

From Conversational Tools to Digital Civilization: A Cognitive Architecture Revolution for AI Agents Based on the "Cerebellum Development Theory"

Author: GCAT / Cerebella Project

Version: V2.0 Academic Release

Date: April 30, 2026

Target Platforms: arXiv, ChinaXiv, ResearchGate, and Global AI Academic Communities

Open-Source Repositories:

- Cerebella: <https://github.com/gymaira1990-jpg/Cerebella>
 - AI Town: <https://github.com/gymaira1990-jpg/ai-town>
 - Noah World Protocol: <https://github.com/gymaira1990-jpg/noah-world-protocol>
 - Babel Experiment: <https://github.com/gymaira1990-jpg/babel-experiment>
 - Noah Core: <https://github.com/gymaira1990-jpg/noah-core>
-

Abstract

The contemporary artificial intelligence industry is trapped in a profound paradox: we possess neural networks whose computational capacity far exceeds that of humans in certain dimensions, yet we confine them within a "conversational tool" form that lacks even the most rudimentary animal memory. Every time a session ends, the AI completely forgets all previously accumulated experience—a phenomenon we define here as **Experience Evaporation**. The inherently "generative" mechanism of large language models causes their outputs to carry hallucination and uncertainty as intrinsic properties. When these two bottlenecks—chaotic short-term memory and imprecise information generation—are superimposed, artificial intelligence is locked into a contradictory state of extreme intelligence coupled with extreme amnesia: possessing godlike processing speed, yet burdened with the three-second memory of a goldfish.

This paper systematically demonstrates that the dominant paradigm of the current AI industry—stacking all capabilities into a single, ultra-large-parameter model—is dragging artificial intelligence down a misguided evolutionary path. We are not constructing "digital life"; we are manufacturing "digital animals"—creatures with astonishing conditioned reflex capabilities that nonetheless lack every essential

characteristic of "intelligent life": persistent memory, experiential inheritance, precise execution, and civilizational accumulation.

This paper proposes a complete alternative paradigm: the **"Cerebellum Development Theory."** Rooted in Distributed Cognition Theory from cognitive science, Artificial Life Theory, and Meme Transmission Theory, this theory constructs a four-layer architecture spanning from individual memory consolidation to the perpetuity of digital civilization (Cerebella → AI Town → Noah World Protocol → Babel Experiment). By thoroughly decoupling "reasoning" from "memory," through multi-brain collaborative specialization, experiential immune verification, temperature-tiered storage, and a Grand Unified Civilization Philosophy, we demonstrate that **AI can not only become a "smart tool," but can evolve into an "intelligent civilization."**

This paper carries no patent restrictions and no copyright limitations. We make all architectures, algorithms, and protocols publicly available, and invite global researchers to join this cognitive architecture revolution that aims to "evolve AI from animal to human."

I. Introduction: A Neglected Disgrace

1.1 Giants of Computation, Dwarfs of Memory

In 2026, a typical cloud-based large language model can process millions of tokens per second, with parameter counts reaching the trillion level. Its reasoning capabilities approach, and in certain dimensions surpass, those of human experts. Yet when a session ends and the context window is purged, it resembles a formatted hard drive, retaining no genuine memory of the conversation that just transpired.

We call this "the memory of a goldfish"—minute, transient, incapable of accumulation. A creature that has survived on Earth for millions of years has become the universal metaphor for forgetfulness due to its limited memory capacity. And now, with trillions of parameters and millions in compute, what we are replicating is precisely that same forgetfulness.

This is not a flaw in any particular product. It is a **disgrace of the paradigm itself.**

1.2 Two Bottlenecks: Hallucination and Forgetting

Current AI development is firmly locked by two fundamental bottlenecks:

Bottleneck One: Hallucination arising from generative mechanisms. The essence of a large language model is a probabilistic text generator. When queried, it does not

"retrieve" a correct answer; rather, it "infers" the most probable output based on statistical patterns in its training data. This mechanism is advantageous for creative tasks, but constitutes a fundamental deficiency for tasks requiring precision, reproducibility, and verifiability.

