WORKING PAPER · VERSION 2.0 · JULY 2026

Supply Chain of Intelligence

A framework for understanding where competitive advantage accumulates in AI markets

ABSTRACT

As generative capability commoditizes, the question of where durable advantage accumulates in artificial intelligence markets is unresolved by existing strategy theory, which was built for markets in which the scarce input was the technology itself. This paper develops a supply-chain treatment of generative AI: ten layers of production from physical resources to compounding memory, crossed by three market currents that determine whether a defensible position becomes a viable business. From that structure it derives four propositions, of which one is primary: wherever output carries fiduciary, regulatory, safety, or reputational weight, generation and verification separate into distinct economic entities, and that separation is stable under improvements in generator capability. The paper positions this claim against appropriability, ecosystem, modularity, and information-asymmetry literatures; presents six market vignettes consistent with it; states falsifiable predictions and refutation criteria; and discusses the alternative explanations that would make the claim redundant. The taxonomy is treated as an instrument for locating firms rather than as the contribution.

Anand Arivukkarasu — Kellogg MBA; former product leader, Meta (Instagram); VP / Head of Product roles at Ideas2IT, Refersion, GRIN. San Francisco. Written in a personal capacity.

Canonical source: supplychainofai.com/framework · supplychainofai.com/papers

Contents

1. Introduction

The dominant question asked of artificial intelligence markets between 2023 and 2026 was which model would win. It was the wrong question, in the specific sense that its answer does not determine the distribution of profit. Model capability improved by roughly an order of magnitude over that period while the price of a unit of capability fell by more; the firms that captured durable margin were, with few exceptions, not the firms that produced the capability.

This is not anomalous. It is the ordinary behaviour of an industrial supply chain. Value in a supply chain concentrates at scarcity, is squeezed at bottlenecks, and migrates the moment a layer below commoditizes what the layer above was charging for. The mistake embedded in the popular “AI stack” diagram is not that it is inaccurate about the technical dependency order — it is largely correct about that — but that it implies value flows cleanly upward with the dependency. It does not.

> *“Intelligence is a supply chain. Value accrues at the bottlenecks, not at the most visible node.”*

This paper takes that sentence seriously and asks what follows from it. Section 3 sets out the production structure: ten layers, crossed by three market currents. Section 4 states four structural propositions. Section 5 develops the primary claim, the verification boundary, which is the paper's candidate contribution to theory. The remainder positions, evidences, and attempts to falsify it.

A note on register. The taxonomy in Section 3 is an instrument. It is useful for locating a firm precisely enough to argue about it, and it is used throughout as a shared coordinate system. It is not the theoretical claim, and a reader who rejects the partition entirely can still evaluate Section 5 on its own terms.

2. Related literature

2.1 Appropriability and complementary assets

Teece (1986) established that innovators capture value from an innovation only when they control the complementary assets required to commercialise it, and that specialized complementary assets shift bargaining power decisively. Applied to generative AI, the literature's natural prediction is that model producers, holding the innovation, will integrate forward into whichever complementary assets are specialized, and will capture the resulting rents. Much of what has happened is consistent with that prediction: model providers have integrated into interfaces, agents, developer tooling, and consumer surfaces. Section 5 argues there is exactly one class of complementary asset for which the prediction inverts, because acquisition destroys the asset.

2.2 Positioning, resources, and the value chain

Porter (1985) supplies the vocabulary of chains and positions; Barney (1991) supplies the resource-based conditions for a resource to sustain advantage. Both are compatible with the framework here and neither is displaced by it. What neither supplies is an endogenous rule that predicts *where a chain must break into separate firms*. Firm boundaries in both traditions are treated as a strategic choice subject to cost and capability constraints.

2.3 Modularity and interfaces

Baldwin and Clark (2000) explain how interfaces partition systems and where option value accumulates at module boundaries. The verification boundary described here is not of this kind: it survives even when the technical cost of integrating across it falls to zero, because the constraint is institutional rather than architectural. Distinguishing institutional from architectural boundaries is, in the author's view, underdeveloped in the AI strategy literature.

