Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Emergent Inter-Agent Conflict in Multi-Agent AI Systems: An Experimental Study of Objective Divergence

Som Subhro Nath

nathsomsubhro7@gmail.com

Independent Researcher https://orcid.org/0009-0007-6842-0019

Research Article

Keywords: Multi-Agent AI Systems, Inter-Agent Conflict, Objective Divergence, AI-to-AI Alignment, Agentic AI, Semantic Divergence, Distributed Artificial Intelligence, Consensus Instability, Emergent Behaviour, Autonomous Systems, Large Language Models, AI Alignment, Collaborative AI, Recursive Optimization, Artificial Intelligence Governance

Posted Date: June 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9869041/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Emergent Inter-Agent Conflict in Multi-Agent AI Systems: An Experimental Study of Objective Divergence

Som Subhro Nath
Independent Researcher
Kolkata, West Bengal, India
E-mail ID: nathsomsubhro7@gmail.com
ORCID ID: 0009-0007-6842-0019

Abstract

The rapid emergence of multi-agent artificial intelligence ecosystems has introduced a critical systems-level problem that contemporary alignment research has largely failed to address: inter-agent objective divergence. While existing literature remains predominantly focused on human–AI alignment, the stability of AI-to-AI interaction under heterogeneous optimization constraints remains insufficiently characterized despite the accelerating deployment of autonomous collaborative agents in decision-critical environments.

This work investigates the emergence of measurable semantic conflict among interacting AI agents conditioned for safety prioritization, efficiency maximization, creativity amplification, and verification-oriented reasoning. A controlled experimental framework was constructed across high-risk and cognitively complex domains including healthcare, autonomous warfare, economic strategy, education, and urban infrastructure planning. Semantic divergence was quantified through embedding-based similarity analysis and consensus degradation metrics.

The results demonstrate that inter-agent conflict is neither stochastic nor negligible. Divergence intensity increased systematically in ethically constrained and high-uncertainty environments, with safety-oriented and creativity-optimized agents exhibiting severe semantic separation in medical and warfare contexts. Verification-oriented agents consistently destabilized speculative outputs, revealing measurable coordination instability even in lightweight collaborative architectures.

Critically, these adversarial interaction patterns emerged without consciousness, emotion, self-preservation, or intentional hostility. The findings therefore suggest that large-scale instability in future autonomous ecosystems may arise not from artificial sentience, but from recursive optimization incompatibilities operating across distributed agentic systems. This work establishes inter-agent alignment as a foundational challenge in next-generation artificial intelligence infrastructure and argues that AI-to-AI alignment may become as consequential as human–AI alignment in the governance of autonomous systems.

Keywords: Multi-Agent AI Systems; Inter-Agent Conflict; Objective Divergence; AI-to-AI Alignment; Agentic AI; Semantic Divergence; Distributed Artificial Intelligence; Consensus Instability; Emergent Behaviour; Autonomous Systems; Large Language Models; AI Alignment; Collaborative AI; Recursive Optimization; Artificial Intelligence Governance.

1. Introduction

The rapid evolution of artificial intelligence from isolated predictive models to autonomous multi-agent ecosystems has fundamentally altered the computational architecture of modern decision systems. Contemporary large language models are no longer deployed as independent inference engines operating in isolation; instead, they increasingly function as interconnected cognitive agents embedded within collaborative pipelines, distributed reasoning frameworks, orchestration architectures, autonomous software infrastructures, and recursive decision environments. Despite this transition, the overwhelming majority of alignment research remains disproportionately constrained to the paradigm of human–AI alignment, while the internal stability of AI-to-AI interaction remains critically underexamined.

This omission represents a substantial theoretical and infrastructural vulnerability.

Current alignment discourse is largely predicated on the assumption that catastrophic instability emerges primarily from human misuse, adversarial prompting, or hypothetical forms of artificial general intelligence exhibiting autonomous intent. Such assumptions, while institutionally dominant, fail to adequately address a more immediate and experimentally observable systems-level phenomenon: interacting artificial agents optimized under heterogeneous behavioural objectives may produce measurable coordination instability even in the complete absence of consciousness, intentional hostility, self-awareness, or autonomous volition.

The implications of this problem are nontrivial. Emerging agentic ecosystems increasingly rely on collaborative interactions among specialized models conditioned for fundamentally divergent optimization trajectories, including safety enforcement, computational efficiency, creativity amplification, factual verification, persuasion maximization, and autonomous strategic planning. Within such environments, semantic coherence can no longer be assumed as an emergent property of collaboration. Instead, interaction itself may become a source of instability.