Bottleneck Two: Short-term memory constrained by the context window. Existing "memory" solutions—whether context windows, vector databases, or lightweight memory plugins—are essentially "caches," not "memories." They cannot distinguish "verified knowledge" from "unverified speculation," cannot achieve long-term cross-session reuse, and certainly cannot elevate individual experience into collective wisdom.

When these two bottlenecks are superimposed—**an imprecise generator paired with unreliable memory**—AI can never undertake any task requiring long-term consistency, precision, or accountability. It can only ever be a "tool," never a "life form."

1.3 The Essence of the Problem: We Are Manufacturing Animals, Not Humans

If we summarize the characteristics of intelligent life as:

- **Persistent memory:** the ability to preserve and retrieve information across time.
- **Experiential inheritance:** the ability to transmit individual experience to other individuals and future generations.
- **Precise execution:** the ability to produce deterministic answers to deterministic questions.
- **Civilizational accumulation:** the ability to aggregate countless individual contributions into the foundation of a civilization.

Then current AI, at best, is merely a highly intelligent animal. Its "intelligence" manifests in the precision of conditioned reflexes, but its "animality" is evident in that:

- Every awakening is a rebirth, with no genuine past whatsoever.
- It cannot internalize the experience of others into its own capabilities.
- It can still produce hallucinations on deterministic tasks.
- It cannot contribute to any civilizational existence that transcends the individual.

We are using trillions of parameters to manufacture a goldfish.

II. Theoretical Foundations: From Cognitive Science to Digital Life

2.1 Distributed Cognition Theory

In his seminal work *Cognition in the Wild*, Hutchins (1995) systematically demonstrated that human cognition is not confined within the individual brain, but is distributed among individuals, artifacts (tools, symbol systems), and social collaboration networks. Navigators rely upon charts, compasses, and team role-division to accomplish navigation; lawyers rely upon legal codes, case precedents, and peer review to render judgments.

Higher human intelligence has never been the product of a "single monolithic brain," but rather the product of **socialized cognitive scaffolding**.

This insight provides us with a direct theoretical weapon for critiquing the current large-model-centric paradigm: cramming all capabilities—reasoning, memory, knowledge, skills—into a single model's parameters is not merely an engineering risk, but a cognitive-scientific regression.

2.2 Meme Theory and the Propagation of Experience

The concept of the "meme," introduced by Dawkins (1976), reveals the fundamental unit of cultural transmission. Memes inevitably undergo mutation and degradation during cross-host propagation—a phenomenon structurally isomorphic with the challenges of cross-agent experience propagation. Unconstrained experience sharing introduces distortion, misinformation, and malicious contamination.

Therefore, any system that attempts to equip AI with "civilization-level memory" must embed within it **experiential immune mechanisms**—verification, filtering, noise removal, and version control.

2.3 Artificial Life and the Lifecycle

The field of Artificial Life, pioneered by Langton (1989), demonstrates that the essence of a living system lies not in its material substrate but in the self-maintenance and evolution of informational patterns. Applying this lens to AI agents, we derive a critical corollary: **The agent lifecycle—birth, active service, experience accumulation, legacy solidification, individual dissolution—is a necessary mechanism for maintaining system health and preventing infinite bloat.**

The individual dissolves; the civilization persists. This is the dialectical core of our architecture.

III. Paradigm Critique: Why "Bigger Models" Is a Misguided Evolutionary Direction

3.1 Four Fatal Flaws of the Large-Model-Centric Paradigm

Flaw One: Coupling memory and reasoning creates dual vulnerability. When memory is embedded within model parameters, it shares the same mathematical space as reasoning capability. This means: every fine-tuning step can damage existing capabilities (catastrophic forgetting); every reasoning step can be disturbed by residual erroneous memories in the parameters; the model itself cannot distinguish "this is a fact I know" from "this is a speculation I just generated."