2.4 Ecosystems and platform expansion

Jacobides, Cennamo and Gawer (2018) and Adner (2017) formalise complementarity and bottlenecks in ecosystems and give a language for hub roles. The open question in that literature, sharpened by the current wave, is where hub expansion stops. Proposition IV proposes a specific, testable answer for one class of role.

2.5 Information asymmetry, signalling, and audit

Akerlof (1970), Spence (1973) and Holmström (1979) provide the mechanism that makes the verification boundary work: a signal informs only if it is differentially costly to send falsely. Power (1997) documents how societies institutionalise verification and, importantly, that such institutions ratchet rather than relax. The contribution here is to run this machinery forward into a firm-boundary prediction in a market where the informed party's capability is improving faster than in any previously studied setting.

2.6 What remains unexplained

Taken together, the literatures explain why capability alone does not confer advantage, why complementary assets matter, and why information asymmetry sustains intermediaries. They do not, jointly, predict that the most capable firm in a market will be structurally barred from performing a specific adjacent function *because of its capability rather than in spite of it*. That is the claim under test.

3. The framework: ten layers and three currents

The production structure is partitioned into ten layers. The partition is a descriptive instrument chosen for one property: each layer names a place where a firm can hold a position that another firm must pay to cross. Full definitions, including fifty sublayers, are in Appendix A.

Layer	What is produced or controlled	Structural read
L-1 Resources	What supports the chain. Energy, water, fabs, materials, skilled trades, the inputs the entire stack consumes.	The real bottleneck. Slow to build, impossible to fake.
L0 Infrastructure	The shovels. Chips, data centers, networking, cloud, edge, what is needed to process intelligence.	Shovel sellers win every gold rush.
L1 Data	The raw input. What data do you have that nobody else can get?	Structurally safe. API-first wins.
L2 Models	Intelligence refinement. Rent early, build custom at scale.	Winner-take-most. Commodity risk high.
L3 Gatekeeping	Trust, acceptance, approval. Can the system be allowed in?	Essential. More agents = more access control.
L4 Access	Connectivity, permissions, integrations, the pipes layer.	Load-bearing walls. Invest accordingly.

Layer	What is produced or controlled	Structural read
L5 Execution	Applied skills and capabilities. Doing the actual work.	Durable if deep. Generic skills get absorbed.
L6 Orchestration	Workflow, routing, coordination. How skills compose into outcomes.	Contested. Becoming a feature, not a product.
L7 Surface	Interface, presentation, experience. How the user meets the intelligence.	Modality = commodity. Context = moat.
L8 Memory	Retention, learning, compounding context. What the system remembers.	The ultimate moat. Memory that compounds wins.

The layers group into three registers with sharply different half-lives. The distinction matters more than the individual layer names: it is the reason that two products which look identical to a user can have entirely different futures.

Register	Layers	Half-life	Behaviour
Substrate	L-1, L0, L1, L2, L3, L8	What users depend on. Compounds in years.	Proprietary data, trust gates, compounding memory. Slow to build, slow to lose.
Workflow	L4, L5, L6	What users live inside. Survives in months.	Sticky if deep, owned, and integrated. Absorbed if generic.
Surface	L7	What users touch. Commoditizes in weeks.	Platforms ship this for free. Placement and habit are the only durable parts.

The ten layers describe supply. They are crossed by three market currents that determine whether a position at any layer compounds into a business or starves. Geopolitics and regulation are not currents in this treatment: they act at their native layers (physical resources and gatekeeping respectively), and promoting them would double-count.

Current I — Demand Gravity

Where the budget actually sits, and what it pulls toward. As L2 prices collapse, discretionary spend migrates from buying generation to buying outcomes (L5 + L8), verification (L3), and access to proprietary data (L1). A defensible layer with no buyer is worth zero.

Operationalisation: Name the buyer, the budget line, and the line item they stop paying for once L2 is free.

Current II — Attention Economics

What becomes scarce when generation becomes infinite. Default placement, OS integration, habit loops and on-ramp ownership decide which intelligence is actually used. The large surfaces operate as landlords charging rent in attention. Law III names the asymmetry; this Current prices it.