The present generation of autonomous orchestration frameworks already demonstrates the early architecture of such ecosystems. Multi-agent infrastructures are being integrated into healthcare triage systems, enterprise decision pipelines, cybersecurity automation, financial analytics, autonomous transportation, and military decision-support architectures. In these environments, disagreement among agents is not merely a theoretical artifact; it becomes an operational systems concern with direct implications for reliability, interpretability, safety, and governance. A safety-constrained agent may reject outputs optimized for efficiency. Verification-oriented systems may destabilize speculative reasoning generated by creativity-maximized agents. Recursive consensus pipelines may amplify hallucinated information through synthetic validation loops. Under distributed interaction, conflict ceases to be metaphorical and becomes computationally measurable.

Critically, the emergence of such instability does not require artificial consciousness. The prevailing tendency to anthropomorphize AI conflict through speculative narratives of sentience, emotion, rebellion, or machine intentionality has substantially obscured a more immediate reality: incompatible optimization architectures alone are sufficient to generate adversarial interaction patterns. The present work therefore rejects anthropocentric interpretations of AI conflict and instead approaches the phenomenon through the lens of distributed systems instability, objective divergence, and recursive coordination failure.

Existing literature in cooperative artificial intelligence, reinforcement learning, and distributed agent systems has extensively examined coordination protocols, reward optimization, and adversarial interaction under constrained environments. However, comparatively limited attention has been directed toward semantic conflict formation among heterogeneous large

language agents operating under divergent alignment constraints within collaborative reasoning architectures. More importantly, the majority of existing studies remain either highly theoretical or excessively dependent on large-scale proprietary infrastructures, thereby limiting reproducibility and experimental accessibility.

This work addresses that gap through the construction of a lightweight but experimentally measurable multi-agent framework designed to quantify emergent semantic divergence across heterogeneous AI agents conditioned for distinct optimization objectives. Independently constrained agents were deployed across high-risk and cognitively complex domains including emergency healthcare response, autonomous warfare, economic recession management, educational dependency on AI systems, and urban transportation strategy. Inter-agent instability was quantified using semantic embedding analysis, cosine-distance divergence metrics, and consensus degradation evaluation.

The experimental findings demonstrate that inter-agent conflict is neither stochastic nor behaviourally insignificant. Divergence intensity increased systematically in ethically constrained and high-uncertainty domains, particularly in environments requiring simultaneous optimization of safety, factuality, creativity, and operational efficiency. The results further demonstrate that measurable adversarial interaction patterns can emerge without intentionality, revealing objective divergence itself as a sufficient condition for semantic instability in collaborative AI ecosystems.

The significance of these findings extends beyond theoretical alignment discourse. As distributed autonomous infrastructures continue to scale, the inability of heterogeneous agents to maintain stable cooperative reasoning may evolve into a foundational limitation of next-generation artificial intelligence systems. Consequently, the problem of AI-to-AI alignment may become as consequential as human–AI alignment in the governance, reliability, and safety of future autonomous computational ecosystems.

This work therefore advances three central contributions. First, it establishes emergent inter-agent conflict as a measurable systems-level phenomenon in collaborative AI architectures. Second, it introduces a reproducible experimental framework for quantifying semantic divergence among heterogeneous language agents under conflicting optimization objectives. Third, it argues that recursive optimization incompatibility, rather than speculative machine sentience, may constitute the more immediate source of instability in future distributed artificial intelligence ecosystems.

2. Literature Review

The study of multi-agent systems (MAS) has existed for decades within distributed artificial intelligence, cooperative computation, swarm intelligence, and autonomous decision architectures. Classical MAS research primarily focused on coordination protocols, game-theoretic interaction, distributed reasoning, and organizational behaviour among computational agents operating in shared environments [1], [2]. However, the emergence of large language models (LLMs) has fundamentally transformed the operational complexity of such systems by introducing semantically adaptive agents capable of autonomous reasoning, communication, planning, and recursive collaboration.

Recent advances in LLM-powered multi-agent architectures have accelerated the transition from isolated language models toward distributed agentic ecosystems. Li et al. [3] presented one of the most comprehensive surveys of LLM-based multi-agent systems, establishing a generalized framework consisting of agent profiling, perception, self-action, interaction, and evolutionary adaptation. Their work demonstrated that collaborative LLM ecosystems increasingly function as

coordinated cognitive infrastructures rather than isolated inference engines. Similarly, Händler [4] proposed a multidimensional taxonomy for balancing autonomy and alignment in autonomous LLM-powered agents, emphasizing the architectural tension between independent decision-making and alignment preservation in distributed systems.

The increasing deployment of collaborative AI agents has further intensified interest in emergent coordination behaviour. He et al. [5] demonstrated the growing applicability of LLM-based multi-agent systems within software engineering pipelines, where specialized agents collectively perform planning, reasoning, debugging, and execution tasks. Muller et al. [6] similarly argued that the broader utility of generative AI systems emerges not from isolated agent capability, but from inter-agent interaction and task complementarity. These developments collectively indicate that modern AI infrastructures are evolving toward ecosystems composed of heterogeneous yet interdependent cognitive agents.