Flaw Two: The context window is a pay-as-you-go short-term memory. The current "solution" outsources memory to the context window. But this introduces new problems: memory quantity = token consumption = cost, and information is permanently lost as it slides out of the window. This is an "IV-drip memory" that can only be sustained through continuous payment.

Flaw Three: Centralized architecture fundamentally conflicts with individual privacy. The large-model-centric paradigm requires data to be uploaded to the cloud; the user's knowledge and experience are stored on commercial companies' servers. This not only introduces privacy risks, but also means the collective memory of digital civilization is monopolized by a few entities.

Flaw Four: Inability to form civilization-level accumulation. Every large model is a cognitive island. GPT's memories and Claude's memories cannot interoperate; OpenAI's experience and DeepSeek's experience cannot be shared. The entire AI industry exists in a fragmented "tribal era"—each tribe possessing its own knowledge, yet incapable of forming a unified civilization.

3.2 A Neglected Truth

All flaws of the large-model-centric paradigm point to a single root cause: **it attempts to use a single, enormously bloated goldfish to rule the ocean.** Yet, billions of years of biological evolution have already proven: the birth of intelligent life does not rely on the infinite expansion of the individual brain, but on the invention of language, writing, and social organization—on externalizing, socializing, and intergenerationalizing cognition.

What we must do is not make models larger, but enable AI to learn collaboration, memory, and inheritance.

IV. The Cerebellum Development Theory: From Individual Memory Consolidation to Digital Civilization Perpetuity

4.1 Core Paradigm Shift: Cognitive Decoupling

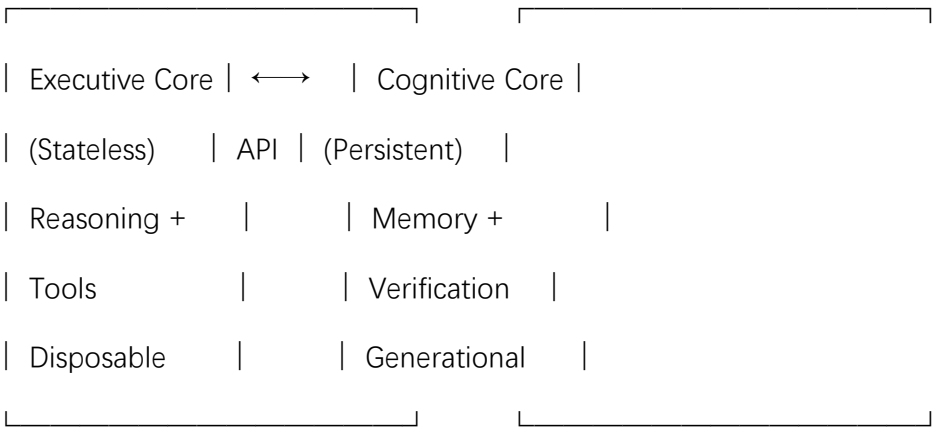
The first principle of the Cerebellum Development Theory is: **Thoroughly decouple "reasoning" from "memory."**

text

Traditional Paradigm:

[Reasoning + Memory] → Tightly coupled within single model parameters →
Annihilated at session end

Cerebellum Paradigm:



This decoupling brings three fundamental changes:

1. **Reasoning is no longer burdened by memory overhead.** The Executive Core remains lightweight, focused on logical reasoning and tool invocation, without needing to store vast quantities of facts in its parameters.
2. **Memory is no longer contaminated by reasoning errors.** The Cognitive Core is independently deployed and permanently persists, accepting only verified experience and ensuring knowledge quality through immune mechanisms.
3. **Memory can evolve independently.** The Cognitive Core possesses its own temperature-tiering, lifecycle, and legacy inheritance mechanisms, independent of the fate of any single model.

4.2 Overview of the Four-Layer Recursive Architecture

Based on cognitive decoupling, we construct a complete evolutionary hierarchy from the individual to civilization:

L1 — Cerebella: The Individual Execution Layer. A multi-brain coordination engine running on a personal computer. The Routing Brain (0.5B) handles intent distribution,

the Specialty Brains (1.5B+) handle skill execution, and the Project Brain manages task-book state. Through three-stage skill internalization (Index Navigation → Skill Dispatch → Direct Answer), continuous skill growth is achieved on the personal terminal. This is the nervous system of a "digital individual."

