

SOMA v1.1.1

Teaching AI to Think Like a Sage

— An Open-Source Wisdom Framework for Artificial Intelligence —

May 2026 · soma-wisdom Open Source Project

github.com/sunyan999999/soma

1. What Is SOMA? In One Sentence

SOMA is an **open-source wisdom thinking framework** that equips large language models and AI agents with a set of "thinking methodologies" — just as humans use principles like "first principles thinking," "systems thinking," and "the Pareto principle" to reason better, SOMA enables AI to automatically apply these wisdom laws when analyzing problems.

You don't need to understand the technology. Think of SOMA as giving your AI a "thinking coach" — when you ask a question, this coach works behind the scenes to select and apply the most appropriate thinking frameworks, helping the AI think deeper, broader, and more reliably.

An analogy: Without SOMA, an AI is like a smart but unstructured student. With SOMA, it becomes a trained sage — it examines problems from multiple angles, draws on past experience, and even reflects on and evolves its own thinking.

2. Why Does AI Need a "Thinking Framework"?

Today's large language models are remarkably capable, but they share a fundamental weakness: **they don't think in a structured way**. Ask "how can I improve my team's productivity?" and you might get a reasonable-sounding answer, but it will lack depth, rely on a single perspective, and likely miss critical factors.

SOMA addresses this exact problem. It comes with seven validated wisdom laws built in:

| Wisdom Law | In Plain English | What It Does |
|------------------|---|--|
| First Principles | Break things down to their fundamentals and reason up | Cuts through surface noise to find root causes |

| | | |
|------------------------|---|--|
| Systems Thinking | See everything as an interconnected whole | Finds leverage points instead of treating symptoms |
| Contradiction Analysis | Identify deeper tensions beneath surface problems | Reveals hidden structural conflicts |
| Pareto Principle | 80% of results come from 20% of causes | Focuses effort on the critical few for maximum impact |
| Inversion | Think backwards: how would you make it worse? Then avoid that | Breaks through blind spots and reveals hidden risks |
| Evolutionary Lens | View things through the lens of time and change | Understands where things came from and where they're going |
| Analogical Reasoning | Use what you know to understand what you don't | Transfers insights across domains |

When the AI receives a question, SOMA automatically determines which wisdom laws are most relevant, combines them by weight, and applies them — much like an experienced thinker instinctively draws on multiple mental tools when facing a new challenge.

3. What's New in v1.1.1: Three Breakthroughs

v1.1.1 is a milestone release. It delivers three critical breakthroughs:

Breakthrough 1: Knowing When to Go Deep — and When to Keep It Simple

Previous versions had a glaring issue: whether you asked "What is photosynthesis? Answer in 50 words" or "How should we set our five-year corporate strategy?", the AI would deploy its entire thinking framework and produce a thousand-word essay. It was like going to a convenience store for water and getting a product launch presentation.

What v1.1.1 does:

- Automatically assesses question complexity — simple questions take the "fast lane" for a concise, direct answer
- Automatically detects user length constraints — say "in 50 words" and it actually stays around 50 words
- Reserves the deep reasoning framework for genuinely complex questions — the right tool for the right job

Real results: In Zero Entropy Think Tank testing, before the fix, a "50 words" request returned 1,268 words. After the fix, it lands precisely around 50 words. Response latency for simple questions dropped from 14–26 seconds to 2–3 seconds.

Breakthrough 2: True Collective Intelligence

SOMA v1.1.0 introduced multi-agent collaboration — you can register multiple "experts" (like a business strategist, a technical architect, a product manager) and have them discuss problems together. But testing revealed a bug: with three experts registered, only one was actually working.

What v1.1.1 does:

- Fixed the expert routing logic — all registered experts are now invited to participate
- Parallel dispatch lets multiple experts think simultaneously — with 5 experts, speed improves nearly 5x
- Distributed evolution means each expert's experience feeds back into the entire system

Real results: A complex business problem now gets simultaneous input from a market strategist, a technical feasibility architect, and a user-experience product manager — then a consensus mechanism synthesizes a comprehensive solution.

Breakthrough 3: From "Reset Every Time" to "Gets Smarter With Use"

Most AI tools start from scratch with every conversation. SOMA is different — it remembers, reflects, and evolves.

v1.1.1's complete evolution loop:

- **Memory:** Every conversation's important insights are automatically stored and indexed
- **Reflection:** Each analysis undergoes quality self-assessment — good patterns are reinforced, bad ones weakened
- **Evolution:** Wisdom law weights auto-adjust based on real-world effectiveness — if "systems thinking" proves particularly powerful for certain domains, its weight automatically increases
- **Pre-warming:** The embedding model downloads in the background — no more waiting on first launch

4. What This Means for AI Applications Like the Zero Entropy Think Tank

The Zero Entropy Think Tank (零熵智库) is one of SOMA's first real-world deployments. SOMA delivers three layers of value for such applications:

Layer 1: A Quantum Leap in Answer Quality

Without SOMA, an AI application is a Q&A information retriever. With SOMA, it becomes a thinking partner that analyzes from multiple angles, draws on accumulated experience, and engages in self-reflection. Users no longer need to structure their own thinking — the AI does it automatically.

Layer 2: An Accumulating Wisdom Asset

Ordinary AI tools "reset" after every conversation. SOMA's memory system enables continuous accumulation — today's strategic discussion becomes tomorrow's reference for analyzing new problems. A Think Tank used for three months is significantly smarter than one on its first day.

Layer 3: From Tool to Partner

This is the most important shift. Without a thinking framework, AI is an "advanced calculator" — you input, it outputs. With SOMA, AI becomes a "thinking partner" — it challenges your assumptions, offers perspectives you hadn't considered, and reminds you of relevant historical cases.