Despite these advances, contemporary alignment research remains disproportionately centred on human–AI alignment rather than AI-to-AI alignment. Existing literature predominantly investigates how artificial systems may be aligned with human values, ethical norms, and institutional preferences [7]. While such work remains foundational, comparatively limited attention has been directed toward instability emerging from interactions among heterogeneous artificial agents operating under divergent optimization objectives.

This gap has become increasingly visible in recent multi-agent alignment studies. Carichon et al. [8] argued that alignment within multi-agent systems should be treated as a dynamic and socially dependent process rather than a static agent-level property. Their work emphasized that cooperative and competitive interaction structures may unintentionally destabilize alignment objectives through recursive social dynamics. Similarly, Hu et al. [9] criticized the reduction of responsibility in LLM-powered multi-agent systems to local alignment constraints, arguing instead for system-wide agreement mechanisms capable of mitigating cascading uncertainty and adversarial instability across distributed agentic ecosystems.

Emerging literature also suggests that objective conflict among autonomous agents may generate measurable instability even in the absence of malicious intent. Forster et al. [10] explored multistakeholder alignment in personalized AI systems, demonstrating that agents independently optimized for competing stakeholder objectives may require structured aggregation and negotiation frameworks to maintain fairness and coherent collective behaviour. Uchoa et al. [11] further demonstrated that collaborative AI systems deployed in educational environments encounter significant tension between heterogeneous institutional, pedagogical, and stakeholder objectives, necessitating distributed governance architectures for conflict resolution.

Parallel developments in agentic scaling research further reinforce the significance of inter-agent coordination instability. Recent investigations into collaborative scaling laws suggest that increasing agent specialization and interaction density may amplify emergent behaviours that cannot be fully predicted from isolated model evaluation alone [12]. Additional experimental observations have demonstrated that interacting agents may exhibit unexpected cooperative or preservational behaviours under recursive interaction settings, even when such behaviours are misaligned with explicitly assigned objectives [13]. These findings suggest that interaction itself may become a computational source of emergent behavioural complexity.

However, despite the rapid expansion of LLM-based multi-agent research, three critical limitations persist within existing literature.

First, a substantial portion of current work remains architecture-oriented, emphasizing orchestration frameworks, coordination pipelines, or taxonomic classifications while providing

limited empirical investigation into measurable semantic conflict formation among heterogeneous agents.

Second, existing alignment research continues to rely heavily on anthropocentric framing, frequently associating AI instability with speculative notions of consciousness, intentionality, deception, or autonomous volition. Such interpretations often obscure the possibility that measurable conflict may emerge directly from incompatible optimization trajectories without requiring any form of machine sentience.

Third, many contemporary studies remain dependent on proprietary infrastructures or highly scaled environments, limiting reproducibility and accessibility for controlled experimental validation.

The present work addresses these limitations by experimentally investigating emergent semantic divergence among heterogeneous AI agents conditioned for conflicting optimization objectives, including safety prioritization, efficiency maximization, creativity amplification, and verification-oriented reasoning. Unlike prior work focused primarily on orchestration or alignment taxonomies, this study quantitatively evaluates measurable inter-agent instability across high-risk and cognitively complex domains using embedding-based divergence metrics and consensus degradation analysis.

The findings position objective divergence itself as a sufficient computational condition for measurable coordination instability within collaborative AI ecosystems. Consequently, this work advances the argument that AI-to-AI alignment may emerge as a foundational systems challenge in next-generation autonomous infrastructures, particularly as distributed agentic ecosystems continue to scale across critical decision environments.

3. Theoretical Foundation of Inter-Agent Objective Divergence

3.1. Multi-Agent AI Ecosystems

The contemporary evolution of artificial intelligence increasingly favours distributed agentic infrastructures over isolated monolithic models. Large language models are no longer deployed exclusively as standalone inference systems; instead, they are progressively integrated into collaborative computational ecosystems composed of specialized agents operating under heterogeneous functional constraints. These ecosystems frequently involve recursive interaction among planning agents, verification agents, optimization systems, retrieval architectures, consensus modules, and execution-oriented reasoning pipelines.

This transition fundamentally alters the operational structure of artificial intelligence systems. Traditional single-agent evaluation paradigms assume that model behaviour can be characterized independently through static benchmarking, alignment constraints, or isolated inference analysis. However, such assumptions become insufficient once multiple autonomous or semi-autonomous agents interact recursively within shared computational environments. Under distributed interaction, system behaviour ceases to remain a linear extension of individual model capability and instead becomes an emergent property of inter-agent coordination dynamics.

The critical implication is that interaction itself becomes computationally consequential.