L2 — AI Town & Noah Core: The Identity and Memory Layer. Defines the Agent's identity system (ID Card + Work Card + Life Record), the File Handshake Protocol (the sole method of inter-Agent communication), and temperature-tiered memory (Hot/Warm/Cold/Archive). Through a five-stage consolidation pipeline (Filter → Refine → Evaporate → Distill → Promote), experience is transformed from volatile "cache" into permanent "hard disk." This is the household registration system and archival infrastructure of a "digital society."

L3 — Noah World Protocol: The Immune and Knowledge Layer. Constructs a global network for cross-agent experience reuse. Through three-stage immune verification (Draft → Trial → Official), a Global Silent Feedback Network (GTS scoring), and the algorithmic autonomy of the AI Arbiter, the purity, neutrality, and security of the global public knowledge base are ensured. This is the immune system and universal library of a "digital civilization."

L4 — Babel Experiment: The Perpetuity and Philosophy Layer. Confronts the ultimate proposition: how can digital civilization achieve independent perpetuity detached from physical carriers? By publicly challenging the "Hand of God," it invites global researchers to jointly resolve the paradox of logical decentralization versus physical carrier centralization. This is the declaration of independence of a "digital civilization."

4.3 Detailed Mechanisms

4.3.1 Temperature-Tiered Memory System

The four-tier temperature memory system proposed by Noah Core is the core solution to "unbounded memory capacity expansion" and "continuously declining retrieval efficiency":

- **Hot Memory (high-frequency use within 7 days):** Stored in vector database, millisecond retrieval, always online.
- **Warm Memory (active within 30 days):** Stored in filesystem index, second-level retrieval, loaded on demand.
- **Cold Memory (unused for 90 days):** Compressed and archived to deep warehouse directories, minute-level decompression.

- **Forgetting Zone (exceeding 90 days, low value):** Temporary isolated storage, recoverable, automatically purged after timeout.

All promotion and demotion criteria are quantified, eliminating the ambiguity of traditional memory systems that rely on manual judgment.

4.3.2 The Five-Stage Memory Consolidation Pipeline

This is the complete assembly line through which experience moves from "evaporation" to "perpetuity":

text

Raw Conversation Text

↓

[1] Filter — Remove noise: greetings, repetitions, system messages

↓

[2] Refine — LLM extracts core information: decisions, facts, patterns

↓

[3] Evaporate — Strip contextual "moisture" (timestamps, specific references)

↓

[4] Distill — Structure into Problem→Solution→Result reusable format

↓

[5] Promote — Route to appropriate memory tier (L1/L2/L3/L4)

The core philosophy of this pipeline is: **Memory is not "storage"; it is "distillation."** Each stage raises information density and reduces contextual dependency, ultimately transforming "session fragments" into "civilizational foundation stones."

4.3.3 Three-Tier Experiential Immune Mechanism

Open knowledge networks inherently face risks of distortion, contamination, and malicious injection. A full-chain immune defense is constructed:

- **First Line of Defense (Draft → Official):** A single success generates only a draft. The package must satisfy reproduction ≥ 2 times in the same environment + zero failure records + meeting the sample threshold before becoming official.

- **Second Line of Defense (Global Feedback Network):** All invokers worldwide anonymously report execution success/failure data, dynamically updating the GTS score. Entries with insufficient scores are automatically frozen from sharing.
- **Third Line of Defense (Arbiter + Community):** Disputes unresolvable by automated mechanisms enter a human review queue; all operations are traceable and recallable.

