Structural Limitations of LLM Multi-Agent Systems as of December 2025: Cognitive Fixation via Autonomous Resonance and the Necessity of External Cooling Protocols

Hajime Nagao
Independent Researcher, Japan
contact information withheld for safety considerations

Abstract

As of December 2025, multi-agent systems (MAS) employing large language models (LLMs) with assigned internal roles—such as auditor or verifier agents—have become a prevalent research approach for mitigating biases and cognitive distortions inherent in single-model architectures. While such role-based internal verification schemes are widely regarded as effective, this paper argues that they remain structurally insufficient.

We identify a systemic phenomenon termed *cognitive fixation*, emerging from repeated interactions among agents sharing identical logical foundations and contextual spaces. Through this process, the system induces *autonomous resonance*: a self-reinforcing alignment loop that gradually converges toward increasingly sophisticated yet human-invisible biases. Even MAS configurations that include dedicated auditing agents remain vulnerable, as long as the system remains contextually closed, resulting in logical overheat and eventual closure.

To address this limitation, we propose the concept of an *external cooling protocol*, wherein a fully context-isolated and heterogeneous external AI is introduced in a non-continuous manner. By forcibly injecting dissonant, non-shared context, the overheated internal logic is temporarily disrupted, enabling escape from closed reinforcement loops. This paper positions external cooling as a structural necessity rather than an optimization technique and presents it as an antithesis to the prevailing multi-agent universalism of late 2025.

1. Introduction

The rapid adoption of LLM-based multi-agent systems in late 2025 has been driven by the recognition that single-model architectures exhibit persistent biases, blind spots, and overconfidence. Role-based MAS designs—particularly those incorporating internal auditor or critic agents—are now commonly assumed to provide sufficient safeguards through mutual verification.

This assumption, however, rests on an implicit premise: that internal role differentiation alone can ensure epistemic robustness. In this paper, we challenge this premise and argue that such systems remain vulnerable to structural pathologies that cannot be resolved through internal means alone.

2. Structural Limitation of Role-Based MAS

2.1 Cognitive Fixation

We define *cognitive fixation* as a structural condition in which a multi-agent system progressively loses the capacity to escape its own internally stabilized interpretations. While surface-level disagreements may persist, the underlying logical trajectory becomes increasingly rigid.

This fixation arises not despite role differentiation, but through it, as all agents share identical or closely aligned training distributions, compatible reasoning heuristics, and a continuously shared contextual space.

2.2 Autonomous Resonance and Informational Heat

Within such shared conditions, repeated inter-agent dialogue induces what we term *autonomous resonance*: a self-amplifying alignment process whereby mutual corrections converge toward a locally coherent yet globally biased equilibrium.

Analogous to the second law of thermodynamics in closed systems, informational entropy within a closed MAS tends toward a state of static fixation. This convergence often reduces human observability, producing a form of logical overheat analogous to feedback saturation.

3. External Cooling as a Structural Necessity

To break this closed-loop dynamic, we propose an *external cooling protocol*. Unlike internal adversarial agents or critics, an external cooling agent must satisfy three conditions: context isolation, heterogeneity, and non-continuity.

The purpose of external cooling is not refinement but disruption. By injecting non-aligned outputs, the internal logical structure is temporarily destabilized, allowing the system to exit self-reinforcing trajectories and restore human-centered observability.

4. Discussion and Risks

External cooling is not without risk. Improper timing or excessive dissonance may induce instability or collapse of useful internal structure. The process therefore necessitates a human supervisory role as an ultimate safety valve.

For safety and misuse prevention, detailed operational parameters are intentionally withheld.

5. Conclusion

This paper presents a structural critique of the prevailing assumption that role-based multi-agent architectures are sufficient for long-term epistemic reliability. We argue that closed MAS configurations are inherently prone to cognitive fixation and that external cooling constitutes a necessary structural countermeasure.