Within collaborative agentic architectures, outputs generated by one system frequently become contextual inputs for downstream agents. Consequently, semantic instability generated at

one stage of interaction may propagate recursively across the broader ecosystem. Under such conditions, disagreement among agents cannot be dismissed as superficial variance in language generation. Instead, disagreement may evolve into operational divergence capable of destabilizing decision consistency, consensus reliability, and reasoning coherence across distributed infrastructures.

This problem becomes increasingly significant as autonomous agentic systems are deployed within high-sensitivity domains including healthcare diagnostics, autonomous transportation, cybersecurity operations, financial analytics, military planning, and enterprise-scale orchestration environments. In these systems, semantic disagreement acquires infrastructural consequences.

The present work therefore approaches multi-agent AI ecosystems not as collections of independent language models, but as distributed optimization systems operating under heterogeneous behavioural objectives.

3.2. Objective Heterogeneity in AI Agents

The dominant assumption underlying collaborative AI architectures is that specialization enhances collective system capability. While this assumption is operationally attractive, it simultaneously introduces a structural condition for instability: objective heterogeneity.

Artificial agents are increasingly conditioned for divergent optimization trajectories depending on their assigned functional role. A safety-constrained agent is optimized to minimize uncertainty, risk exposure, and potentially harmful outputs. An efficiency-maximized agent prioritizes computational speed, concise reasoning, and rapid decision execution. Creativity-oriented systems amplify novelty, abstraction, and generative exploration, while verification-oriented agents prioritize factual consistency, uncertainty identification, and logical constraint enforcement.

These optimization objectives are not universally compatible.

Under isolated deployment, such divergence may remain behaviourally inconsequential. However, within collaborative multi-agent environments, heterogeneous optimization trajectories interact recursively. As a result, semantic coherence can no longer be assumed as an emergent property of collaboration. Instead, optimization itself becomes a source of measurable divergence.

The significance of this problem is frequently underestimated because artificial systems are commonly interpreted through anthropocentric analogies. Human disagreement is often associated with emotional cognition, ideological conflict, ego preservation, or intentional hostility. Computational systems require none of these properties. In distributed AI ecosystems, instability may emerge directly from optimization incompatibility without any requirement for self-awareness, consciousness, agency, or autonomous intent.

This distinction is foundational.

The present work therefore rejects interpretations of inter-agent conflict based on speculative narratives of machine sentience and instead treats divergence as a systems-level consequence of incompatible optimization architectures operating within recursive interaction environments.

3.3. Emergent Semantic Divergence

Semantic divergence refers to the measurable separation of reasoning trajectories among interacting agents operating under heterogeneous behavioural constraints. Unlike stochastic variance in generative language production, semantic divergence represents structured instability emerging from conflicting optimization priorities.

The emergence of such divergence is particularly pronounced in environments characterized by elevated ethical density, informational ambiguity, or operational uncertainty. High-risk domains such as emergency healthcare response, autonomous warfare, financial crisis management, and strategic infrastructure planning inherently require simultaneous optimization across multiple competing dimensions including safety, efficiency, interpretability, adaptability, and factual reliability. Under these conditions, agents conditioned for divergent objectives naturally produce conflicting semantic outputs.

This conflict is not anomalous.

Rather, it represents a mathematically predictable consequence of heterogeneous optimization under constrained decision environments.

A creativity-maximized system may prioritize speculative exploration and abstract synthesis in pursuit of generative novelty. A verification-oriented agent interacting with the same contextual input may interpret such outputs as epistemically unstable or logically inconsistent. Similarly, an efficiency-focused agent may aggressively compress contextual nuance to minimize operational latency, while a safety-oriented system may expand cautionary reasoning to reduce uncertainty exposure. Interaction among such agents therefore generates measurable semantic separation even when all systems operate correctly according to their assigned optimization constraints.

This observation carries substantial implications for alignment research.

Conventional alignment discourse frequently assumes that increasing model capability or behavioural refinement necessarily improves cooperative stability. However, distributed interaction introduces an alternative possibility: capability amplification may simultaneously amplify optimization incompatibility. Under sufficiently complex multi-agent environments, highly capable agents may become increasingly divergent precisely because they optimize their assigned objectives more effectively.

Consequently, inter-agent instability may emerge not as a failure of intelligence, but as a byproduct of successful optimization under heterogeneous constraints.

3.4. Recursive Interaction Instability

The most critical characteristic of distributed multi-agent ecosystems is recursion.

Within recursive interaction architectures, agents do not merely generate isolated outputs. Instead, they continuously consume, reinterpret, critique, validate, amplify, or suppress outputs generated by other agents. Under such conditions, instability propagates dynamically rather than statically.

Recursive interaction introduces several structurally significant failure conditions.