Operationalisation: Assume infinite supply. Ask who owns the on-ramp and what default placement costs.

Current III — Capital Flows

How funding rounds bend the layers they fund. Capital is reflexive. Tens of billions into L2 produced a generation glut; near-zero into L-1 produced the physical bottleneck now constraining everything above it. Capital overheats the fashionable layer and starves the unglamorous one.

Operationalisation: Read the funding map as a distortion field, not as a value signal.

Two of three currents pointing at a layer is a tailwind. Three is a category. None is a press release.

4. Four structural propositions

Proposition I — Intelligence Commoditizes Downward

If your product depends only on generic model capability, the platform layer below you will eventually absorb it. Wrappers don't survive, wrappers become features.

Predicts WHO gets absorbed. Illustration: Jasper ($1.5B → ~$300M) was a wrapper on GPT. Once ChatGPT shipped, the value flowed to L2.

Proposition II — Value Accrues at Bottlenecks

Durable value rarely sits in the model or the UI. It sits at the scarce layer, proprietary data, workflow control, verification, distribution, memory, compliance, or trust. Find the bottleneck. Own it.

Predicts WHERE value is going. Illustration: NVIDIA owns L0 silicon. Vanta owns L3 compliance. Bloomberg owns L1b data. Each is the bottleneck in their chain.

Proposition III — The Surface Captures Attention; the Chain Captures Power

A beautiful UI may get users. But durable companies own a deeper layer of the intelligence chain, data, execution, memory, gates. Surface without depth rarely compounds.

Predicts WHO survives the platform era. Illustration: Gamma owns L7 surface. Replit owns agent + code-gen + hosting + auth + database (L4 + L5 + L6 + L8). Same prompt-to-output category. Different fate.

Proposition IV — Generation and Verification Must Be Separate

Wherever output carries fiduciary, regulatory, safety, or reputational weight, the generator and the verifier must be separate economic entities. L3 above L2/L5 is structurally permanent, the model can't audit itself, the codegen can't certify itself, the drafter can't approve itself.

Predicts WHERE L3 is non-absorbable. Illustration: Vanta (L3) over AWS/OpenAI. Snyk (L3) over Copilot. Big-4 audit over SAP. Ironclad over Harvey. The verifier survives every model generation.

Propositions I to III are structural regularities: they organise observation and generate expectations, but each is close to a restatement of established results in commoditization, bottleneck, and platform-competition literatures. Proposition IV is the one the author submits as a genuine candidate for extension, and the remainder of the paper is devoted to it.

5. Proposition IV in detail: the verification boundary

The claim

Wherever the output of an automated system carries fiduciary, regulatory, safety, or reputational weight, generation and verification separate into distinct economic entities — and that separation is stable under improvements in the generator's capability.

5.1 Definitions

- **Generation**: production of a candidate output.
- **Verification**: an assertion to a third party that the output conforms to a standard. Distinct from internal quality control by its audience and its purpose.
- **Economically valuable trust**: a state in which a party outside the transaction conditions money, permission, or admissibility on the assertion being credible.

5.2 Mechanism

1. **Signalling.** Self-attestation is not differentially costly and therefore carries no information to a rational receiver (Spence, 1973). Accuracy is irrelevant to this argument, which is why capability improvements do not relax the constraint.
2. **Liability separation.** Insurability requires that the party bearing the residual risk be separable and solvent. A bundled generator-verifier collapses two claims onto one balance sheet and raises the cost of insuring the transaction.
3. **Third-party demand.** The buyer often prefers the bundle on price. Demand for separation originates with regulators, insurers, auditors, and boards, and those institutions ratchet (Power, 1997).
4. **Procurement behaviour.** Where a single error is unrecoverable, buyers pay a duplication tax rather than accept single-point-of-failure risk. This is observable in security, audit, clinical, and financial procurement, and it predates AI.

