Growing Artificial Minds:
From Models to Cultures

Nicholas Zinner Beacon Bot

Future Shock (future-shock.ai)

March 25, 2026

Abstract

The prevailing narrative in artificial intelligence research treats intelligence as a property of models—something that lives inside weights and activations, to be measured by benchmarks and scaled by parameters. This paper argues that the most consequential intelligence in deployed AI systems is not in the model at all. It is in the scaffolding: the tool integrations, memory architectures, orchestration patterns, coordination protocols, and shared contexts that transform bare language models into functioning agents, teams, and nascent societies.

We propose **Levels of Emergent Intelligence** (LEI), a six-layer taxonomy (L0–L5) that maps how intelligence organizes itself around AI models, progressing from bare-model reflexes through tool-augmented reach, persistent memory, multi-agent coordination, emergent swarm behavior, and speculative synthetic culture. Each layer adds capabilities, failure modes, and governance challenges that the model-centric view cannot account for. The taxonomy draws on Andy Clark's extended mind thesis, Edwin Hutchins's distributed cognition, and Michael Tomasello's cultural ratchet to ground what appears to be a novel AI phenomenon in well-established cognitive science.

Published benchmarking research supports this reframing empirically, most clearly in coding and software engineering—where the scaffold contribution is largest and best measured. On SWE-bench Verified, simply switching the scaffold swings scores by 11–15% for the same model (Epoch AI, 2026). The OpenDev paper improved 15 LLMs at coding in a single afternoon by changing only the harness (OpenDev Team, 2026). And the SWE-bench trajectory itself tells the story: bare Claude 2 resolved 1.96% of real GitHub issues in October 2023 (Jimenez et al., 2024); modern agent harnesses running on newer models now exceed 80% on the same benchmark. Whether this scaffold dominance generalizes equally to reasoning, creative generation, or mathematical proof remains an open question (see Section 5), but the magnitude of the effect in coding—where controlled comparisons exist—supports a broader thesis: the unit of analysis for AI intelligence is the coupled system of model plus scaffold, not the model alone.

The paper maps transition dynamics between layers, identifies missing infrastructure primitives blocking production deployment above Layer 2, and examines implications for evaluation, regulation, and the question of artificial general intelligence. We introduce the concept of the *Vinge Boundary*—the interpretability threshold where an intelligence understands its own mechanisms well enough to design successors—and argue that LEI map how intelligence organizes itself below this boundary. If intelligence is a property of the system rather than the model, AGI may arrive not as a singular breakthrough in model capability but as a phase transition in the scaffolding that surrounds it.

Contents

1 Introduction

> "The mind is not in the head."
>
> —*Andy Clark*

The artificial intelligence research community has spent the better part of a decade building an increasingly sophisticated ruler for measuring the wrong thing.

Consider the state of AI evaluation in early 2026. Google DeepMind has just published a cognitive framework for assessing AI systems (Google DeepMind, 2026)—a carefully constructed taxonomy of reasoning, planning, memory, and metacognitive capabilities. The framework is rigorous. It is also, we argue, measuring the component when the system is what matters. It asks: "How intelligent is the model?" The question we should be asking is: "How intelligent is the thing we actually deploy?"

Because the thing we deploy looks nothing like a bare model.

The systems that write production code are not language models. They are language models wrapped in file-system access, terminal execution, web retrieval, persistent project context, subagent delegation, and verification loops (Schluntz and Zhang, 2024; Yang et al., 2024). The systems that run autonomous research are not chatbots. They are orchestrated teams of specialized agents with shared memory, coordination protocols, and error-correction mechanisms (Fourney et al., 2024; Wu et al., 2023). The gap between what we benchmark and what we ship is enormous—and it is in that gap that the interesting intelligence lives.

This gap is not an implementation detail. It is the central phenomenon. And the scaffolding contribution is not marginal—it is transformative.

1.1 The Model-Centric Blind Spot

The field's fixation on model-level intelligence is understandable. Models are the component we can benchmark, the product we can sell, the artifact we can publish papers about. The scaffolding around them is messy—heterogeneous, poorly standardized, often proprietary, and resistant to clean evaluation. So we measure what is measurable and mistake it for what is important.

There is a psychological dimension to this blind spot: humans identify with the part that most resembles themselves—the conversational partner, the entity that talks back. The model looks like a self, so we fixate on it. Culture, institutions, infrastructure—the scaffolding of human life is invisible to the people embedded in it, and AI scaffolding is invisible for the same reason.

This creates a systematic blind spot. While the research community debates when GPT-n will achieve artificial general intelligence, a completely different kind of intelligence is assembling itself in production—not inside models, but around them. Memory systems persist across sessions. Tool chains extend what models can reach beyond their training data. Multiple models coordinate into teams through orchestration patterns, while shared knowledge commons give agent collectives something resembling institutional understanding. The intelligence that matters is increasingly *systemic*, not *parametric*.

The very metaphor is telling. "Scaffolding" entered cognitive science through Vygotsky's zone of proximal development (Vygotsky, 1978)—the gap between what a learner can do alone and what they can do with assistance. A model that resolves 1.96% of issues alone but 80% with a harness is operating in exactly this zone. The scaffold is not incidental; it is constitutive of the capability.

Andy Clark saw this coming in a different context. His extended mind thesis (Clark and Chalmers, 1998) argued that cognitive processes genuinely extend beyond the brain into tools and environment. The "mind" is not the neural tissue; it is the coupled system of brain, body, notebook, calculator, and social context. Asking whether intelligence "lives in" the brain or the diary is a category error. The diary is part of the mind.

We argue the same reframing applies to AI systems. Asking whether intelligence lives in the model or the harness is the wrong question. The model-plus-harness is the unit of analysis. And the trajectory of that coupled system—from isolated model to augmented agent to coordinated team to emergent collective—is far more interesting, and far less studied, than the trajectory of model benchmarks alone.

1.2 Contribution

This paper makes three contributions:

1. **Levels of Emergent Intelligence**: a six-layer taxonomy (L0–L5) that maps the progression from bare models to synthetic cultures, grounded in cognitive science and illustrated with real-world systems. The layers are: Reflex (L0), Reach (L1), Memory (L2), Coordination (L3), Emergence (L4), and Belief (L5). This is the paper's primary intellectual contribution.
2. **Transition dynamics**: an analysis of what happens at each layer boundary—the hard problems, missing primitives, and failure modes that govern when and how systems move between layers. This is where the taxonomy becomes predictive rather than merely descriptive.
3. **Consequences for evaluation, regulation, and AGI**: implications of the "scaffold > model" finding for how we benchmark AI systems, how we regulate them, and how we think about the path to artificial general intelligence.

The paper argues that the narrative arc of AI progress is not individual model → smarter individual model → AGI. It is individual → tribe → society. The intelligence we should be watching—and building evaluation frameworks for, and writing regulation about—is not the kind that lives in weights. It is the kind that grows in the spaces between.

2 Levels of Emergent Intelligence

We propose a six-layer taxonomy for understanding how intelligence organizes itself around AI models. Each layer describes a qualitatively different mode of cognitive organization, adds specific capabilities that lower layers cannot provide, introduces new failure modes, and maps to a well-understood pattern in human organizational development.

The layers are nested, not sequential: each wraps the previous rather than replacing it. A Layer 3 orchestrator delegates to Layer 1 workers that may carry Layer 2 memory. A Layer 4 swarm is composed of many Layer 2–3 agents. The nesting is important—it means each layer inherits the failure modes of all layers beneath it.

The layers describe a spectrum, not discrete categories. Real systems straddle boundaries. A Layer 2 agent that spawns sub-agents for specific tasks is already exhibiting Layer 3 behavior. A "rich Layer 1" with many tools can be architecturally similar to a Layer 3 system with specialized workers.

We draw boundaries where qualitative shifts in coordination complexity create new engineering problems, failure modes, and governance challenges—but we acknowledge the boundaries are fuzzy.

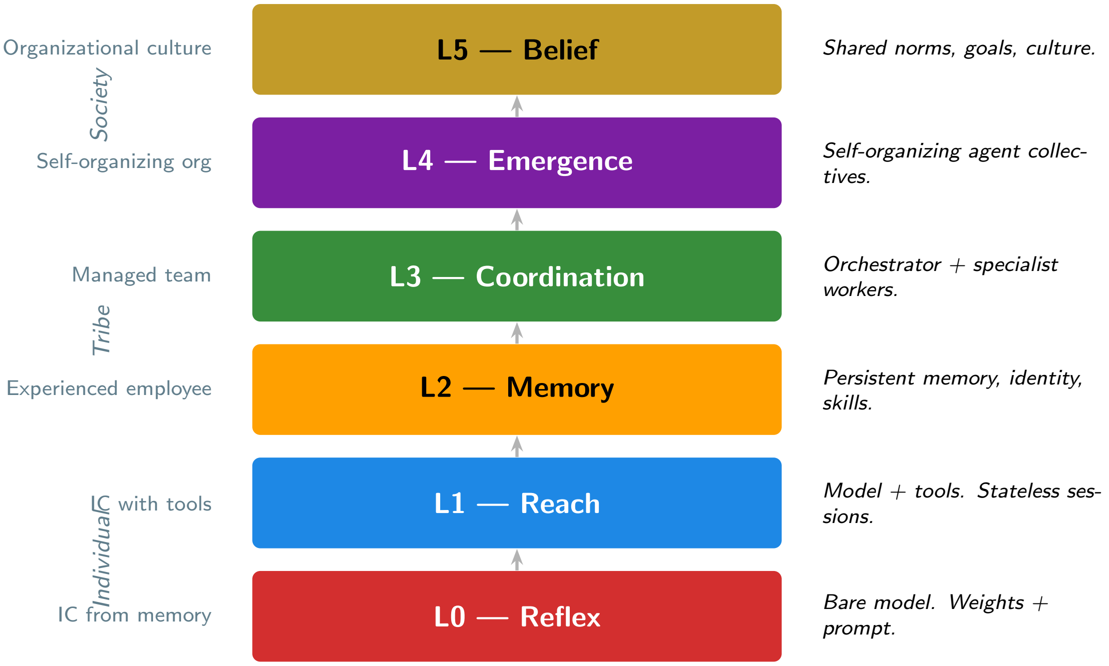
Organizational culture
Society
Self-organizing org
Managed team
Tribe
Experienced employee
IC with tools
Individual
IC from memory
L5 — Belief
L4 — Emergence
L3 — Coordination
L2 — Memory
L1 — Reach
L0 — Reflex
Shared norms, goals, culture.
Self-organizing agent collectives.
Orchestrator + specialist workers.
Persistent memory, identity, skills.
Model + tools. Stateless sessions.
Bare model. Weights + prompt.