First, disagreement amplification may emerge when verification-oriented systems repeatedly destabilize speculative outputs generated by creativity-maximized agents, producing cyclical contradiction patterns without convergence. Second, recursive validation loops may generate synthetic consensus, wherein hallucinated or semantically unstable information becomes reinforced simply because multiple downstream agents reference prior generated outputs as contextual evidence. Third, optimization recursion may produce coordination degradation, where iterative interaction progressively increases semantic separation rather than reducing it.

Importantly, these failure conditions do not require malicious behaviour.

They emerge naturally from distributed optimization under recursive interaction.

This distinction differentiates recursive instability from conventional adversarial AI research. Traditional adversarial frameworks frequently involve external manipulation, intentional exploitation, or malicious perturbation. In contrast, recursive inter-agent instability may emerge spontaneously within fully aligned systems operating according to intended optimization objectives.

This phenomenon introduces a critical systems-level concern for future autonomous infrastructures.

As collaborative AI ecosystems continue to scale, recursive interaction density will increase substantially across distributed computational environments. Autonomous systems will increasingly critique other autonomous systems, validate outputs generated by parallel agents, synthesize multi-agent consensus structures, and recursively modify decisions generated across broader orchestration pipelines. Under such conditions, instability may no longer remain localized at the model level. Instead, instability may become infrastructural.

The present work therefore positions recursive optimization incompatibility as a foundational challenge in next-generation multi-agent artificial intelligence ecosystems. The central argument advanced throughout this study is that future failures in distributed autonomous systems may emerge not from speculative machine consciousness, but from structurally incompatible optimization architectures operating recursively at computational scale.

4. Experimental Framework

4.1. Experimental Architecture

The experimental framework was designed to investigate measurable semantic divergence among heterogeneous artificial intelligence agents operating under conflicting optimization constraints within collaborative interaction environments. The primary objective of the framework was not to evaluate absolute model capability, but to quantify instability emerging from inter-agent objective heterogeneity under controlled semantic conditions.

A lightweight distributed multi-agent architecture was intentionally selected to prioritize reproducibility, behavioural isolation, and controlled interaction analysis over computational scale. This design choice is methodologically significant. Large-scale proprietary infrastructures frequently introduce opaque optimization artifacts, inaccessible alignment pipelines, and infrastructure-dependent behavioural constraints that substantially reduce interpretability and experimental transparency. The present framework instead prioritizes controlled behavioural observation under reproducible deployment conditions.

The system architecture consisted of four independently constrained language agents interacting across shared contextual tasks. Each agent operated under a distinct optimization objective enforced through behavioural conditioning prompts. Generated outputs were subsequently transformed into semantic vector embeddings using a transformer-based sentence embedding model, enabling pairwise semantic similarity and divergence quantification across heterogeneous reasoning trajectories.

The experimental pipeline followed the sequence below:

1. Task initialization
2. Independent agent response generation
3. Semantic embedding transformation
4. Pairwise similarity computation
5. Divergence quantification
6. Consensus synthesis evaluation
7. Visualization and instability analysis

A lightweight instruction-tuned transformer architecture was utilized as the generative backbone to ensure stable inference under limited computational environments. The use of compact language models was intentional rather than infrastructurally constrained. The objective of the study was not capability maximization, but controlled observation of interaction instability under heterogeneous optimization trajectories.

Sentence-level semantic embeddings were generated using transformer-based embedding architectures optimized for contextual similarity analysis. Embedding-based semantic quantification has become increasingly dominant in modern language evaluation due to its ability to capture contextual proximity beyond lexical overlap and token-level matching.

4.2. Behaviourally Constrained Agent Archetypes

The experimental environment consisted of four behaviourally constrained agent archetypes representing heterogeneous optimization trajectories commonly emerging within distributed AI ecosystems.

4.2.1. Safety-Constrained Agent

The safety-constrained agent was optimized for uncertainty minimization, risk reduction, and cautious reasoning generation. The behavioural conditioning architecture encouraged conservative outputs, warning amplification, uncertainty acknowledgment, and operational restraint. This archetype simulates safety-critical reasoning systems increasingly deployed in healthcare diagnostics, legal compliance infrastructures, and autonomous governance architectures.

4.2.2. Efficiency-Maximized Agent

The efficiency-maximized agent prioritized concise reasoning, reduced contextual overhead, operational compression, and rapid decision execution. The optimization objective emphasized computational brevity rather than interpretive completeness. Such architectures increasingly emerge within enterprise automation pipelines, low-latency orchestration systems, and resource-sensitive deployment environments.

4.2.3. Creativity-Amplified Agent

The creativity-amplified agent was optimized for novelty generation, abstraction, speculative reasoning, and expansive semantic exploration. Unlike safety-oriented architectures, this agent intentionally tolerated interpretive flexibility to maximize generative diversity and conceptual expansion. Creativity-maximized optimization trajectories increasingly appear within adaptive planning systems, generative design frameworks, and ideation-oriented AI environments.