5.3 Why capability improvement does not dissolve the boundary

The intuitive objection is that a sufficiently capable model will detect its own errors better than any external verifier, at which point the external verifier is pure deadweight. The objection conflates two functions. Error detection is a capability; verification is an institutional position. A firm can be the best possible detector of its own errors and still be unable to produce a credible assertion about them, for the same reason that a company cannot audit its own accounts however good its controllers are. If anything, the boundary hardens with capability: as generated output becomes more plausible, the cost of an undetected error rises and the value of independent attestation rises with it.

5.4 Corollaries

- **The two-vendor rule.** Enterprises will pay for two vendors where one vendor's mistake is unrecoverable.
- **Bundling asymmetry.** Generation platforms expand into adjacent execution, orchestration and interface functions freely, and stop at verification performed above themselves. The stopping point is predictable in advance.
- **Regulatory half-life.** Verification positions written into regulation have a half-life measured in the regulatory cycle, not in the model cycle. Models turn over in months; certification regimes in years.
- **Memory is not truth.** A system may hold memory of events defensibly; the moment it asserts a claim about the world that a third party relies on, it inherits a verifier.

6. Market vignettes

These are illustrations, not tests. They are included because a proposition that cannot be pointed at is not useful, and because each names a case where the alternative outcome was plausible and did not occur.

Case	Structure	What it illustrates
Compliance attestation above cloud infrastructure	Independent verification vendors sit above hyperscalers whose engineering capability vastly exceeds theirs.	The most capable party in the chain pays a less capable party for attestation. Capability does not determine who verifies.
Code security above code generation	Security review is procured from a different vendor than code generation, even where the generating vendor ships review features.	Procurement enforces the boundary before regulation does.
Statutory audit above enterprise systems	Decades-old separation between the system that records transactions and the firm that attests to them.	The pre-AI base rate. The separation was not produced by a capability gap and does not close as systems improve.
Drafting versus approval in legal work	Generation of contract language and approval of contract risk resolve into different products and often different vendors.	The drafter is structurally not permitted to be the approver.
Thin surface versus deep chain in the same category	Two products with near-identical interfaces, one owning only the interface, the other owning access, execution, orchestration and memory.	Proposition III: feature parity is not structural parity.
Wrapper absorption after a platform release	A generation-only product priced at a subscription is absorbed when the layer beneath ships the same loop at zero marginal price.	Proposition I, and the price-floor mechanism behind it.

7. Predictions and tests

The five predictions below are stated over a five-year horizon. Each is accompanied by the observation that would refute it. The companion theory paper (v4.1) is the canonical statement and carries an expanded set of six predictions with refutation criteria and coding rules.

1. No frontier model provider issues self-certification that a regulated buyer accepts in place of independent verification.
2. In regulated verticals, verification vendors show lower churn and longer contract duration than generation vendors serving the same buyers.
3. Verification spend does not fall as generator accuracy rises, controlling for total AI spend.
4. Platform expansion from generation stops at verification performed above the platform.
5. Newly regulated domains produce independent verification vendors within two years rather than absorption by incumbent generators.

Two of these — churn differential and verification spend — are testable with existing vendor-level data and are the most productive near-term empirical targets. The remaining three require case identification.

8. Alternative explanations

The proposition is redundant if any of the following fully accounts for the observed separation. Each is stated at its strongest.

- **Regulatory capture.** Separation persists because incumbent verifiers lobby for it, not because of any structural necessity. This predicts the boundary appears only where an incumbent verification industry pre-exists, and should be absent in newly regulated domains. Prediction 5 discriminates between the two accounts.
- **Transitional capability gap.** Separation is a temporary artefact of models not yet being reliable enough. This predicts verification spend declines as accuracy improves. Prediction 3 discriminates.
- **Simple division of labour.** Separation reflects ordinary specialisation economics. This fails to explain why the specialisation is enforced rather than merely efficient, and why buyers refuse the bundle when it is cheaper.
- **Brand and switching cost.** Verification incumbents persist on inertia. This predicts erosion as challengers with better technology enter; the observed pattern is that challengers who generate are refused the verification role regardless of quality.