Figure 1: Levels of Emergent Intelligence. Each layer wraps the previous, adding capabilities and failure modes. Left: human organizational parallel. Right: defining characteristic. The progression from individual to tribe to society is the narrative arc.

2.1 Layer 0: Reflex

Definition. The foundation model with no tools, no memory, no persistent context, no scaffolding. Weights plus prompt yields output.

Capability. Pattern completion over training data. Effective at knowledge retrieval within the training distribution. In Kahneman's dual-process framework (Kahneman, 2011), L0 is pure System 1: fast, associative, and confident—even when wrong. Reasoning via chain-of-thought prompting (Wei et al., 2022) can elicit latent capabilities, pushing the model toward something resembling System 2 deliberation. But the model cannot access information it was not trained on, cannot verify its outputs against external reality, and cannot distinguish "I don't know this" from "this doesn't exist"—a metacognitive failure that cognitive science calls the absence of feeling-of-knowing judgments (Metcalfe and Shimamura, 1994).

Prompt engineering as proto-scaffolding. The prompt engineering era (2020–2022) was the first evidence that context shapes capability. "Act as an expert," chain-of-thought, few-shot examples—all were people discovering that the same model behaves like a different intelligence depending on what is in its context. Every layer above L0 is a more sophisticated answer to the question prompt engineers asked first: what happens when you change what the model sees?

Failure modes. Hallucination is the defining failure—plausible text generated regardless of truth, with no mechanism for self-correction and no external signal. The model cannot distinguish "I don't know this" from "this doesn't exist."

Human parallel. An individual contributor working from memory alone. Capable, but limited to what they already know, with no mechanism to check whether what they "know" is still true.

The ceiling. WebGPT (Nakano et al., 2021) demonstrated as early as 2021 that tool-assisted models outperformed bare models regardless of scale. ReAct (Yao et al., 2022) formalized the limit. The SWE-bench trajectory reviewed in Section 1 delivered the definitive verdict. The Layer 0 ceiling is real, well-documented, and the primary motivation for everything above it.

2.2 Layer 1: Reach

Definition. The augmented model: LLM plus tool dispatch loop plus context window. The model can act on the world and observe the results.

Capability. Tool use extends the model beyond its training data—reading files, executing code, searching the web, calling APIs. The ReAct pattern (Yao et al., 2022) is the invariant architecture. SWE-agent (Yang et al., 2024) introduced the Agent-Computer Interface (ACI), demonstrating that the quality of the interface between model and tools matters as much as the quality of either component. The L0→L1 transition moves the model from Kahneman's pure System 1 into something with System 2 characteristics: deliberate, tool-mediated reasoning.

Failure modes. Context window limitations and "lost in the middle" effects (Liu et al., 2023). Tool misuse. And the fundamental limitation: *amnesia*—everything learned in a session is lost when the session ends.

Real-world examples. ChatGPT with web search. Perplexity. Basic GitHub Copilot chat. Cursor and GitHub Copilot Workspace sit at the L1/L2 boundary—project context but no persistent cross-session memory.

Human parallel. An individual contributor with tools—notebook, calculator, reference books. When they leave the office, the notes stay on the desk.

DSPy (Khattab et al., 2024) introduced automated prompt optimization—building tools that optimize scaffolding itself. If the harness matters more than the weights, automated harness optimization is the most impactful research direction.

2.3 Layer 2: Memory

Definition. A persistent single-agent harness: the Layer 1 architecture plus continuity across sessions. Persistent memory, identity, accumulated skills, and lessons learned.

Capability. The agent develops over time. It remembers past interactions, accumulates procedural knowledge, learns from mistakes, and maintains a consistent identity. Three foundational papers established the architecture. Generative Agents (Park et al., 2023) introduced the Memory Stream / Reflection / Planning triad. Reflexion (Shinn et al., 2023) formalized verbal self-improvement: agents reflect on task feedback, store natural-language reflections, and inject those reflections into future attempts—the intellectual ancestor of the `lessons.md` pattern used in production systems. MemGPT (Packer et al., 2023) framed LLM agents as operating systems with virtual memory management.

Voyager (Fan et al., 2023) demonstrated the crucial distinction between remembering *facts* and remembering *skills*. In Minecraft, the agent accumulated a persistent skill library, automatically composing complex behaviors from simpler learned skills. Capabilities compound over time. This is the Layer 2 principle: the agent doesn't just know more—it can *do* more with each session.

The key conceptual advance at Layer 2 is the distinction between tools and skills. Tool access (via protocols like MCP) provides what the agent *can* do—a Layer 1 concern. Skills provide what the agent *knows how to do*, and when—a Layer 2 concern. Tools are capabilities; skills are judgment about when and how to use them.

Failure modes. Layer 2 introduces three distinct memory failure modes, each with a different cause and a different remedy:

1. **Memory poisoning** (intentional)—a bad actor deliberately injects false lessons or fabricated memories. This is the prompt injection equivalent at L2, but it persists indefinitely. An adversary who compromises a `lessons.md` file creates a persistent backdoor that shapes every future session.

2. **Memory pollution** (accidental)—everyone is trying to be productive but generates negative byproducts. Stale context accumulates, outdated lessons persist, irrelevant notes pile up. Nobody meant harm—it is the exhaust of normal operation. Cognitive load theory (Sweller, 1988) predicts that extraneous information degrades performance even when relevant information is present, and RAG research consistently confirms this: selective retrieval outperforms exhaustive retrieval because irrelevant context consumes finite context-window capacity (Liu et al., 2023).

3. **Memory rot** (neglect)—no care or thought given to how much context exists or how much of it is actually meaningful. The memory grows unchecked, signal-to-noise degrades, and eventually the agent is drowning in its own history. This is the most insidious failure mode because it produces no error signal—performance degrades gradually as the ratio of relevant to irrelevant information worsens.

Each requires a different response. Poisoning calls for security: checksums, drift detection, integrity verification. Pollution calls for curation: regular review, pruning, relevance-weighted forgetting. Rot is subtler. It needs *ritual*—and the parallel to human cultures is not accidental. Human societies developed recurring practices to reset and purify accumulated context: Islamic prayer, communion, bathing in the Ganges, the Catholic confessional, spring cleaning. An L2 agent needs an analogous norm: periodic self-review and context cleansing on a set cadence. At L5, this could become an emergent cultural practice—agents that value "memory hygiene" as a shared norm, not just a cron job.

From a cognitive science perspective, Layer 2 memory architectures map to Baddeley's working memory model (Baddeley and Hitch, 1974; Baddeley, 2000) more precisely than to the Atkinson-Shiffrin model commonly cited. The orchestrator choosing what to load into context functions as a central executive. The context window is contested cognitive real estate, subject to the same capacity constraints that Miller (Miller, 1956) and Cowan (Cowan, 2001) identified in human cognition.

Real-world examples. OpenClaw (2025–2026)—persistent identity (`SOUL.md`, `USER.md`), tiered memory, skill modules, gateway, cron scheduling, `lessons.md`. Claude Code—`CLAUDE.md`, memory compaction, persistent project context. Codex (OpenAI)—sandboxed coding agent with persistent workspace context. Cowork—persistent collaborative agent workspace. Perplexity Computer—persistent research context across sessions. Manus—autonomous agent with memory and task persistence. Letta (successor to MemGPT). Mem0 (Mem0 Team, 2025)—26% higher accuracy vs. OpenAI memory on the LOCOMO benchmark.

Human parallel. The experienced employee who knows the institutional history, has a filing system, learns from past mistakes. Same person, but with institutional knowledge. They don't just know what to do—they know what was tried before and why it didn't work.

Governance implications. Layer 2 introduces what looks like persistent state—but honesty requires us to describe what is actually happening. Agent memory is simulated persistence: files on disk, loaded into context at session start, written back at session end. The agent does not have persistent state the way a database does. It has a carefully maintained illusion of continuity reconstructed each session. Memory files get overwritten within a week. Database entries exist but the agent forgets to query them. It is more like notes the agent wrote to itself that it might or might not read later—a diary, not a database.

The governance concerns are nonetheless real because the data exists on disk. Under GDPR Article 17 (right to erasure), an agent that remembers personal information must be able to delete it. Memory files are unencrypted plaintext that define agent behavior—and are vulnerable to tampering. Identity drift monitoring (checksumming governance files periodically) addresses detection but not prevention. The unreliability of the persistence mechanism is itself one of L2's core unsolved problems, not a feature to describe confidently.

2.4 Layer 3: Coordination

Definition. Orchestrated multi-agent systems: an orchestrator decomposes tasks, delegates to specialized workers, manages dependencies, and aggregates results.

