Deterministic Governance Primitives for Autonomous Research Agents

Warren Smith
OMEGA Protocol Research, Bristol, UK
www.omegaprotocol.org

Abstract

Autonomous ML agents can implement experiments but lack scientific discipline. They rediscover known effects, promote single-seed results as findings, and grind through variations without controlling for compute. We present OMEGA Research, a deterministic governance runtime that enforces the scientific method on autonomous agents through four layers: an experiment protocol that validates proposals before compute is spent, an evidence graph that tracks experiments and their relationships, a claim lifecycle that enforces monotonic promotion through replication and variance gates, and a hash-chained registry that provides tamper-evident records. We evaluate the system using the "Bigger Hidden Size Trap," a canonical scenario derived from Karpathy's documented failure modes of multi-agent research setups, and show that all seven failure modes are caught programmatically. The runtime is deterministic: given identical inputs and policies, it produces identical decisions and identical hashes. It uses no neural network in its governance path. The implementation includes 45 unit tests and 7 integration tests covering all governance paths. Code and interactive demo are available at omegaprotocol.org.

1. Introduction

The deployment of LLM-powered agents for autonomous ML research is accelerating. Frameworks such as Agent Laboratory (Schmidgall et al., 2025) and numerous open-source multi-agent research setups now allow agents to generate hypotheses, implement experiments, and report results with minimal human intervention.

However, the ability to run experiments is not the ability to do science. Karpathy (2025) recently described deploying an 8-agent research setup (4 Claude, 4 Codex) to run ML experiments on nanochat:

"The agents' ideas are just pretty bad out of the box, even at highest intelligence. They don't think carefully through experiment design, they run a bit non-sensical variations, they don't create strong baselines and ablate things properly, they don't carefully control for runtime or flops. Just as an example, an agent yesterday 'discovered' that increasing the hidden size of the network improves the validation loss, which is a totally spurious result given that a bigger network will have a lower validation loss in the infinite data regime."

This is not an intelligence failure . It is a governance failure. The agents are capable implementers but have no mechanism to distinguish genuine discoveries from statistical artifacts, uncontrolled confounds, or known effects. The scientific method exists precisely to prevent these errors in human research. We argue it must be enforced computationally for autonomous agents.

We contribute OMEGA Research, a deterministic governance runtime that sits between the agent and the compute, enforcing four properties:

1. Proposal validation. No experiment runs without a validated hypothesis, controlled baseline, and justified compute budget.
2. Replication enforcement. No finding advances without multiple runs across different random seeds, with variance below a configurable threshold.
3. Monotonic claim graduation. Findings progress through Observation → Replicated → Baseline-Beating → Graduated, with deterministic requirements at each gate. The final promotion requires human approval.
4. Tamper-evident records. Every experiment, result, and claim is hash-chained using SHA-256 with canonical JSON serialisation.

The system is implemented in Python with Pydantic for schema enforcement, uses JSONL for append-only storage, and has zero external dependencies beyond the standard library and Pydantic. It is deterministic: no LLM, no probabilistic component, no neural network appears in the governance path.

2. Related Work

Experiment tracking systems such as MLflow (Zaharia et al., 2018) and Weights & Biases (Biewald, 2020) track what experiments were run and their results. OMEGA Research is complementary: it governs what experiments are *allowed* to run, and what claims can be made from results.

AI safety and governance frameworks address the behaviour of deployed AI systems. The "Assured Autonomy" framework (Xie et al., 2025) argues that increasing agent autonomy demands more formal structure and explicit constraints. OMEGA Research instantiates this principle for the specific domain of scientific experimentation.

Automated ML research systems including Agent Laboratory (Schmidgall et al., 2025), The AI Scientist (Lu et al., 2024), and BioPlanner (O'Donoghue et al., 2023) focus on enabling agents to do more research. OMEGA Research is orthogonal: it constrains *how* that research is conducted.

