toward Expert-Level Medical Question Answering with Large Language Models. *Nature Medicine*, 31.

[42] Henderson, P., Krass, M., Zheng, L., et al. (2024). Pile of Law: Learning Responsible Data Filtering from the Law and a 256GB Open-Source Legal Dataset. *NeurIPS 2024 Datasets and Benchmarks Track*.

[43] Wu, S., Irsoy, O., Lu, S., et al. (2023). BloombergGPT: A Large Language Model for Finance. arXiv:2303.17564.

Appendix A arXiv Submission Metadata

The following metadata is provided to accompany an arXiv submission of this paper. The author acknowledges the limitations of self-categorization and welcomes reclassification by arXiv moderators.

- **Primary category:** cs.AI (Artificial Intelligence)
- **Cross-list categories:** cs.MA (Multi-Agent Systems); cs.LG (Machine Learning)
- **MSC classification:** 68T01 (General topics in artificial intelligence); 68T05 (Learning and adaptive systems in artificial intelligence)
- **ACM classification:** I.2.11 (Distributed Artificial Intelligence — Multiagent systems); I.2.7 (Natural Language Processing)
- **Comments:** Position paper. 4,200 words. 43 references. No experiments; conceptual and architectural contribution. Author corresponds at bill@mindhyve.ai.