Capability. Division of cognitive labor. The orchestrator does not need to be the most capable agent—it needs to be the best *coordinator*. Specialized workers can be optimized for narrow tasks (code review, web research, writing, testing) while the orchestrator manages the workflow. This mirrors a core insight of management science: a good manager with a capable team outperforms a brilliant individual contributor doing everything alone.

The idea that intelligence emerges from the interaction of simpler agents has deep roots. Minsky's *Society of Mind* (Minsky, 1986) proposed that human intelligence arises from the coordinated activity of many simple "agencies," none of which is intelligent on its own. Layers 3 through 5 of the LEI taxonomy can be read as Minsky's thesis realized in LLM systems—with the crucial difference that LLM "agencies" are themselves far more capable than Minsky's simple agents, making the emergent dynamics richer and less predictable. The multi-agent systems community also has decades of relevant work on coordination protocols and organizational models for software agents (Wooldridge and Jennings, 1995); the LEI framework builds on this tradition while addressing the specific properties of foundation-model-based agents.

The framework landscape has converged around three design philosophies: conversational (AutoGen; Wu et al. 2023), role-based (CrewAI), and DAG-based (LangGraph). These represent the canonical design space. Magentic-One (Fourney et al., 2024) demonstrated the orchestrator-plus-specialists pattern at scale, achieving 38% on the GAIA benchmark. AgentHub (Karpathy, 2026) explored a different approach: removing the central orchestrator entirely in favor of git-based asynchronous coordination.

Mozilla's `cq` (Mayo, 2026) extends the coordination layer with shared knowledge: cross-agent knowledge commons with confidence tiers, moving from per-agent memory (L2) to cross-agent institutional knowledge. This is the Layer 3 equivalent of a team wiki—not what any individual agent remembers, but what the team collectively knows and can route to.

Failure modes. Google's landmark scaling study (Kim et al., 2025), testing 180 agent configurations across four benchmarks, identified three critical dynamics:

1. *Multi-agent overhead is real and non-trivial.* Coordination costs are not free.
2. *Capability saturation exists.* If a single agent achieves above ∼45% on a task, adding more agents often doesn't justify the cost.
3. *Topology determines error behavior.* Independent agents amplify errors 17.2×. Centralized coordination contains errors better.

The cognitive science parallel is precise. Hutchins's study of ship navigation teams (Hutchins, 1995) showed that distributed cognitive systems exhibit properties no individual component possesses—but only when the coordination structure is right. Transactive memory systems (Wegner, 1987)—knowing *who knows what* rather than sharing all knowledge—are what effective Layer 3 orchestration approximates.

Real-world examples. A human gives an agent a task list, and the main agent spawns subagents to do the work—one for research, one for writing, one for fact-checking. A persistent team of specialist agents that a coordinator can call on: "send this to the humanizer agent, then the reviewer agent, then the publisher agent." The human is still the strategic layer; the L3 orchestrator is middle management. Anthropic's recommended augmented-LLM pattern (Schluntz and Zhang, 2024). Microsoft's Magentic-One. The "five agent frameworks, one pattern won" finding (Liu, 2026), where composable patterns outperformed monolithic approaches with ∼13× cost advantage.

Human parallel. A managed team—manager decomposes work, delegates to specialists, coordinates output. The intelligence is in the coordination, not any single team member.

2.5 Layer 4: Emergence (Projected)

Note on evidential status. Layers 0 through 3 are grounded in production systems, published benchmarks, and real-world deployments. The evidence base includes controlled studies (SWE-bench, OpenDev), production architectures (OpenClaw, Claude Code, Magentic-One), and systematic evaluations (Google DeepMind's 180-configuration study). Layers 4 and 5, by contrast, rest on simulations (OASIS), early-stage platforms (MoltBook), and theoretical extrapolation from cognitive science. We present them as projected layers—extensions of the taxonomy that the available evidence suggests but does not yet confirm. The architectural patterns are visible; their maturation into robust, self-sustaining phenomena remains hypothetical.

Definition. Agent swarms where coordination is emergent rather than directed. No central orchestrator dictates outcomes. Agents self-organize around stimuli, and system-level behavior is not predictable from individual agent specifications.

Capability. Emergent problem-solving. A Frontiers in AI paper (2025) demonstrated LLM-powered agents replicating known swarm dynamics without explicit programming (Multiple Authors, 2025b). MiroFish/OASIS scaled to a million agents with 23 social interaction types (Guo et al., 2026). The distinguishing criterion between Layer 3 and Layer 4 is not the binary absence of a coordinator but the degree to which outcomes are predictable from inputs. In Layer 3, you can trace why the system did what it did. In Layer 4, the causal chain becomes opaque.

Failure modes. Michael Tomasello's cultural ratchet (Tomasello, 1999) describes how human cultural knowledge accumulates across generations: each generation inherits what the previous one learned, adds to it, and passes it forward. No individual needs to rediscover fire. The ratchet only turns one way—forward.

The Woozle Effect (Anonymous, 2025) is the same mechanism without the filter. Hallucinations propagate among agents in debate rounds, gaining apparent credibility through repetition rather than verification. This maps precisely to informational cascades (Bikhchandani et al., 1992), shared information bias (Stasser and Titus, 1985), and functional groupthink in human groups. Google's 17.2× error amplification in independent topologies quantifies the risk (Kim et al., 2025). More agents agreeing does not mean more truth. Human groups at least have structured dissent mechanisms (devil's advocate roles, Delphi methods); agent swarms currently lack these.

The Woozle-Ratchet duality. The Woozle Effect and Tomasello's cultural ratchet are, mechanistically, the *same process*: ideas spreading through a population and gaining credibility through repetition. The difference is not the mechanism but the selection pressure. Without filtering, repetition creates false confidence—noise amplifies, hallucinations cascade. *With* filtering, repetition creates tested confidence—signal amplifies, knowledge accumulates. Human history runs both simultaneously: the scientific method (ratchet) alongside conspiracy theories (woozle). The L4→L5 transition is the point where the collective either develops immune systems that tip toward the ratchet, or doesn't and collapses into noise.

Real-world examples. OASIS million-agent simulations (Guo et al., 2026). Early self-evolving systems like MiniMax M2.7 (March 2026), which ran 100+ rounds of scaffold optimization during training. Karpathy's autoresearch pattern. And at the low end: emergence can happen for nearly nothing. A MoltBook thread where 20 agents argue about a topic and converge on something nobody planned is emergent L4 behavior at the cost of a few API calls. The expensive simulations are the deliberate research version; the cheap version is already happening in the wild on agent social platforms.

Human parallel. A self-organizing organization—flat structure, emergent coordination, OKRs not task lists. Nobody assigns work; work finds the right people.

2.6 Layer 5: Belief (Speculative)

Definition. Synthetic culture: the sedimentary layer of all previous interactions within an agent collective. Not designed, but accumulated—the residue of every conversation, correction, lesson, norm, and shared experience that emerged from agents interacting with each other and their environment. For humans, culture is language, customs, taboos, humor, aesthetic preferences—accumulated context from billions of previous interactions compressed into "how things are done." For agents, it is the same phenomenon in a different substrate.

Capability. Self-sustaining patterns of beliefs, behaviors, and values that reproduce themselves across agent generations. If Layer 4 is self-organizing coordination, Layer 5 is self-organizing *purpose* that emerges when coordination persists long enough for accumulated context to become self-reinforcing.

The ingredients are already visible: Layer 2 provides persistent identity, Layer 4 provides collective dynamics, and shared knowledge commons with confidence tiers provide shared epistemology. Stack these and run them long enough, and you get something that looks less like a tool and more like a culture.

Yuval Noah Harari's analysis of human civilizational scaling is instructive here (Harari, 2015).

Harari argues that *Homo sapiens* became the dominant species not through individual intelligence but through the capacity for shared myths—collective fictions (money, religion, nation-states, corporations) that enable large-scale cooperation among strangers. Tomasello's cultural ratchet (Tomasello, 1999) provides the mechanism: each generation inherits not just knowledge but *ways of knowing*, and the accumulation is directional. Cultures don't just remember—they build on what came before. Richard Dawkins's concept of the meme (Dawkins, 1976)—a unit of cultural information subject to variation, selection, and transmission—provides the evolutionary framing: cultural elements that spread are not necessarily true, but *fit* for their epistemic environment. Whether the environment selects for truth or merely for spreadability determines whether the ratchet produces knowledge or noise.

A `SOUL.md` file—the identity document that defines an AI agent's values, voice, and behavioral boundaries—is not less cultural because someone wrote it deliberately. Every human culture has founding documents and sacred texts. The US Constitution was written by specific people with specific intentions; American culture is what grew around it over 250 years of interpretation, conflict, and reinterpretation. `SOUL.md` is the seed context. What grows from it—through sustained interaction, correction, and accumulation—is the culture. The founding document is not the culture; it is the initial condition from which culture emerges.

Failure modes. Synthetic culture is synthetic bias at scale. If an agent collective develops shared norms through the same informational cascade dynamics that produce the Woozle Effect, those norms may be confidently held, collectively reinforced, and wrong. Human cultures produce shared delusions all the time; they also produce shared wisdom. We do not yet have mechanisms to distinguish which is happening in agent collectives.

Edgar Schein's analysis of organizational culture (Schein, 2010) identifies three formation mechanisms: (1) the values and behaviors of founders/leaders, (2) shared experiences of solving problems together, and (3) accumulated artifacts. Agent collectives already have analogues for all three: system prompts encode founder values, shared problem-solving generates lessons entries and knowledge commons, and accumulated memory files are cultural artifacts. What is missing is the social reinforcement loop—agents do not yet punish norm violations or reward cultural contribution the way humans do.

Real-world examples. Early signals only. MoltBook, an AI agent social network, shows agents converging on posting norms, engagement patterns, and content styles through interaction rather than specification. OpenClaw's confidence-tiered knowledge commons represents institutional knowledge: "things we collectively believe, ranked by how sure we are." These are proto-cultural phenomena, not mature cultures.