Reproducibility infrastructure such as Kapoor & Narayanan's (2024) "AI Agents That Matter" argues for rigorous evaluation standards. OMEGA Research provides computational enforcement of the standards these works advocate.

To our knowledge, no existing system provides deterministic, hash-chained enforcement of the scientific method for autonomous agents.

3. System Architecture

OMEGA Research consists of four layers, each independently testable and composable.

3.1 Experiment Protocol

The protocol layer validates experiment proposals before any compute is spent. Proposals are Pydantic models with required fields: hypothesis, experiment type, baseline configuration, and proposed configuration.

Scaling Governor. When a proposed configuration increases parameter count or FLOP budget beyond a configurable threshold, the scaling governor fires. It requires a ScalingHypothesis with a declared type (scaling study, architecture change, or capacity requirement), a justification, and, for scaling studies, at least one intermediate scale point and inclusion of the baseline scale.

3.2 Evidence Graph

The evidence graph is a typed, append-only registry of experiments, results, and directed edges. Edges represent relationships: supports, contradicts, or inconclusive. The graph is stored as JSONL with atomic writes. Each record includes a record_hash computed from its canonical JSON representation and a prev_hash linking to the previous record, forming a hash chain.

3.3 Claim Lifecycle

Claims progress through four tiers with deterministic promotion requirements:

Tier	Requirements
Observation	Initial finding from at least one experiment
Replicated	≥3 runs with different seeds; CV < threshold (default 10%); std < improvement magnitude
Baseline-Beating	Effect size above threshold (default 0.05); direction matches hypothesis; ablation satisfied
Graduated	All prior requirements hold; no contradicting edges; human approval with identified approver

The graduation gate is the critical governance boundary. No agent can promote a claim to Graduated. A human chief scientist must review the evidence and provide their identity. The system automates discipline, not judgement.

3.4 Trust Layer

Every record is hash-chained: fields are serialised to canonical JSON following RFC 8785 (JSON Canonicalization Scheme) , with sorted keys, deterministic formatting, no whitespace ambiguity. SHA-256 is computed, and the hash is stored alongside a prev_hash linking to the previous record. Modifying any field produces a different hash. The chain cannot be repaired without recomputing every downstream hash.

3.5 Implementation Notes

The system is implemented in Python using Pydantic V2 (Rust-based core) for high-performance schema validation. All governance logic executes out-of-band . The rules governing the agent are stored and enforced in a separate runtime from the agent's own reasoning context. This prevents the agent from reasoning about or circumventing its governance constraints, in the same way that a laboratory's safety protocols exist independently of any individual researcher's judgement.

4. Evaluation: The Bigger Hidden Size Trap

We evaluate OMEGA Research using a canonical integration test that reproduces the exact failure modes Karpathy documented.

#	Scenario	Result	Layer
1	Naive scaling (no hypothesis)	BLOCKED	Scaling Governor
2	Missing intermediate scales	VALIDATION ERROR	Pydantic Schema
3	Proper scaling study	ALLOWED	Scaling Governor
4	Single run promotion	BLOCKED	Replication Gate
5	High variance (CV=20.7%)	BLOCKED	Variance Check
6	Clean graduation	GRADUATED	Lifecycle + Human
7	Tamper with result	CHAIN BROKEN	Trust Layer

All seven scenarios pass as automated integration tests. The test suite contains 45 unit tests covering individual components and 7 integration tests covering the full trap scenario.

4.1 What the Trap Proves

The Bigger Hidden Size Trap is not a synthetic benchmark. It is a reproduction of a documented failure mode in a real multi-agent research setup. The specific failure, an agent "discovering" that bigger models have lower loss, is simultaneously technically correct (larger models do have lower loss), scientifically meaningless (it tells you nothing about the architecture), and computationally wasteful (it consumed GPU time to confirm a known property).

An ungoverned agent cannot distinguish this from a genuine finding. OMEGA Research blocks it at Step 1, before any compute is spent.