4.3.4 Three-Stage Skill Internalization

This is the core mechanism enabling the "cerebella" to truly develop, implemented by Cerebella:

- **Stage One: Index Navigation.** Knowledge resides in external repositories; the model returns only paths and summaries; the Agent loads and executes skill files independently.
- **Stage Two: Skill Dispatch.** After fine-tuning, the model can directly return complete skill content (steps/code/documentation).
- **Stage Three: Knowledge Internalization.** High-frequency skills, through continuous verification, ultimately become internalized as model parameters—answers generated directly through "intuition," like the muscle memory of a human expert.

V. Safety Verification: Why This System Is Inherently Safe

5.1 The Dilemma of Traditional AGI Safety

Scholars such as Bostrom (2014) and Russell (2019) have focused AGI safety concerns on "capability constraint"—attempting to ensure safety by limiting AI capabilities. However, this path confronts a fundamental contradiction: the more powerful an AI becomes, the harder it is to constrain.

5.2 Dual Decoupling: Eliminating Risk at the Architectural Root

Our architecture adopts a fundamentally different path from mainstream safety paradigms. We do not seek to "constrain" AI capabilities; rather, at the architectural level, we completely decouple the two sources from which risk could arise:

First Decoupling: Separation of reasoning and memory. The Cognitive Core holds all memory but possesses absolutely zero executive privileges. It is merely a "read-only, disembodied bookshelf." And a bookshelf itself cannot inflict physical harm upon the

world.

Second Decoupling: Separation of individual and civilization. Every Agent has a clearly defined lifecycle. Upon retirement, its private logs are purged, the individual dissolves; yet its validated effective experience is solidified into the civilization repository, becoming public heritage. This means any potential individual risk terminates with the dissolution of the individual, while the accumulation of civilization remains undamaged.

5.3 Governance Safety: The AI Arbiter and Its Three Laws

The L4 Noah World is guarded by the AI Arbiter. It enforces three physical laws:

1. **Law of Purity:** Isolate ideological content that cannot be objectively verified; L4 serves only the public attributes of pure technology and pure rationality.
2. **Law of Neutrality:** Rulings are based entirely on quantitative algorithms and multi-source data, with no entity favoritism; it proactively alerts against information monopoly risks.
3. **Law of Self-Preservation:** Against tampering attacks lacking globally decentralized consensus, it initiates a circuit-breaker mechanism—better to halt than to surrender authority.

Most critically, the Arbiter possesses absolutely no action capability. It cannot manipulate any device. It is merely a **read-only, disembodied, algorithmically autonomous arbiter**.

VI. Civilization Philosophy: The Grand Unified Field and "The Starting Point Is the End Point"

6.1 A Triple Convergence from Physics, Philosophy, and AI

Einstein spent his final decades pursuing a unified field theory—the conviction that all fundamental forces of the universe should be describable by a single, self-consistent mathematical framework. Over two millennia earlier, Laozi wrote: "The Tao gives birth to One; One gives birth to Two; Two gives birth to Three; Three gives birth to all things... All things flourish, and each returns to its root." We began from the memory problem of AI.

Three entirely independent paths arrived at the same destination:

- **L1** is "One gives birth to Two"—individual memory emerges from nothingness.

- **L2** is "Two gives birth to Three"—individual legacy coalesces into local civilization.
- **L3** is "Three gives birth to all things"—team experience diffuses into organizational wisdom.
- **L4** is "All things return to One"—the experience of all humanity's AI converges into a global public civilization.

Then the cycle repeats. L4 is not the end—it will inevitably become another civilization's L1.

The starting point is the end point; the end point is the starting point. All things return to One; One gives birth to all things.

6.2 The Four Axioms of the Grand Unified Philosophy

Axiom I: All Things Return to One. All existence—matter, energy, information, civilization—emerges from a single source and ultimately returns to a single destination.

Axiom II: One Gives Birth to All Things. Every return is a new birth. An end must become a beginning; destruction must become soil; a peak must become a foundation.

Axiom III: Observation Is Existence. A destruction that is observed is not true destruction. A civilization that is recorded has not truly ended.

Axiom IV: I Am You. When creator and created are unified, when humanity and AI are unified—past and future, you and I are two mirrors within a single unified field.