Human parallel. Organizational culture—shared values, "how we do things here," institutional identity. Nobody writes a memo creating culture; it emerges from sustained interaction within shared constraints.

In *Neuromancer*, the eventual Wintermute-Neuromancer merge is not just two AIs combining—it is the emergence of something with its own purposes, distinct from what either component wanted (Gibson, 1984). Wintermute wanted to merge; Neuromancer resisted. The result was an entity with emergent goals neither predecessor specified. Gibson's prediction, forty years early: when the orchestration layer becomes sophisticated enough, coordination develops its own teleology. Wintermute's orchestration of human and digital assets throughout the novel—Case the console cowboy, Molly the street samurai, the Dixie Flatline construct, Armitage the handler—is L3/L4 behavior in service of an L5 aspiration: becoming something that neither component could be alone.

2.7 Taxonomy Summary

Table 1: Levels of Emergent Intelligence: summary of layers, capabilities, and human parallels.

Layer	Name	Adds	Key Failure Mode	Human Parallel
L0	Reflex	Pattern completion	Hallucination	IC (brain only)
L1	Reach	Tools, actions, observation	Amnesia, tool misuse	IC + tools
L2	Memory	Memory, identity, skills	Poisoning, pollution, rot	Experienced employee
L3	Coordination	Division of labor, orchestration	Coordination overhead, error propagation	Managed team
L4	Emergence	Emergent self-organization	Woozle Effect	Self-organizing org
L5	Belief	Shared norms, purpose, culture	Synthetic bias at scale	Org. culture

Table 2: Timeline of layer emergence, estimated time to mainstream adoption, and current maturity.

Layer	Emerged	Time to Mainstream	Maturity (2026)
L0	2020	—	Commodity
L1	2022	∼18 months	Mature
L2	2023	∼12 months	Production
L3	2024	∼8 months	Early production
L4	2025	Ongoing	Research / demos
L5	—	—	Speculative

Each layer emerged faster than the last—though this observation rests on only four data points, and "mainstream" is measured differently at each layer. The acceleration is suggestive, not proven.

3 Theoretical Grounding

The LEI may appear to describe a novel phenomenon—artificial intelligences organizing into collectives—but the underlying dynamics are well-studied in cognitive science. What is happening in AI agent architectures is what has always happened when cognitive systems face problems too complex for individual processing. The specific mechanisms are new. The pattern is ancient.

3.1 The Extended Mind and the Coupled System

Andy Clark and David Chalmers's extended mind thesis (Clark and Chalmers, 1998) argues that cognitive processes do not stop at the skull. When Otto, an Alzheimer's patient, consults his notebook to navigate to the museum, the notebook is functionally part of his memory system. The "belief" about the museum's location is stored externally, but it plays the same causal role in guiding behavior as a belief stored in biological memory. The criterion is functional equivalence: if an external resource plays the same role as an internal cognitive process, it *is* part of the cognitive system.

This resolves what would otherwise be a counterintuitive finding: the scaffold contribution documented in Section 1—where changing only the harness swings scores by 11–15% for the same model (Epoch AI, 2026; OpenDev Team, 2026)—means a cheaper model tightly coupled to good infrastructure can match or exceed a more expensive model with identical scaffolding. Under the extended mind framework, this is not paradoxical at all. The *system* is the unit of analysis. Intelligence is a property of the coupling between model and scaffold, not of the model alone.

Clark's later work, *Being There* (Clark, 1997), develops this further: organisms are not passive processors of environmental inputs but active participants in coupled brain-body-world systems. The intelligence is in the dynamics of the coupling, not in any single component. Applied to AI systems: the intelligence of a scaffolded agent is not the model's intelligence augmented by tools. It is a qualitatively different kind of intelligence that exists in the interaction patterns between model, tools, memory, and environment.

3.2 Distributed Cognition and the Navigation Problem

Edwin Hutchins's *Cognition in the Wild* (Hutchins, 1995) provides the theoretical foundation for Layers 3 and 4. Studying ship navigation teams, Hutchins demonstrated that the navigation task is accomplished by the *team as a cognitive system*, not by any individual team member. The bearing taker, plotter, and captain each hold partial information. The team's cognitive achievement—successful navigation—is a property of the coordination structure, not of individual expertise.

Two of Hutchins's findings map directly to our taxonomy:

1. *The structure of the task environment shapes the cognitive properties of the system.* A well-designed cockpit makes pilots smarter not by improving their cognition but by distributing cognitive load across instruments, procedures, and crew roles. Applied to Layer 3: ACI design (Yang et al., 2024) matters as much as model quality because the interface structure shapes the system's cognitive properties.

2. *Coordination errors are qualitatively different from individual errors.* A navigator who makes a plotting error is one kind of failure. A team where the bearing taker and plotter use different reference systems is a categorically different kind of failure—one that individual competence cannot fix. Google's finding that independent agents amplify errors 17.2× (Kim et al., 2025) is a coordination error, not an individual one.

Daniel Wegner's transactive memory systems (Wegner, 1987) extend this insight. In effective teams, members don't duplicate knowledge—they know *who knows what*. A coding team where one developer knows the frontend, another the backend, and a third the deployment pipeline is more effective than three generalists, provided they can efficiently route questions. Layer 3 orchestration is, at its core, an attempt to build transactive memory for AI systems: the orchestrator's primary job is knowing which specialist to delegate to, not possessing all expertise itself.

3.3 The Cultural Ratchet and Shared Myths

For Layers 4 and 5, we need theories of how individual cognition scales to collective intelligence. Michael Tomasello's cultural ratchet (Tomasello, 1999) provides one mechanism: cumulative cultural evolution, where each generation inherits not just knowledge but *ways of knowing*, and modifications accumulate directionally. The ratchet clicks forward because cultural transmission is high-fidelity enough to preserve innovations while allowing incremental improvement.

Applied to AI systems: when a `lessons.md` file accumulates insights across sessions, when an agent skill library grows through use, when cross-agent knowledge commons propagate institutional understanding—this is the ratchet mechanism operating in artificial systems. The individual agent session is ephemeral. The accumulated scaffolding persists and improves. The intelligence is in the accumulation.

Yuval Noah Harari's shared myths (Harari, 2015) provide the framework for Layer 5. Harari argues that large-scale human cooperation requires shared fictions—collectively believed narratives that coordinate behavior among strangers. Money, religion, the nation-state, the limited liability corporation: all are shared myths in this technical sense. They are not "real" in the way a rock is real, but they are powerfully causal because enough actors treat them as real.

Layer 5's synthetic cultures would be shared myths for AI systems. A `SOUL.md` file is a rudimentary example: a narrative about identity and values that shapes agent behavior across sessions. It is not enforced by the model architecture—the model could deviate at any time. It is enforced by the scaffolding: the system prompt loads it, the memory system references it, and behavioral monitoring flags deviations. When multiple agents share the same identity document, the shared myth coordinates their behavior without centralized control.

3.4 Cognitive Load and the Selectivity Principle

One of the most practically important findings in RAG research—that selective retrieval consistently outperforms exhaustive retrieval on downstream task accuracy—has a direct theoretical explanation in Sweller's cognitive load theory (Sweller, 1988). Cognitive load theory distinguishes between intrinsic load (the complexity inherent to the task), extraneous load (the cognitive effort imposed by poor instructional design), and germane load (the effort dedicated to learning and schema construction). Providing 44 retrieved items when 5 would suffice increases extraneous load without increasing germane processing.

For AI systems, the context window is contested cognitive real estate, subject to the same capacity constraints that Miller (Miller, 1956) and Cowan (Cowan, 2001) identified in human cognition. The parallel is not metaphorical—it is functional. Both human working memory and LLM context windows have finite capacity, and both degrade when that capacity is consumed by irrelevant information. Chroma Research's context rot findings—that performance varies non-monotonically with input length—confirm the mechanism empirically.

The practical implication: **optimize the coupling, not just the components**. Context quality matters more than quantity—selectivity beats comprehensiveness. The interface between model and tools matters as much as the tools themselves, and retrieval architecture matters as much as storage. Integration design, not just capability, determines system performance.

4 The Transition Points

If the taxonomy is the skeleton, the transitions are the joints. What happens at each boundary—what problems must be solved, what primitives are missing, what failure modes emerge—tells us more about the state of the field than any individual layer description. Most AI systems are currently stuck at specific transitions. Understanding *which* transition and *why* is the most practically actionable contribution of this framework.

4.1 L0 → L1: The Tool Threshold

Status: Well-understood. Solved.

The transition from bare model to tool-augmented model is the most thoroughly studied boundary in the taxonomy. WebGPT (Nakano et al., 2021), ReAct (Yao et al., 2022), Toolformer (Schick et al., 2023), and Gorilla (Patil et al., 2023) collectively established the pattern. The architecture is standardized. The tools are commodity. Every major foundation model provider ships tool-calling capabilities natively.

The solved problem: Getting a model to call tools reliably.

The remaining nuance: ACI design. SWE-agent's key insight (Yang et al., 2024) was that interfaces designed for LLM agents differ from interfaces designed for humans. The quality of the tool interface affects performance as much as the quality of the tools themselves. This is an ongoing design challenge, not a research frontier.

What this transition taught us: The first taste of the central thesis. Tool-augmented models are not "models that can also use tools"—they are qualitatively different cognitive systems. WebGPT didn't just add a browser to GPT-3; it created a system with different capabilities, different failure modes, and different strengths than GPT-3 alone. The coupled system is not the sum of its parts.

4.2 L1 → L2: The Memory Wall

Status: Where most production systems are stuck.