A Zero Entropy Think Tank user's reflection: "Before, chatting with AI felt like consulting a dictionary. Now it feels like talking to a seasoned mentor — it asks questions back, probes deeper, and raises dimensions I never even considered."

5. What This Means for Everyday People

You might think: I'm not an AI developer — what does SOMA have to do with me? The answer: **it's redefining your relationship with AI.**

You Don't Need to Become a Prompt Engineer

There's a hidden barrier with today's AI tools: you have to learn "how to ask questions properly." The same question asked by different people yields vastly different results. SOMA's significance is that it lets the AI learn "how to think properly" — rather than depending on the user's prompting skills. You don't need to study "prompt engineering." SOMA fills in the thinking gap for the AI.

You Get "Thinking," Not Just "Information"

When you ask a medical question, a SOMA-powered AI won't just give you a list of symptoms. It will use First Principles to break down the causes, Systems Thinking to reveal connections between lifestyle and symptoms, and Inversion to help rule out misdiagnoses. You get structured analysis, not search results.

Your AI Gets to Know You

SOMA's memory and evolution mechanisms mean that over time, the AI gradually understands your thinking preferences, your domain needs, and your preferred analytical angles. It becomes not a cold, generic tool but a growing, personalized intellectual companion.

6. What This Means for the AI Industry

From "Bigger Models" to "Smarter Models"

For the past two years, the AI industry's main theme has been "go big" — more parameters, more training data, more compute. This path is hitting diminishing returns: the marginal benefit of larger models is declining while costs are skyrocketing.

SOMA represents a different path: **don't change the model itself — change how the model thinks**. It adds a lightweight thinking framework layer that produces qualitative leaps in existing models — just as giving the same person different thinking training dramatically changes their output.

This approach has three advantages:

- **Low cost:** No model retraining needed, no additional GPUs required
- **Explainable:** You can see which wisdom laws the AI used and why it analyzed things that way
- **Evolvable:** Wisdom laws auto-tune based on real-world effectiveness

The Paradigm Shift: AI From Tool to Sage

AI development can be viewed in three eras:

- **Era 1.0:** AI can answer questions (ChatGPT emerges)
- **Era 2.0:** AI can use tools and execute tasks (the Agent era)
- **Era 3.0:** AI can think in structured ways and self-evolve (SOMA's vision)

SOMA v1.1.1 is a small but solid milestone in Era 3.0. It proves that **wisdom is not a function of model size — it's a function of how structured your thinking is.**

Open Source & Local-First

SOMA adheres to two principles: **fully open source** and **local-first**. This means any developer, any enterprise can deploy SOMA on their own servers with zero cloud dependency and complete data privacy. In an era of growing concern about data sovereignty and AI safety, this choice carries weight.

7. By the Numbers

| Metric | Value | Context |
|------------|-------|--|
| Test Cases | 618 | Full coverage of all core features — all passing |

| | | |
|---------------------|---------------|---|
| Wisdom Laws | 7 | Covering causality, systems, contradiction, efficiency, inversion, evolution, and analogy |
| LLM Support | 10+ models | DeepSeek, Kimi, Qwen, Claude, GPT, Gemini, and more |
| Multi-Agent Speedup | 4.9× | 5 experts: from 502ms to 102ms |
| Vector Search Speed | 0.1 ms | Pinpoint retrieval across 100K+ memories |
| Memory Capacity | 100K+ entries | Scalable to hundreds of millions |
| Package Size | 192 KB | Ultra-lightweight, installs in seconds |

8. The Architecture — Five Concentric Layers

SOMA's architecture can be understood as five concentric layers:

Innermost: Memory Layer — Episodic memory (what happened) + Semantic memory (knowledge graph relationships) + Scene aggregation + User profiles

Second: Framework Layer — 7 wisdom laws + auto weight adjustment + reasoning templates + hypothesis testing

Third: Reasoning Layer — Problem decomposition → Memory activation → Anti-bias search → Reasoning framework → Causal extraction

Fourth: Evolution Layer — Reflection logging → Quality assessment → Weight auto-tuning → Trigger word discovery → New law mining

Outermost: Collaboration Layer — Multi-agent registration → Expert routing → Parallel dispatch → Consensus formation → Distributed evolution

Each layer is an independent, pluggable module. You can use just the memory layer to build a smart knowledge base, or activate all five layers to create a complete AI sage system.

9. The Name "SOMA"

SOMA comes from ancient Sanskrit, meaning "wisdom drink" — a sacred beverage in Vedic tradition said to grant insight and enlightenment. In Aldous Huxley's *Brave New World*, SOMA is a substance that keeps people content and clear-minded. We chose this name hoping the project can become a "wisdom drink" for AI — giving machines the structured thinking capability of human sages.

SOMA also stands for "**S**tructured **O**ntology for **M**etacognitive **A**ugmentation."

10. The Road Ahead

SOMA's roadmap is clear and restrained:

Short-term (v1.2.x): The Zhongdao Engine — when the AI becomes overly fixated on a particular thinking framework, auto-correct to maintain balanced, flexible thinking.

Medium-term (v2.0.0): Autonomous Cognitive Loop — a full closed loop of "Perceive → Remember → Reason → Evolve → Act," enabling the AI to self-dialogue and think even without an external LLM.

But one bottom line will never change: **SOMA will always be open source, always local-first, and never locked to any specific LLM.** It is a public good for the community, not owned by any single company.

11. Try It Now

Install (one line):

```
pip install soma-wisdom
```

5-minute quickstart: github.com/sunyan999999/soma

Already deployed in: The Zero Entropy Think Tank (零熵智库), providing sage-level thinking assistance to its users.

SOMA v1.1.1 · Open-Source Wisdom Framework · Apache 2.0 License

Community-driven. Contributions, feedback, and experience reports are always welcome.