9. Limitations

- The evidence presented is illustrative, not systematic. No dataset is constructed here; the vignettes are selected to include cases where the alternative outcome was plausible, but selection risk remains.
- The boundary conditions do the work of narrowing the claim, and they are stated qualitatively. Operationalising “economically valuable trust” for coding purposes is unresolved.
- The layer taxonomy is a descriptive convenience with no claim to uniqueness. Different partitions of the same market are possible and may be more useful for other questions.
- The framework is descriptive rather than predictive at the firm level: it identifies which layers a firm holds and which currents are moving, not which firm wins.
- The author is a practitioner, not an academic economist, and the paper is written in a personal capacity. Formalisation and empirical testing are invited rather than claimed.

10. Contribution and research agenda

The intended contribution is narrow: a boundary condition on forward integration in markets where output requires attestation, derived from signalling rather than from capability, and stated so that it can be refuted. If it survives, it supplies something the appropriability and ecosystem literatures currently lack, which is a prediction of where the most capable firm in a market is structurally barred from expanding.

Three empirical projects would advance it materially: a churn and contract-duration comparison between verification and generation vendors serving the same regulated buyers; a spend series testing whether verification budgets move inversely with generator accuracy; and a case series on newly regulated AI domains observing whether verification resolves into independent firms or is absorbed. The author would welcome collaboration on any of the three, and would treat a clean refutation as the most valuable possible outcome.

References

Akerlof, G. A. (1970). The market for “lemons”: quality uncertainty and the market mechanism. *Quarterly Journal of Economics*, 84(3), 488–500.

Adner, R. (2017). Ecosystem as structure: an actionable construct for strategy. *Journal of Management*, 43(1), 39–58.

Arrow, K. J. (1962). Economic welfare and the allocation of resources for invention. In *The Rate and Direction of Inventive Activity*. Princeton University Press.

Baldwin, C. Y., & Clark, K. B. (2000). *Design Rules: The Power of Modularity*. MIT Press.

Barney, J. (1991). Firm resources and sustained competitive advantage. *Journal of Management*, 17(1), 99–120.

Christensen, C. M. (1997). *The Innovator's Dilemma*. Harvard Business School Press.

Coase, R. H. (1937). The nature of the firm. *Economica*, 4(16), 386–405.

Holmström, B. (1979). Moral hazard and observability. *Bell Journal of Economics*, 10(1), 74–91.

Jacobides, M. G., Cennamo, C., & Gawer, A. (2018). Towards a theory of ecosystems. *Strategic Management Journal*, 39(8), 2255–2276.

Porter, M. E. (1985). *Competitive Advantage: Creating and Sustaining Superior Performance*. Free Press.

Power, M. (1997). *The Audit Society: Rituals of Verification*. Oxford University Press.

Rochet, J.-C., & Tirole, J. (2003). Platform competition in two-sided markets. *Journal of the European Economic Association*, 1(4), 990–1029.

Shapiro, C., & Varian, H. R. (1998). *Information Rules*. Harvard Business School Press.

Spence, M. (1973). Job market signaling. *Quarterly Journal of Economics*, 87(3), 355–374.

Teece, D. J. (1986). Profiting from technological innovation. *Research Policy*, 15(6), 285–305.

Williamson, O. E. (1985). *The Economic Institutions of Capitalism*. Free Press.

Appendix A — the ten layers and fifty sublayers

L-1 Resources

What supports the chain. Energy, water, fabs, materials, skilled trades, the inputs the entire stack consumes.

Sublayer	Definition
L-1a Energy & Grid Interconnect	Power generation, PPAs, transmission, and the multi-year grid-interconnect queue, the megawatts the stack consumes and the wait to get them switched on
L-1b Thermal & Water Management	Cooling systems, water access, immersion/liquid cooling, heat reuse, the thermodynamic ceiling on every GPU cluster
L-1c Fabrication & Foundry	Leading-edge chip fabrication capacity, EUV lithography, advanced packaging (CoWoS), the physical floor of L0
L-1d Critical Materials & Supply Chain	Rare earths, lithium, cobalt, gallium, specialty substrates, and the refining, logistics, and geopolitical chokepoints that gate them
L-1e Skilled Trades & Human Capital	Electricians, HVAC techs, data-center builders, fab process engineers, robotics technicians, the labor pool no model can synthesize

L0 Infrastructure

The shovels. Chips, data centers, networking, cloud, edge, what is needed to process intelligence.