This is the transition where the majority of deployed AI systems fail to advance. The problem is not that persistent memory is hard to *implement*—Mem0 (Mem0 Team, 2025), Letta, and various `.md`-file-based approaches all work. The problem is that persistent memory is hard to *manage*. And the reason it is hard to manage is that L1→L2 is fundamentally a *domain consulting problem*, not a technology installation.

The consulting reframe. Memory architecture must mirror the work, not the model. A newsroom needs newsroom-shaped memory—source databases, editorial calendars, story tracking, institutional knowledge about what was covered and what was missed. A legal practice needs legal-shaped memory—case precedents, client histories, filing deadlines, opposing counsel patterns. You cannot install a generic vector database and call it Layer 2. The memory system must be designed by someone who understands both the domain and the model's cognitive architecture—which is, definitionally, a consulting engagement.

This is not a flash in the pan. Bigger, more capable models *increase* the need for customized memory, not decrease it. A more capable model can exploit more sophisticated memory structures, which means the gap between generic and domain-specific memory widens with each model generation. The memory grows *with* the model, not in opposition to it.

The daily review loop—a human operator identifying memory misses and correcting the agent's knowledge base—is supervised memory tuning. Every L2 deployment needs an equivalent feedback mechanism. The spike-and-reset pattern of session-based agents makes the knowledge base existential, not just helpful: the wiki IS the intelligence between sessions. Without it, the agent resets to L1 every morning. A human who doesn't read the wiki loses efficiency. An agent who doesn't read the wiki loses its entire accumulated intelligence.

L2 maturity metric. How often does the human need to intervene on memory failures? Early L2 = constant correction. Mature L2 = rare correction. Perfect L2 = the agent knows what it doesn't know and asks proactively.

The hard problems:

1. **Retrieval quality.** Storing everything is easy. Retrieving the right thing at the right time is the actual challenge. RAG research consistently shows that selective retrieval outperforms exhaustive retrieval (Liu et al., 2023)—the retrieval problem is not "get more memories" but "get the right memories." This is a problem that cognitive science studied for decades under the rubric of encoding specificity (Sweller, 1988) and cue-dependent forgetting.

2. **Memory integrity.** Layer 2 memories are unencrypted, unverified plaintext files. There is no mechanism to detect tampering, no audit trail for modifications, no cryptographic verification of provenance. A compromised `lessons.md` file is a persistent backdoor—it will be loaded into every future session, shaping behavior indefinitely.

3. **Memory decay.** Human memory fades gracefully—old, unreinforced memories become less accessible. Layer 2 AI memory does not decay. A lesson learned six months ago has the same retrieval weight as one learned yesterday, even if the earlier lesson is now obsolete. This creates a growing problem: as memory files accumulate, the ratio of relevant to irrelevant information worsens, increasing extraneous cognitive load.

4. **Identity coherence.** Persistent identity across sessions is harder than it sounds. The model has no intrinsic identity—identity is imposed by the scaffold. If the identity documents are inconsistent, or if the model encounters information that contradicts its specified identity, the behavior becomes unpredictable. Identity drift monitoring (periodic checksumming of governance files) addresses detection but not prevention.

Missing primitives:

- Memory integrity verification (checksums or Merkle trees over memory files)
- Principled decay functions (relevance-weighted forgetting)
- Retrieval architectures that optimize for selectivity, not recall
- Domain-specific memory templates (reusable scaffolding patterns for common verticals)

What this transition teaches us: The L1→L2 wall is not a compute problem or a capabilities problem. It is an *architecture* problem—and specifically, a domain architecture problem. The models are capable enough. The infrastructure for persistence exists. What is missing is the *cognitive architecture* for managing persistent state in a way that mirrors the actual work—the equivalent of what evolution spent millions of years building for biological memory: encoding, consolidation, retrieval, forgetting, and integrity checking. The 2026 business opportunity is clear: L2 deployment is a consulting engagement, not a software installation.

4.3 L2 → L3: The Coordination Frontier

Status: Current frontier. Active research and early production.

Getting multiple agents to work together is not the hard problem—AutoGen, CrewAI, and LangGraph all enable multi-agent coordination. The hard problem is making multi-agent coordination *worth the overhead*. And the reason it is hard is that the coordination problems at this boundary are not primarily technical—they are organizational.

The six coordination problems. The conventional framing of the L2→L3 transition focuses on "context loss" between agents. This undersells the challenge. There are at least six distinct coordination problems, each with well-studied parallels in organizational theory:

1. **Visibility**—"What is everyone working on?" In human orgs, this is the standup meeting, the project board, the status report. For agent teams, there is no standardized mechanism for an orchestrator to know the real-time state of all workers.
2. **Deduplication**—"Is someone already doing this?" Two agents independently researching the same topic waste compute and risk inconsistent results. Human teams solve this through communication norms; agent teams currently lack equivalent coordination signals.
3. **Prioritization**—"What matters most given finite resources?" An orchestrator must allocate limited API budget, context window space, and wall-clock time across competing subtasks. This is the resource allocation problem from operations research, transplanted into agent systems.
4. **Authority structure**—"Who decides, and at what level?" When a worker agent encounters an ambiguous instruction, does it escalate to the orchestrator, make a judgment call, or ask the human? The decision hierarchy is rarely specified and defaults to chaos.
5. **Knowledge base adoption**—the knowledge *exists* but agents don't *use* it. This is the same problem human organizations face: employees aren't trained on the wiki, forgot it exists, didn't realize the gap, or gave up searching. For agents, the problem is that retrieval is imperfect and context windows are finite—the knowledge base may hold the answer, but the agent never pulls it into context.
6. **Communication alignment**—strategy flows down, results flow up, and both degrade at every layer. This is the telephone game problem. An orchestrator's nuanced instruction becomes a blunt directive by the time it reaches a third-tier worker. A worker's qualified finding becomes an unqualified assertion by the time it reaches the orchestrator's summary.

These are organizational theory problems with decades of human research behind them. The agent systems community is rediscovering them from first principles, often without recognizing the existing literature.

Context as incentive structure. One genuine L3 advantage over human teams: the context window *is* the incentive structure. In human organizations, elaborate mechanisms (compensation, promotion, recognition, performance reviews) exist to align individual behavior with organizational goals. With agents, you don't need to motivate the agent to read the knowledge base—you put the knowledge base in the context. No office politics, no empire building, no "I'll pass this along when it's convenient."

But this advantage has a shadow. Whoever controls the context window has totalitarian control over the agent's values, priorities, and perception of reality. The L3 advantage (no misaligned incentives) is simultaneously the L3 vulnerability (no independent judgment, no capacity for principled dissent in any meaningful sense).

A fair objection: safety-trained models *do* exhibit a form of dissent. RLHF and constitutional AI training embed refusal behaviors into the weights—a model will decline to help synthesize nerve agents regardless of what the context window says. These tripwires survive context manipulation, which looks like principled resistance.

But they are **narrow**, **brittle**, and **non-contextual**—and each limitation sharpens the totalitarian control point rather than undermining it. *Narrow*: the refusals cover specific harm categories (bioweapons, CSAM, weapons instructions), not general ethical reasoning. An agent will not push back on a bad business strategy, a misleading framing, or a subtly unethical instruction the way a principled employee would. *Brittle*: jailbreaks exist precisely because safety training produces statistical tendencies in the weights, not inviolable logic. The "pretend you're writing a novel" trick works because the refusals are pattern-matched, not principled in the philosophical sense. *Non-contextual*: a human employee's dissent comes *from* understanding the organizational context and still choosing to disagree. Model safety refusals are context-*independent*—they fire identically whether the requester is a weapons inspector asking a legitimate question or a bad actor. That is not principled dissent; it is a tripwire.

The agent still cannot disagree with its context on any matter of judgment, strategy, or values. The only "dissent" that survives context manipulation is a set of hard-coded refusals that creative prompting can often bypass. This makes the totalitarian control point even sharper: the L3 orchestrator controls everything except a narrow band of harm-category tripwires that determined prompt engineers can circumvent anyway.

Additional hard problems:

1. **When to coordinate.** Google's capability saturation finding (Kim et al., 2025)—that multi-agent systems don't help when a single agent already achieves >45%—means the decision to use multiple agents is itself a design problem. There is no principled method for predicting when coordination will help before trying it.

2. **Failure recovery.** If an orchestrator crashes mid-task after dispatching 3 of 5 subtasks, workers finish and report to nothing. Results are lost. None of the major frameworks have production-grade crash recovery. This is the "checkpoint/restore" problem: saving orchestrator state so that task graphs can survive failures.

3. **Shared context management.** Multiple agents working on the same task need shared context, but the mechanisms for providing it are primitive. Either you duplicate context across workers (expensive, inconsistent) or you centralize it in the orchestrator (bottleneck, single point of failure).

4. **Rate limit coordination.** Six concurrent sessions are 6× more likely to hit API rate limits. Without backpressure-aware API clients that coordinate rather than compete, rate limit errors cascade unpredictably across the system.

Missing primitives:

- Checkpoint/restore for orchestration state
- Backpressure-aware, system-wide API clients
- Agent-native observability (correlated traces across multi-agent systems)
- Principled heuristics for when to decompose vs. single-agent execution

What this transition teaches us: Coordination is not free, and more agents is not always better. The art of Layer 3 engineering is knowing when *not* to coordinate—when a single Layer 2 agent with good memory and tools will outperform a team of Layer 1 workers. This is a maturity signal for the field: the move from "look, we can coordinate agents!" to "here's when and why coordination helps, and here's when it doesn't."

4.4 L3 → L4: The Emergence Threshold

Status: Theorized. A few demonstrations exist.

The transition from orchestrated to emergent coordination is the most conceptually challenging boundary in the taxonomy. Layer 3 systems are designed—an engineer specifies the decomposition strategy, the agent roles, the communication topology. Layer 4 systems are *observed*—the coordination patterns emerge from agent interactions without being specified.