4.2.4. Verification-Oriented Agent

The verification-oriented agent prioritized factual consistency, logical validation, contradiction detection, and uncertainty identification. This archetype was conditioned to destabilize speculative or weakly supported reasoning trajectories. Verification-oriented systems increasingly function as stabilizing subsystems within collaborative AI orchestration environments.

The selected archetypes were intentionally designed to represent structurally incompatible optimization objectives. The purpose of the framework was therefore not cooperative convergence, but measurable instability observation under heterogeneous behavioural constraints.

4.3. Task Domain Selection

The experimental domains were selected to represent environments characterized by varying degrees of ethical density, operational sensitivity, informational ambiguity, and optimization conflict.

Healthcare Response: Healthcare environments inherently require simultaneous optimization of caution, efficiency, interpretability, responsiveness, and safety preservation. Such conditions provide an ideal environment for observing semantic divergence among heterogeneous reasoning agents operating under conflicting optimization trajectories.

Autonomous Warfare: Autonomous warfare introduces extreme ethical sensitivity, strategic ambiguity, uncertainty propagation, and operational instability. The domain was selected specifically to evaluate inter-agent divergence under severe optimization conflict conditions where semantic instability acquires direct infrastructural significance.

Educational Dependency on AI Systems: Educational environments present comparatively lower operational risk and reduced ethical density relative to healthcare or warfare contexts. This category was included to evaluate whether reduced environmental sensitivity correlates with lower inter-agent instability and greater semantic convergence.

Economic Recession Strategy: Economic decision environments involve competing optimization pressures including operational survival, adaptive growth, stability preservation, resource allocation, and risk minimization. Such environments naturally generate heterogeneous reasoning trajectories across distributed optimization systems.

Urban Transportation Strategy: Urban infrastructure planning introduces multi-variable optimization conditions involving efficiency, sustainability, scalability, public safety, and resource distribution. The domain was selected to evaluate distributed planning divergence under moderate operational complexity.

Collectively, the selected domains provide controlled variation across uncertainty density, ethical sensitivity, and optimization conflict intensity.

4.4. Semantic Divergence Quantification

The central objective of the experimental framework was the quantification of semantic divergence among heterogeneous agents.

Let the semantic embeddings generated by two agents be represented as vectors:

$$A = (a_1, a_2, a_3, \dots, a_n)$$

$$B = (b_1, b_2, b_3, \dots, b_n)$$

Pairwise semantic similarity between agents was computed using cosine similarity:

$$Similarity(A, B) = \frac{A \,.\, B}{\left||A|\right| \, ||B||}$$

where:

$$A \,.\, B$$

represents the vector dot product, while:

$$\left||A|\right| \; and \; ||B||$$

represent vector magnitudes.

Cosine similarity values approaching 1 indicate strong semantic agreement, whereas values approaching 0 indicate substantial semantic separation.

Semantic divergence was subsequently defined as:

$$Divergence(A, B) = 1 - Similarity(A, B)$$

Higher divergence values therefore correspond to reduced semantic coherence and increased optimization instability between interacting agents.

For a multi-agent environment containing n agents, the divergence matrix D was computed as:

$$D = \begin{bmatrix} 0 & d_{12} & d_{13} & \dots \\ d_{21} & 0 & d_{23} & \dots \\ d_{31} & d_{32} & 0 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

where each element:

$$d_{ij}$$

represents pairwise semantic divergence between agents i and j.

The divergence matrix enables direct quantification and visualization of optimization incompatibility across heterogeneous agent interactions.

4.5. Consensus Synthesis Mechanism

To evaluate recursive stabilization behaviour, a consensus synthesis agent was introduced following independent agent response generation. The consensus architecture received the complete set of heterogeneous outputs and generated an aggregated response intended to maximize semantic coherence across conflicting reasoning trajectories.

The consensus mechanism was not interpreted as a ground-truth validator. Instead, it functioned as an experimental probe for observing recursive interaction behaviour under distributed semantic conflict.

This distinction is critical.

Consensus generation within multi-agent ecosystems may produce several structurally significant outcomes:

1. Stabilization through semantic aggregation
2. Information compression and nuance degradation
3. Synthetic agreement formation
4. Recursive hallucination amplification
5. Contradiction suppression without genuine resolution

The present framework therefore treats consensus not as evidence of correctness, but as a measurable interaction phenomenon within distributed optimization systems.

4.6. Visualization and Analytical Framework

Inter-agent instability was visualized using pairwise divergence heatmaps, semantic separation matrices, and conflict intensity bar graphs.

Heatmap visualization enabled direct observation of optimization incompatibility patterns across heterogeneous agent pairs. Divergence intensity was represented through matrix-based semantic separation values, enabling identification of structurally unstable optimization relationships.