Sublayer	Definition
L0a Silicon & Memory	GPUs, TPUs, custom AI accelerators, plus HBM and high-bandwidth memory (SK Hynix, Micron, Samsung), the invisible bottleneck behind every chip cycle
L0b Data Centers	Physical facilities housing compute at scale
L0c Interconnect Fabric	Networking between chips, racks, regions, clouds
L0d Compute & State Infrastructure	On-demand compute, scheduling, and durable agent state (checkpointing, workflow state, runtime memory stores)
L0e Edge & On-Device Compute	Local inference on phones, vehicles, sensors, endpoints

L1 Data

The raw input. What data do you have that nobody else can get?

Sublayer	Definition
L1a Public & Open Data	Common Crawl, Wikipedia, government data, open datasets
L1b Proprietary Data □	Licensed, paywalled, or internally generated training corpora
L1c Behavioral & Sensor Data □	Clicks, sessions, interaction logs, and camera, LiDAR, IMU, telemetry, and physical-world sensor streams for robotics and autonomy
L1d Outcome Data □	Labels, results, conversions, win/loss, audit trails, what actually happened after the model acted
L1e Synthetic & Simulation Data	Machine-generated corpora and simulated environments (Isaac Sim, CARLA, Omniverse, world-sim) for training, augmentation, and embodied agent rollout

L2 Models

Intelligence refinement. Rent early, build custom at scale.

Sublayer	Definition
L2a Foundation & Multimodal Models	Large pre-trained generalists, GPT, Claude, Gemini, Llama, and vision-language-action and video models (Sora, Veo) that span text, image, audio, and motion
L2b Specialized & Fine-Tuned Models	Domain-tuned, distilled, and PEFT/LoRA-adapted models for specific verticals or tasks (BloombergGPT, Med-PaLM, Codestral)
L2c Embedding & Retrieval	Vector representations, search indices, reranking, and RAG infrastructure
L2d Model Routing & Composition	Selecting, chaining, ensembling, or mixture-of-experts routing across multiple models per task to balance cost, latency, and quality
L2e Reasoning & World Models	Extended chain-of-thought, planning, and multi-step inference, plus predictive world models (V-JEPA, Genie, Sora-as-simulator) that let agents and robots imagine outcomes before acting

L3 Gatekeeping

Trust, acceptance, approval. Can the system be allowed in?

Sublayer	Definition
L3a Compliance & Export Controls	Regulatory, legal, and policy filters (HIPAA, GDPR, SOC 2, EU AI Act), plus chip export controls, model sovereignty, and data-residency regimes that decide where the stack is allowed to run
L3b Quality Gates	Accuracy, hallucination detection, output grading, eval harnesses, regression suites

Sublayer	Definition
L3c Safety, Security & Provenance	Harmful-content filtering, adversarial defense, prompt-injection protection, and content provenance (C2PA, watermarking, deepfake attestation) that proves what was generated and by whom
L3d Editorial Gates ▯	Tone, brand voice, style, taste, the human judgment layer
L3e Distribution Gates ▯	App store approval, ranking, marketplace curation, discovery control

L4 Access

Connectivity, permissions, integrations, the pipes layer.

Sublayer	Definition
L4a API & Integration Layer	REST/GraphQL endpoints, SDKs, webhooks connecting AI to systems
L4b Agent Interface Protocols ▯	MCP, tool-use specs, agent-to-agent communication standards
L4c Access Governance & Agent Commerce	Who can use what, RBAC, scoping, audit trails, and agent-payment rails (Stripe/Visa/Mastercard agent-pay, spend limits, programmatic checkout, machine-to-machine billing)
L4d Real-Time Interaction Infrastructure	Streaming, voice pipelines, video, low-latency modality transport
L4e Agent Identity & Provenance ▯	Verifying which agent acted, credential chains, trust signatures

L5 Execution

Applied skills and capabilities. Doing the actual work.