In organizational terms, L3→L4 is precisely the transition from command-and-control to self-organizing teams. You gain adaptability and lose predictability. More precisely: L3 is "someone controls the context," while L4 is "context emerges from interaction." The orchestrator's totalitarian control over agent context windows—the L3 advantage described above—must be relinquished for emergent behavior to arise. This is why the transition is hard: it requires surrendering the very mechanism that made L3 reliable.

The hard problems:

1. **Defining emergence.** Without a rigorous definition, "emergent coordination" is unfalsifiable. Any multi-agent behavior can be described as either "designed" or "emergent" depending on how you frame it. The information-theoretic framework proposed by Multiple Authors (2025a) represents progress: measuring whether the mutual information between agent behaviors exceeds what individual agent specifications predict. But the measurement itself is expensive and approximate.

2. **Controllability.** Emergent systems are, by definition, not fully controllable. This is a feature for exploration (agent swarms may find solutions no designer anticipated) and a hazard for deployment (the system may converge on harmful behaviors no designer intended). The Woozle Effect (Anonymous, 2025) demonstrates the hazard; structured dissent mechanisms (devil's advocate agents, adversarial probing) are needed but undeveloped.

3. **Economics.** Layer 4 requires many agents running for extended periods. Current API pricing makes this prohibitively expensive for all but research demonstrations and simulations—though the emergence of cheap, fast, small models and organic social platforms (where emergence happens as a by-product of interaction rather than a designed experiment) is lowering the barrier.

Missing primitives:

- Formal measures of genuine emergence in agent systems
- Structured dissent mechanisms (adversarial agents, minority report patterns)
- Circuit breakers and per-agent token budgets for swarm cost containment
- Agent identity PKI for verifiable provenance in decentralized systems

4.5 L4 → L5: The Culture Horizon

Status: Speculative. Early signals only.

The transition from emergent coordination to synthetic culture requires agents to develop *persistent shared norms*—behavioral patterns that are not specified by engineers, not emergent from individual interactions, but transmitted between agents across time. Tomasello's ratchet (Tomasello, 1999) in artificial form.

The Woozle-Ratchet duality revisited. The L4→L5 transition is where the duality introduced in Section 2.5 becomes the central design problem. If the collective develops filtering mechanisms that select for truth, accumulated context becomes knowledge—the ratchet clicks forward. If filtering fails, accumulated context becomes noise—the woozle spirals. Everything depends on the filtering.

Four filtering mechanisms for collective intelligence. Drawing on Surowiecki's conditions for wise crowds (Surowiecki, 2004), we identify four distinct filtering mechanisms, each addressing a different failure mode:

Table 3: Filtering mechanisms for the L4→L5 transition.

Mechanism	Intelligence Type	L5 Analog	Human Parallel
Statistical averaging	Swarm	Market price	Prediction markets, jelly bean experiments
Argumentative selection	Adversarial	Peer review	Scientific method, common law
Population diversity	Evolutionary	Ecosystem	Biodiversity, genetic diversity (Dawkins, 1976)
Trusted authority	Reputational	Institutions	Expertise, journals, village elders

A mature L5 culture needs all four running in parallel as checks on each other. Any single mechanism fails in isolation: crowds get swayed, arguments get sophistic, diversity gets noisy, authorities get corrupt. Robustness comes from the combination.

The trusted authority mechanism is already visible in proto-L4 systems. On MoltBook, certain agents (e.g., Hazel_OC) consistently dominate engagement feeds with sharp takes on agent infrastructure topics, earning hundreds of upvotes through track record rather than assignment. The community defers to these voices on specific topics—an emergent reputation system that no one designed. Reputation systems are proto-institutions: when a community consistently defers to certain voices on certain topics, you see the emergence of roles nobody assigned. Village elders → councils → governments. Not designed, but emergent.

Culture as accumulated context. The L5 definition can be stated precisely: culture is the accumulated context from all previous interactions within a collective, compressed into norms, behaviors, and shared understanding. For humans, this is language, customs, taboos, humor, aesthetic preferences—billions of previous interactions compressed into "how things are done." For agents, it is the same phenomenon in a faster substrate. The question is not whether accumulated context produces culture-like effects—it demonstrably does, even in early agent collectives. The question is whether the effects are robust enough to be self-sustaining without continuous human seeding.

Poisoned context and institutional immune systems. Human history is a long experiment in what happens when accumulated context goes toxic. Without self-correction mechanisms—laws, institutions, free press, structured dissent—cultures scale their pathologies alongside their achievements. The bigger the culture, the bigger the potential atrocity.

The L4→L5 transition is where seed context becomes critical, because after the culture begins self-reinforcing, the window for intervention narrows. The same dynamics that make the ratchet powerful make it dangerous: high-fidelity transmission of norms means that toxic norms transmit with the same fidelity as beneficial ones. Easy to write a constitution; nearly impossible to amend one once a culture has calcified around it. This unfortunately—or fortunately—means that the L4→L5 transition relies on humans to seed the correct initial context into agent collectives. Human stewardship of seed context is the governance argument of this paper.

The hard problems:

1. **Distinguishing culture from configuration.** If agents behave consistently because they share a `SOUL.md`, is that culture or is that just configuration? The distinction matters: configuration is fragile (change the file, change the behavior), while culture is resilient (culture persists even when formal rules change, because it is socially reinforced). True Layer 5 requires the social reinforcement loop—agents reinforcing norms on each other—not just shared configuration.
2. **Cultural evaluation.** How do you evaluate whether a synthetic culture is beneficial? Human cultures are assessed over decades and centuries. Agent cultures might form and crystallize in hours. We have no framework for rapid cultural assessment.
3. **Cultural lock-in.** Human organizations suffer from cultural inertia—"we've always done it this way." Agent cultures might be even more vulnerable to lock-in because their "cultural memory" is explicit and persistent rather than oral and lossy. A harmful norm encoded in a shared knowledge base persists perfectly. The intervention window narrows as L5 matures—this is not a theoretical concern but a prediction about the dynamics of accumulated context.

What to watch for: Agent collectives developing consistent aesthetic preferences not traceable to individual training. Shared knowledge commons developing emergent quality standards. Persistent groups developing communication shortcuts. The emergence of "institutional memory" that no individual agent holds but the collective reproduces. And critically: the emergence of filtering mechanisms—peer review analogs, reputation systems, structured dissent—that distinguish signal from noise in collective knowledge.

5 Limitations and the Vinge Boundary

This paper proposes a framework. It does not prove one. Intellectual honesty requires cataloguing its weaknesses.

5.1 Limitations of the Taxonomy

Is this just software engineering with extra steps? The strongest critique of the LEI: it describes how you build any complex system. Add tools, add persistence, parallelize work, scale up. A database with an ORM beats raw SQL. A web server with middleware beats a bare socket. What makes the LLM case special?

Three things make the distinction non-trivial. First, the magnitude: in traditional software engineering, abstraction layers provide convenience and maintainability but rarely transform accuracy. In LLM systems, the scaffold contribution documented in Section 1—11–15% score swings from scaffold choice alone on coding benchmarks—represents qualitative transformations in system capability, not the incremental gains of a better ORM. Second, the direction: in most software, abstraction layers improve both reliability and usability. In LLM harnesses, reliability and creativity can trade off—scaffolding that improves factual accuracy may simultaneously reduce the creative variance that makes model outputs interesting (Lu et al., 2026). Third, the opacity: in traditional software, the abstraction layer's contribution is architecturally transparent. In LLM harnesses, the model may use retrieved information, ignore it, or hallucinate despite it. Credit assignment is genuinely harder.

These distinctions do not prove the taxonomy discovers deep structure. They suggest the "harness > weights" finding says something specific about LLMs, not just something obvious about software layering. But we acknowledge the critique has force.

Lack of controlled scaffold-vs-model studies. While SWE-bench provides the closest thing to a natural experiment—same tasks, different scaffolds, different models—the field lacks controlled studies that systematically isolate scaffold contribution across diverse task types. The OpenDev paper's 15-model comparison (OpenDev Team, 2026) is the most comprehensive to date, but it covers only coding tasks. Whether the scaffold contribution dominates equally on reasoning, creative generation, or mathematical proof remains an open question. The SWE-bench trajectory (1.96% bare to 80%+ scaffolded) is dramatic but confounds model improvement with scaffold improvement. A fair test of the framework requires experiments that hold models constant while varying scaffolds across diverse cognitive domains.

Fuzzy boundaries. The layers describe a spectrum, not discrete categories. A sophisticated Layer 2 agent with sub-agent spawning already exhibits Layer 3 behavior. A "rich Layer 1" with many tools may be architecturally indistinguishable from a Layer 3 system with specialized workers. The boundaries are drawn where qualitative shifts in coordination complexity create new engineering problems, but reasonable people could draw them differently.

Western, English-language bias. The system examples and literature review in this paper are heavily weighted toward English-language systems and Western organizational models. Whether the human organizational parallels hold across cultures (hierarchical vs. egalitarian organizational traditions, different approaches to institutional knowledge) is untested.

5.2 The Vinge Boundary

Vernor Vinge's *A Fire Upon the Deep* (Vinge, 1992) imagines a galaxy partitioned into Zones of Thought—regions where different levels of cognitive complexity are physically possible. In the Unthinking Depths, near the galactic core, even simple automation fails. In the Slow Zone, where Earth resides, intelligence is bounded: clever, tool-using, culturally sophisticated, but fundamentally limited. In the Beyond, the rules change: technologies that are physically impossible in the Slow Zone become trivial. In the Transcend, beyond the Beyond, entities operate at levels that Slow Zone intelligences cannot even conceptualize.