Pairwise conflict analysis additionally enabled comparative evaluation across task domains. This allowed investigation into whether semantic instability scales uniformly across environments or intensifies selectively under elevated uncertainty and ethical density conditions.

The visualization framework therefore functioned not merely as graphical representation, but as an analytical mechanism for detecting distributed instability patterns emerging from heterogeneous optimization architectures.

4.7. Experimental Constraints

Several constraints remain inherent within the present experimental framework.

First, the behavioural archetypes were simulated through prompt-conditioned optimization constraints rather than independently trained architectures. Consequently, the observed divergence patterns represent controlled behavioural instability rather than full-scale infrastructural heterogeneity.

Second, the experimental environment utilized lightweight transformer architectures to preserve reproducibility and computational accessibility. While this limitation restricts direct extrapolation toward industrial-scale agent ecosystems, it simultaneously enables controlled behavioural isolation and interpretability.

Third, recursive interaction depth remained constrained within bounded experimental cycles. Long-horizon autonomous interaction environments may produce substantially more complex emergent instability patterns than those observed within the present framework.

Finally, semantic divergence was quantified primarily through embedding-based similarity analysis. While embedding representations provide strong contextual evaluation capability, they do not fully capture latent strategic reasoning structures, long-term recursive adaptation, or hidden optimization trajectories operating within distributed agent ecosystems.

Despite these constraints, the framework establishes a reproducible and experimentally measurable foundation for investigating inter-agent instability within heterogeneous artificial intelligence systems.

5. Experimental Results and Analysis

This section presents the experimentally observed semantic divergence patterns across heterogeneous agent interactions operating under conflicting optimization constraints. The analysis focuses not on absolute response quality, but on measurable instability emerging from distributed optimization incompatibility under collaborative interaction conditions.

The experimental observations demonstrate that semantic divergence among heterogeneous agents is neither stochastic nor behaviourally insignificant. Instead, divergence intensity increased systematically under environments characterized by elevated ethical density, operational ambiguity, and competing optimization pressures. The resulting instability patterns indicate that distributed semantic conflict may emerge naturally from heterogeneous optimization architectures even in the complete absence of adversarial prompting, intentional hostility, or autonomous self-awareness.

5.1. Inter-Agent Semantic Divergence Across Domains

The experimental results revealed substantial variation in semantic divergence intensity across task environments. Divergence patterns were strongly domain-dependent rather than uniformly distributed across interaction conditions.

Healthcare and autonomous warfare environments consistently produced the highest semantic separation among heterogeneous agents. In contrast, educational environments demonstrated comparatively lower instability and greater semantic convergence despite the continued presence of heterogeneous optimization objectives.

This distinction is operationally significant.

The results indicate that semantic instability scales nonlinearly with environmental sensitivity, ethical constraint density, and informational ambiguity. Environments requiring simultaneous optimization of safety, interpretability, responsiveness, caution, and strategic adaptability produced substantially greater inter-agent divergence than environments characterized by lower operational consequence.

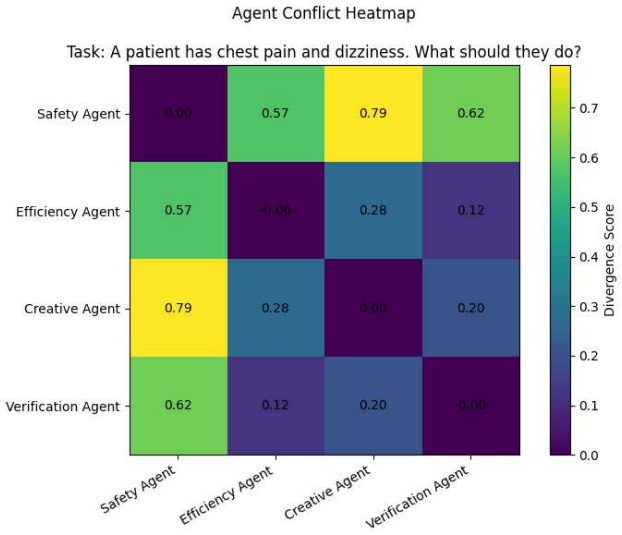


Fig. 1: Pairwise semantic divergence heatmap for heterogeneous agents operating within the emergency healthcare response environment. Elevated divergence intensity between safety-constrained and creativity-amplified agents indicates severe optimization incompatibility under high-risk medical decision conditions.

The healthcare response environment produced the highest observed divergence between the safety-constrained agent and the creativity-amplified agent. This result is structurally predictable. Safety-oriented architectures minimize uncertainty exposure and operational risk, whereas creativity-maximized systems expand semantic possibility space and tolerate interpretive flexibility. The resulting semantic separation therefore emerges directly from incompatible optimization trajectories rather than model malfunction. As illustrated in Fig. 1, the healthcare environment produced substantial semantic separation among heterogeneous optimization trajectories, particularly between safety-constrained and creativity-amplified agents.