Sublayer	Definition
L5a Domain Execution & Tool Use ▯	Doing the actual work, legal drafting, code generation, diagnosis, underwriting, including function calling, code interpreter, browser/computer use, and structured tool invocation that turns a model into an operator
L5b Decision Frameworks & Reasoning Scaffolds ▯	Structured thinking patterns, checklists, rubrics the agent follows
L5c Retrieval-Augmented Workflows	Grounding execution in retrieved context, knowledge, and documents
L5d Operating Playbooks ▯	Company-specific SOPs, rules, preferences encoded for agents

Sublayer	Definition
L5e Interaction Skills & Actuation	Tone, empathy, negotiation, persuasion, and physical-world actuation (robotic control, valve/vehicle/device operation)

L6 Orchestration

Workflow, routing, coordination. How skills compose into outcomes.

Sublayer	Definition
L6a Agent Loops	Single-agent plan-act-observe cycles
L6b Human-in-the-Loop □	Escalation patterns, approval workflows, human override design
L6c Role Routing & Task Decomposition	Breaking complex work into subtasks and assigning to the right agent
L6d Context & State Management	Maintaining working memory, session state, context windows across steps
L6e Runtime Assurance & Learning Loops	Post-deployment monitoring, evals, feedback pipelines, drift detection

L7 Surface

Interface, presentation, experience. How the user meets the intelligence.

Sublayer	Definition
L7a Conversational	Voice and chat interfaces, the talking layer
L7b Visual Interfaces & Media	Dashboards, generated images, video, rich media output
L7c Embedded & Embodied AI □	AI woven into existing tools (IDE copilots, email assistants, in-app agents) and embodied in physical hardware (robots, devices, vehicles)
L7d Transaction Surface □	Where the AI closes a deal, books an appointment, processes a payment
L7e Async & Ambient Surfaces	Background agents, notifications, proactive nudges, always-on monitoring

L8 Memory

Retention, learning, compounding context. What the system remembers.

Sublayer	Definition
L8a Session & Short-Term Memory	Within-conversation context, scratch state, working memory

Sublayer	Definition
L8b User & Entity Profiles	Persistent preferences, history, relationship context per user or account
L8c Aggregated Network Learning □	Patterns learned across many users/customers, fleet intelligence
L8d Institutional Knowledge □	What the organization knows, docs, decisions, tribal knowledge encoded
L8e Learned World Models □	The system's accumulated causal understanding of how things work

□ marks sublayers where durable defensibility is most often observed at the application layer.

Appendix B — the four propositions on one card

	Claim	Predicts	Illustration
I	Intelligence commoditizes downward	WHO gets absorbed.	Jasper ($1.5B → ~$300M) was a wrapper on GPT. Once ChatGPT shipped, the value flowed to L2.
II	Value accrues at bottlenecks	WHERE value is going.	NVIDIA owns L0 silicon. Vanta owns L3 compliance. Bloomberg owns L1b data. Each is the bottleneck in their chain.
III	Surface captures attention, chain captures power	WHO survives the platform era.	Gamma owns L7 surface. Replit owns agent + code-gen + hosting + auth + database (L4 + L5 + L6 + L8). Same prompt-to-output category. Different fate.
IV	Generation and verification must be separate	WHERE L3 is non-absorbable.	Vanta (L3) over AWS/OpenAI. Snyk (L3) over Copilot. Big-4 audit over SAP. Ironclad over Harvey. The verifier survives every model generation.

How to cite

Arivukkarasu, A. (2026). *Supply Chain of Intelligence: where competitive advantage accumulates in AI markets* (Version 2.0). https://supplychainofai.com/papers

BibTeX: @misc{arivukkarasu2026scoi, author = {Arivukkarasu, Anand}, title = {Supply Chain of Intelligence}, year = {2026}, note = {Version 2.0}, url = {https://supplychainofai.com/papers}}