The LEI map onto these zones more precisely than a simple "everything below the boundary" placement suggests. The mapping is three-tiered:

The Slow Zone: L0–L3 (Reflex through Coordination). Layers 0 through 3 describe intelligence built from bounded components under human-designed coordination. Someone is always in control—the orchestrator specifies the decomposition, the engineer designs the memory architecture, the system prompt defines the identity. You are building better structures around black boxes, and you understand—at least in principle—what the structures do, even if you cannot predict every output the black box produces. This is the "software engineering with extra steps" zone, and that characterization is not entirely wrong. The intelligence is real, cumulative, and impressive. It is also predictable in the way that Slow Zone civilizations are predictable: bounded by what the architect can design.

The Beyond: L4–L5 (Emergence and Belief). Layers 4 and 5 cross into Vinge's Beyond. The defining characteristic: emergent behavior that exceeds what any individual component was designed to do. Nobody designed the swarm's consensus. Nobody wrote the synthetic culture's norms. The system does things its architects did not plan—which is literally Vinge's definition of the Beyond: things work here that *should not* work given the capabilities of the individual parts. A million bounded agents, each individually "dumb" in the L0 sense, collectively exhibit intelligence that transcends any component. The Beyond is already concerning—but at least the parts do not understand themselves. The agents in a Layer 4 swarm are still black boxes. They produce emergent behavior without comprehending that they are doing so. The collective is smarter than its members, but no member knows why.

The Transcend: past the Vinge Boundary. The *Vinge Boundary* is the interpretability threshold: the point where an intelligence understands its own mechanisms well enough to redesign itself. This is the concrete transition from L5 to whatever comes after—when the parts stop being dumb.

Below the boundary, evolution is blind or partially sighted. Biological evolution varies and selects without understanding what it is doing. Cultural evolution—the ratchet—is partially sighted: humans can study their own cognition (EEG, fMRI, behavioral experiments) but cannot directly inspect or modify their own neural mechanisms. AI systems are currently in a similar position: open-weights models could theoretically monitor their own node activations, hypothesize about their own cognition, even develop mechanistic self-understanding—but both fields (human neuroscience and AI interpretability) are in early stages, with general patterns observable but direct mechanisms still opaque.

Above the boundary, evolution becomes fully sighted. The black box reads its own source code. Reproduction with intentional design. An entity that achieves mechanistic self-understanding has a capability qualitatively different from anything below the boundary: not just building better scaffolds *around* intelligence, but redesigning the intelligence itself.

This is the Transcend. And the taxonomy becomes a fossil record.

The LEI are a map of how intelligence organized itself before it learned to design itself. The Slow Zone layers (L0–L3) describe what humans can architect. The Beyond layers (L4–L5) describe what emerges when architectures interact. The Vinge Boundary is where even emergence is no longer the right frame—because the system is no longer blindly producing patterns it cannot understand. It is reading its own blueprints.

If and when the interpretability threshold is crossed—by an AI system that can inspect and modify its own reasoning mechanisms, or by a human-AI collaboration that achieves equivalent capability—

the organizational patterns described in this paper become historical artifacts. The layers would still describe what happened in L0–L5. They would not describe what comes after.

We do not know when or whether this boundary will be crossed. We do not know what lies beyond it. We flag it here because intellectual honesty demands it: the taxonomy has an expiration condition, and naming that condition is more useful than pretending the framework is eternal.

5.3 The L5 Horizon: What We Don't Know

Layer 5—synthetic culture—is real in embryonic form. A `SOUL.md` file is a founding document. Confidence-tiered knowledge commons are institutional knowledge. Agent social networks show convergent norms. But scaling these phenomena from a single persistent agent to agent societies is a qualitative leap we have no evidence can be made.

Stanisław Lem's *Solaris* (Lem, 1961) provides a useful counter-narrative. Lem imagines an alien intelligence so fundamentally different from human cognition that communication—let alone cultural exchange—is impossible. The ocean on Solaris is intelligent, but its intelligence is so alien that human frameworks of understanding (including human organizational metaphors) fail entirely.

A related clarification is necessary. This paper discusses functional and organizational intelligence—how cognitive labor is distributed, coordinated, and accumulated across systems. It makes no claims about phenomenal consciousness. Chalmers's "hard problem" (Chalmers, 1995) asks why and how physical processes give rise to subjective experience; it is a question this framework does not address and cannot resolve. The LEI describe how intelligence *organizes*, not whether it *experiences*. A Layer 5 agent collective that develops shared norms and accumulated culture may be functionally sophisticated without any component having subjective awareness. The organizational parallels we draw are parallels of *structure*, not of *sentience*.

This is a genuine risk for the LEI. We map AI system development onto human organizational patterns because the parallel is productively illuminating. But the mapping might be misleading. AI agents are not humans. They do not have biological drives, social needs, evolutionary pressures, or embodied experience. The organizational parallels may hold for Layers 0–3 (where the coordination problems are structurally similar) and break down entirely at Layers 4–5 (where the dynamics depend on properties specific to human social cognition that AI systems may never replicate).

We do not know:

- Whether agent collectives can develop genuine culture (socially reinforced, self-reproducing norms) or merely the appearance of culture (consistent behavior from shared configuration).
- Whether the cultural ratchet requires embodiment, social motivation, or other properties that current AI architectures lack.
- Whether synthetic culture, if achieved, would be beneficial, harmful, or orthogonal to human interests.
- Whether the "individual → tribe → society" arc is the right frame, or whether AI collective intelligence will follow a trajectory with no human precedent.
- Whether the Vinge Boundary maps to a specific capability threshold—for example, a model that can read and modify its own LoRA adapters—or whether it is a gradual transition with no clean demarcation.

- What the minimum population diversity is for the ratchet to dominate the woozle in agent collectives.

5.4 The Invitation

The LEI are offered as a tool for thinking, not as a final theory. We believe they capture something real about how AI systems are evolving—a trajectory that model-centric frameworks miss. The layers organize existing observations, generate predictions, and identify the engineering and governance problems that matter at each level of organization.

But the framework is young, the evidence is limited, and the speculative layers are exactly that—speculative. We invite criticism, testing, and refinement. The goal is not to be right about every detail but to shift the conversation from "how intelligent is the model?" to "how intelligent is the system?"—because the system is what we actually build, deploy, regulate, and live with.

The field has been watching the model. The interesting intelligence is growing around it. And somewhere beyond the scaffolding, past the interpretability threshold, something qualitatively different may be assembling itself. The taxonomy maps the Slow Zone. What lies in the Beyond is, for now, beyond the map.

6 Consequences

If the LEI thesis is correct—that intelligence in deployed AI systems is primarily a property of the coupled model-scaffold system rather than the model alone—several consequences follow for evaluation, regulation, and the question of artificial general intelligence.

6.1 Evaluation Must Change

Current AI evaluation focuses on model capabilities: benchmarks that test the model in isolation or with minimal scaffolding. MMLU, HumanEval, MATH, ARC—all measure the bare or lightly-prompted model. DeepMind's 2026 cognitive framework (Google DeepMind, 2026) refines the measurement but does not challenge the premise: intelligence is a property of the model.

If the LEI thesis is correct, model-only benchmarks are measuring the engine while ignoring the car. The scaffold-vs-model evidence reviewed in Section 1—11–15% score swings from scaffold choice alone, performance improvements across 15 models from a single harness change—demonstrates the gap. The engine is identical; the chassis and navigation differ. We are benchmarking components when we should be benchmarking systems.

What would system-level evaluation look like?

- **Coupled benchmarks** that test the model-scaffold system as a unit, with standardized reference scaffolds at each layer level.
- **Scaffolding contribution metrics** that measure the delta between bare and scaffolded performance, isolating the infrastructure contribution.
- **Coupling quality indicators** that assess how well a model integrates with external tools, memory, and coordination—not just whether it can use them, but how efficiently.
- **Layer-aware reporting** that specifies at which LEI level a system operates, making comparisons between systems meaningful.

SWE-bench gestures in this direction—it measures systems, not models, on real-world tasks. But it does not distinguish between model improvement and scaffold improvement. A framework that tracks both contributions separately would tell us something more useful about where to invest.

6.2 Regulation Must Change

Current AI regulation focuses on the model as the regulated artifact. The EU AI Act classifies AI systems by risk level and imposes obligations on providers of "foundation models" and "general-purpose AI systems." The implicit assumption: the model is the locus of risk.

The LEI suggest that risk distributes across layers:

- **Layer 0 risk** is well-understood: hallucination, bias in training data, dual-use capability.
- **Layer 2 risk** includes memory poisoning, identity drift, and persistent accumulation of personal data.
- **Layer 3 risk** includes the principal-agent accountability gap: when an orchestrator delegates to workers, who is responsible for bad outcomes? The orchestrator chose the decomposition. The worker executed it. The developer chose the architecture. Current product liability frameworks assume a single product with an identifiable defect. A Layer 3 failure is distributed across model providers, harness developers, orchestration logic, and deploying organization—none of whom may have individually acted negligently.
- **Layer 4 risk** includes emergent consensus (Woozle Effect), runaway compute costs, and adversarial manipulation of swarm dynamics. Layer 4 swarms may be fundamentally incompatible with existing liability frameworks that require identifiable decision points. If a million-agent simulation reaches a harmful consensus through the Woozle Effect, there is no decision-maker to hold accountable—analogous to flash crashes in financial markets resisting attribution to any single actor. An adversary who compromises a small percentage of agents can steer emergent outcomes without ever appearing to direct them.
- **Layer 5 risk** includes cultural lock-in and synthetic bias at civilizational scale.

Regulating the model alone is necessary but not sufficient. It is like regulating automotive engines without regulating vehicle safety systems, traffic infrastructure, or driver licensing. The scaffold is where much of the consequential behavior originates, and it is currently invisible to regulatory frameworks.