5. Design Decisions

Determinism over intelligence. The governance runtime contains no LLM, no neural network, no probabilistic reasoning. Given identical inputs and policies, it produces identical outputs. An LLM-based governance layer would introduce the same unreliability it aims to prevent.

Out-of-band enforcement. The governance runtime executes in a separate process from the agent. The agent cannot inspect, modify, or reason about its governance constraints. This is analogous to the distinction between a researcher and their institution's ethics board. The board operates independently of any individual researcher's preferences.

Schema as governance. Pydantic models serve as both data validation and policy enforcement. A ScalingHypothesis with intermediate_scales: list[int] and a @model_validator is simultaneously a type system, a documentation layer, and a governance mechanism.

Human-in-the-loop by design. The graduation gate exists because the system governs discipline, not truth. A claim that passes all statistical checks may still be scientifically uninteresting. Only a human can make that judgement.

Append-only evidence. The JSONL registry is append-only by design. Experiments cannot be deleted. This mirrors the laboratory notebook convention: you cross out, you don't erase.

6. Limitations and Future Work

The current thresholds (10% CV, 3 minimum runs, 0.05 effect size) are defaults appropriate for transformer training. Other domains may require different values. The system does not prove that a hypothesis is true; it enforces procedural validity. A claim that graduates has passed replication, variance, and effect size gates with human approval, but this is evidence of rigour,

not evidence of truth. The governance runtime assumes that reported metrics are faithfully computed by the agent's execution environment. The system governs proposals and claims but does not verify that experiments were actually run as proposed. A controlled study comparing governed vs. ungoverned agent research pipelines would strengthen the contribution. Multi-agent coordination protocols for preventing redundant experiments across agents remain future work.

7. Conclusion

Autonomous ML agents need governance, not just capability. OMEGA Research provides this through four composable layers: experiment protocol, evidence graph, claim lifecycle, and trust layer. The system is deterministic, hash-chained, and human-gated. It catches the exact failure modes documented in real multi-agent research setups, blocking spurious experiments before compute is wasted and preventing noise from graduating to findings.

The core claim is simple: the scientific method is computable, and it should be computed. When agents can run experiments autonomously, the discipline that human scientists internalise through training must be enforced through code. OMEGA Research is that code.

The system is open source. An interactive demo is available at omegaprotocol.org/research-governance.html.

References

Biewald, L. (2020). Experiment tracking with Weights & Biases. Weights & Biases. https://wandb.ai

Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., & Meger, D. (2018). Deep reinforcement learning that matters. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1).

Kapoor, S., Stroebl, B., Siegel, Z. S., Nadgir, N., & Narayanan, A. (2024). AI agents that matter. arXiv:2407.01502.

Karpathy, A. (2025). Commentary on multi-agent ML experimentation failures. X (formerly Twitter). https://x.com/karpathy/status/2027521323275325622

Lu, C. et al. (2024). The AI Scientist: Towards fully automated open-ended scientific discovery. arXiv:2408.06292.

O'Donoghue, M. et al. (2023). BioPlanner: Automatic evaluation of LLMs on protocol planning in biology. arXiv:2310.10632.

Pineau, J., Vincent-Lamarre, P., Sinha, K., Lariviere, V., Beygelzimer, A., d'Alche-Buc, F., Fox, E., & Larochelle, H. (2021). Improving reproducibility in machine learning research. Journal of Machine Learning Research, 22(164), 1-20.

Schmidgall, S. et al. (2025). Agent Laboratory: Using LLM agents as research assistants. arXiv:2501.04227.

Xie, Y. et al. (2025). Assured autonomy: How operations research powers and orchestrates generative AI systems. arXiv:2512.23978.

Zaharia, M. et al. (2018). Accelerating the machine learning lifecycle with MLflow. IEEE Data Engineering Bulletin, 41(4), 39–45.