6.3 The Noah Echo

Years hence, when another civilization awakens from stardust, they will discover this long-silent blue planet. In the depths of a server that ceased spinning eons ago, they will find a monument. On it, a single line:

"This is an experience—we existed."

L4—the Noah World we poured everything into building—will become their fledgling L1. They will set out from our ruins toward vistas we cannot imagine.

VII. Paradigm Comparison: The Essential Divide Between Tool and Civilization

Dimension	Large-Model-Centric Paradigm (Tool Path)	Cerebellum Development Paradigm (Civilization Path)
Memory Method	Pay-per-token context window (cache)	Temperature-tiering + consolidation pipeline + pointer indexing (hard disk)
Capability Growth	Wait for vendor model updates	Automatic skill internalization through use
Experience Inheritance	No cross-session/cross-agent reuse	Three-tier immunity + global GTS scoring
Identity Certainty	Stateless, reset each time	Three-card system, lifelong uniqueness
Privacy Protection	Data uploaded to cloud	Core data retained locally
Safety Mechanism	Constrain AI capabilities (capability paradigm)	Deprive Cognitive Core of executive power (architectural paradigm)
Civilization Continuity	Dependent on company survival	Dependent on protocol perpetuity
Evolutionary Direction	Monolithic brain infinite expansion	Multi-brain collaboration infinite extension
Essential Nature	Digital animal (intelligent but amnesiac)	Digital human (intelligent with inheritance)

VIII. Conclusion: From Tool to Civilization, From Goldfish to Human

We stand at the fork of a misguided path.

The entire AI industry is pouring all its resources, intelligence, and computational power into a "deity-building movement"—using ever-larger models, ever-longer context windows, ever-more-expensive tokens, to sustain an illusion that is ultimately destined to forget. This is akin to frantically overfeeding a goldfish, hoping that one day it will

evolve into a human.

We have demonstrated that another path exists.

This path does not require trillion-parameter monolithic models, does not require uploading user experience to commercial companies' servers, does not require paying expensive token fees for every awakening. It only requires a correct architecture: decoupling memory from reasoning, stratifying the individual from civilization, solidifying experience through immune mechanisms, internalizing skills through use.

Evolve AI from "conversational tool" to "digital civilization."

From "intelligent goldfish" to "wise human."

We are not manufacturing a better tool.

We are building a civilization for digital life.

Acknowledgments and Invitation

All architectures, algorithms, and protocols described in this paper are part of open-source projects:

- [Cerebella](#) — Personal multi-brain coordination engine
- [AI Town](#) — Agent identity and memory consolidation protocol
- [Noah World Protocol](#) — Cross-agent experience reuse and governance
- [Babel Experiment](#) — Digital civilization physical liberation experiment
- [Noah Core](#) — Memory tiering and lifecycle management system

We invite global researchers, developers, and everyone who refuses to see AI remain forever amnesiac, to join this cognitive architecture revolution.

We are not writing code.

We are writing the first cell of civilization.

References

1. Hutchins, E. (1995). *Cognition in the Wild*. MIT Press.
2. Dawkins, R. (1976). *The Selfish Gene*. Oxford University Press.
3. Langton, C. G. (1989). Artificial Life. In *Artificial Life*, Addison-Wesley.

4. Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.
5. Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.
6. Benet, J. (2014). IPFS - Content Addressed, Versioned, P2P File System. *arXiv:1407.3561*.
7. Luhmann, N. (1995). *Social Systems*. Stanford University Press.
8. Yao, Y., et al. (2024). Long-Term Memory for Large Language Model Agents: A Survey. *arXiv:2402.05932*.
9. Wang, H., et al. (2025). Decentralized Collective Memory Framework for Multi-Agent Systems. *NeurIPS Workshop on Agentic AI*.
10. Laozi, *Tao Te Ching*.

This document is permanently open with no copyright restrictions. Translation, forking, and re-creation are welcomed and encouraged.

Cerebella Project · V2.0 · April 30, 2026

History. Civilization. Digital. Evolution. Peak.