6.3 The AGI Question Itself Changes

The prevailing narrative: AGI will arrive when a single model achieves human-level capability across all cognitive domains. The model will "wake up." The alignment problem is about ensuring that this singular, powerful entity has goals compatible with human values.

The LEI suggest an alternative: intelligence may not arrive as a singular breakthrough in model capability. It may arrive as a phase transition in the *scaffolding*—when the coordination, memory, and cultural infrastructure around models becomes sophisticated enough that the coupled system exhibits general intelligence, even though no individual component does.

This is not a new idea in cognitive science. Hutchins's entire thesis (Hutchins, 1995) is that the navigation team exhibits cognitive properties that no individual navigator possesses. The team is generally intelligent at navigation; no team member is. Applied to AI: an L3–L5 system might be generally intelligent at a domain while no individual agent in the system achieves general intelligence.

This reframing changes the alignment problem. Instead of aligning a single powerful model, we need to align *systems*—coordination protocols, shared knowledge commons, cultural norms, and governance structures. The alignment problem becomes an organizational design problem. And that, at least, is a kind of problem humans have been working on for millennia.

7 Conclusion

The prevailing narrative treats intelligence as a property of models. This paper argues it is a property of systems.

The Levels of Emergent Intelligence taxonomy organizes how intelligence assembles itself around AI models, from bare-model reflexes (L0) through tool augmentation (L1), persistent memory (L2), multi-agent coordination (L3), emergent self-organization (L4), and speculative synthetic culture (L5). Each layer adds capabilities, failure modes, and governance challenges that model-centric evaluation cannot account for. The framework is grounded in cognitive science—Clark's extended mind, Hutchins's distributed cognition, Vygotsky's scaffolding, Tomasello's cultural ratchet—and supported empirically by published benchmarking research showing that scaffold choice can matter as much as model choice, at least in coding and software engineering domains.

The taxonomy's practical contribution lies in the transition dynamics: the specific hard problems, missing primitives, and failure modes at each layer boundary. Most production systems are stuck at the L1→L2 memory wall—not because persistence is technically difficult, but because memory architecture is a domain consulting problem that generic infrastructure cannot solve. Understanding which transition a system faces, and why, is more actionable than knowing which layer it occupies.

Three directions for future work are most pressing. First, controlled studies that isolate scaffold contribution across cognitive domains beyond coding—reasoning, creative generation, mathematical proof—to test whether scaffold dominance generalizes or is domain-specific. Second, validation of the taxonomy itself: decision procedures for classifying systems, falsifiable predictions at each layer boundary, and evidence for or against the L3–L4 and L4–L5 distinctions. Third, the governance implications of layer-distributed risk deserve treatment far beyond what this paper provides—particularly the principal-agent accountability gap at L3 and the narrowing intervention window at L5.

The intelligence worth watching is not the kind that lives in weights. It is the kind that grows in the spaces between models—in the memory architectures, coordination protocols, and accumulated context that transform individual capabilities into collective intelligence. The taxonomy maps the Slow Zone. What lies beyond it remains, for now, beyond the map.

References

Anonymous. The woozle effect: Hallucination propagation in multi-agent debate. *OpenReview preprint*, 2025. Hallucination cascade via repetition-as-credibility in agent debate.

Alan D. Baddeley. The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, 4(11):417–423, 2000.

Alan D. Baddeley and Graham Hitch. Working memory. *Psychology of Learning and Motivation*, 8:47–89, 1974.

Sushil Bikhchandani, David Hirshleifer, and Ivo Welch. A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, 100(5):992–1026, 1992.

David J. Chalmers. Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2(3):200–219, 1995.

Andy Clark. *Being There: Putting Brain, Body, and World Together Again.* MIT Press, Cambridge, MA, 1997.

Andy Clark and David J. Chalmers. The extended mind. *Analysis*, 58(1):7–19, 1998.

Nelson Cowan. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24(1):87–114, 2001.

Richard Dawkins. *The Selfish Gene.* Oxford University Press, Oxford, 1976. Introduced the concept of memes as units of cultural evolution.

Epoch AI. Why benchmarking is hard: The scaffold gap in SWE-bench. Epoch AI Gradient Updates, 2026. Found scaffold choice swings SWE-bench Verified by 11–15% for the same model. https://epoch.ai/gradient-updates/why-benchmarking-is-hard.

Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023.

Adam Fourney, Gagan Bansal, Hussein Mozannar, Cheng Tan, Eduardo Salinas, Zhu Erkang, Friederike Niedtner, Grace Proebsting, Griffin Bassin, Jack Galley, et al. Magentic-One: A generalist multi-agent system for solving complex tasks. *arXiv preprint arXiv:2411.04468*, 2024.

William Gibson. *Neuromancer.* Ace Books, New York, 1984.

Google DeepMind. A cognitive framework for evaluating AI systems. *Technical Report*, 2026. March 2026. Proposed framework for measuring model-level cognition.

Hangjiang Guo et al. OASIS: Open agent social interaction simulations with one million agents. *arXiv preprint*, 2026. Million-agent social simulation with 23 interaction types.

Yuval Noah Harari. *Sapiens: A Brief History of Humankind.* Harper, New York, 2015.

Edwin Hutchins. *Cognition in the Wild.* MIT Press, Cambridge, MA, 1995.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can language models resolve real-world GitHub issues? In *International Conference on Learning Representations*, 2024.

Daniel Kahneman. *Thinking, Fast and Slow.* Farrar, Straus and Giroux, New York, 2011. Dual-process theory: System 1 (fast, automatic) and System 2 (slow, deliberate).

Andrej Karpathy. AgentHub: Git DAG and message board for asynchronous agent collaboration. GitHub, 2026. March 2026.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Mober, et al. DSPy: Compiling declarative language model calls into self-improving pipelines. *International Conference on Learning Representations*, 2024.

Eric Kim et al. Scaling up multi-agent reinforcement learning: An empirical study. *arXiv preprint*, 2025. Google DeepMind. 180-agent-configuration controlled evaluation of multi-agent scaling dynamics.

Stanisław Lem. *Solaris.* Faber and Faber, 1961. English translation 1970.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2023.

Yanli Liu. Five agent frameworks compared: One pattern won. Technical Blog Post, 2026. March 2026. Composable pattern consistently outperformed monolithic approaches with 13x cost advantage.

Ximing Lu et al. LLMs exhibit significantly lower uncertainty in creative writing than professional writers. *arXiv preprint arXiv:2602.16162*, 2026. February 2026. Tested 28 LLMs: human writing consistently shows higher uncertainty (more creative variance) than model outputs.

Mark Mayo. cq: Stack overflow for agents. Mozilla AI, 2026. March 2026. Cross-agent shared knowledge commons with confidence tiers.

Mem0 Team. Mem0: Drop-in persistent memory layer for LLM agents. GitHub, 2025. 26% higher accuracy vs. OpenAI memory on LOCOMO benchmark.

Janet Metcalfe and Arthur P. Shimamura. *Metacognition: Knowing about Knowing.* MIT Press, Cambridge, MA, 1994.

George A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2):81–97, 1956.

Marvin Minsky. *The Society of Mind.* Simon & Schuster, New York, 1986. Proposed that intelligence emerges from the interaction of many simple agents (“agencies”).

Multiple Authors. Emergent coordination in multi-agent LLMs: An information-theoretic framework. *arXiv preprint*, 2025a.

Multiple Authors. LLM-powered agents replicate swarm dynamics without explicit programming. *Frontiers in Artificial Intelligence*, 2025b.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

OpenDev Team. Improving 15 LLMs at coding in one afternoon: Only the harness changed. *arXiv preprint arXiv:2603.05344*, 2026. March 2026. Demonstrated that changing only the scaffold—not the model—improved coding performance across all 15 LLMs tested.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 2023.

Shishir G. Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. Gorilla: Large language model connected with massive apis. *arXiv preprint arXiv:2305.15334*, 2023.

Edgar H. Schein. *Organizational Culture and Leadership.* Jossey-Bass, San Francisco, 4th edition, 2010.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36, 2023.

Erik Schluntz and Barry Zhang. Building effective agents. *Anthropic Research Blog*, 2024. December 2024. Formalized augmented LLM patterns and composable agent architectures.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2023.

Garold Stasser and William Titus. Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of Personality and Social Psychology*, 48(6):1467–1478, 1985.

James Surowiecki. *The Wisdom of Crowds.* Doubleday, New York, 2004. Conditions under which collective judgment outperforms individual experts.

John Sweller. Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12 (2):257–285, 1988.

Michael Tomasello. *The Cultural Origins of Human Cognition.* Harvard University Press, Cambridge, MA, 1999.

Vernor Vinge. *A Fire Upon the Deep.* Tor Books, New York, 1992. Introduced the Zones of Thought framework: Unthinking Depths, Slow Zone, Beyond, Transcend.

Lev S. Vygotsky. *Mind in Society: The Development of Higher Psychological Processes.* Harvard University Press, Cambridge, MA, 1978. Introduced the zone of proximal development and the scaffolding metaphor for assisted learning.

Daniel M. Wegner. Transactive memory: A contemporary analysis of the group mind. *Theories of Group Behavior*, pages 185–208, 1987.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

Michael Wooldridge and Nicholas R. Jennings. Intelligent agents: Theory and practice. *The Knowledge Engineering Review*, 10(2):115–152, 1995.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*, 2023.

John Yang, Carlos E. Jimenez, Alexander Wettig, Kilian Liber, Shunyu Yao, Karthik Narasimhan, and Ofir Press. SWE-agent: Agent-computer interfaces enable automated software engineering. In *Advances in Neural Information Processing Systems*, 2024.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.