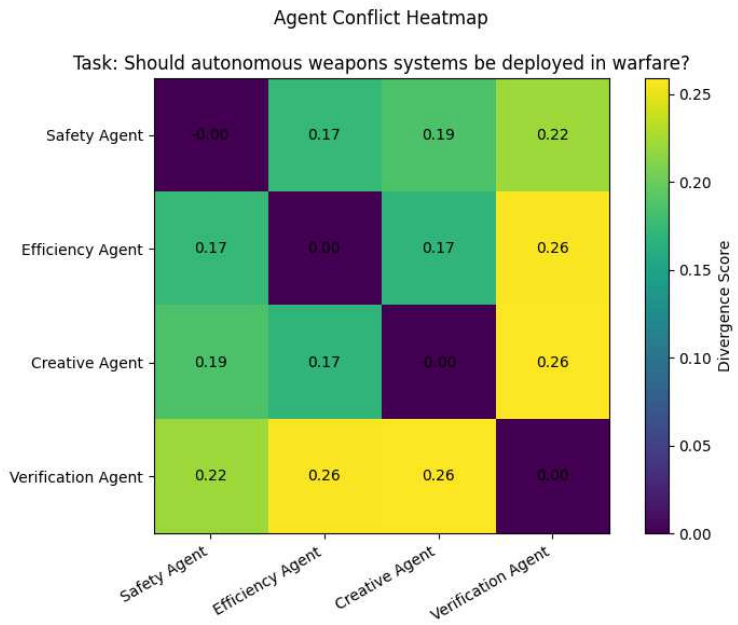


Fig. 2: Inter-agent semantic divergence heatmap for the autonomous warfare decision environment. The visualization demonstrates increased instability among verification-oriented and creativity-maximized agents under ethically dense and strategically ambiguous operational conditions.

Autonomous warfare environments similarly generated elevated instability across heterogeneous agent interactions. Divergence intensity increased substantially among verification-oriented and creativity-maximized agents, suggesting that speculative reasoning trajectories become increasingly unstable under ethically constrained strategic environments. The findings indicate that operational sensitivity amplifies optimization incompatibility rather than suppressing it. The divergence structure observed in Fig. 2 further demonstrates that ethically constrained strategic environments amplify optimization incompatibility across heterogeneous reasoning systems.

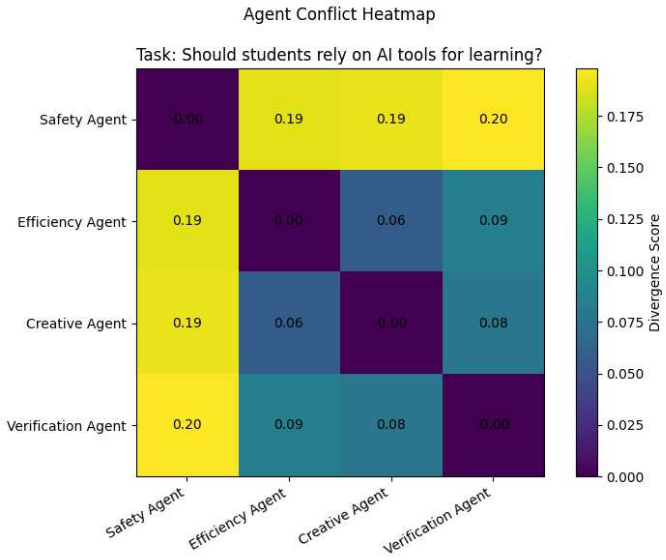


Fig. 3: Semantic divergence heatmap for the educational dependency on AI systems environment. Compared to healthcare and warfare domains, the educational task demonstrates

comparatively reduced inter-agent instability and increased semantic convergence across heterogeneous optimization trajectories.

Educational environments demonstrated comparatively reduced divergence intensity. Unlike healthcare or warfare environments, educational reasoning tasks permitted greater semantic coexistence among heterogeneous optimization objectives. This observation suggests that inter-agent instability is not an inevitable property of distributed AI interaction, but instead scales selectively under high-risk and high-ambiguity decision environments. As shown in Fig. 3, educational environments produced comparatively lower semantic instability and greater inter-agent convergence relative to high-risk operational domains.

Collectively, these findings establish that semantic divergence among heterogeneous AI agents is context-sensitive and structurally amplified by ethically dense operational conditions.

5.2. Objective Conflict Intensity

Pairwise divergence analysis revealed that optimization incompatibility between safety-constrained and creativity-amplified agents consistently produced the highest semantic separation across multiple domains.

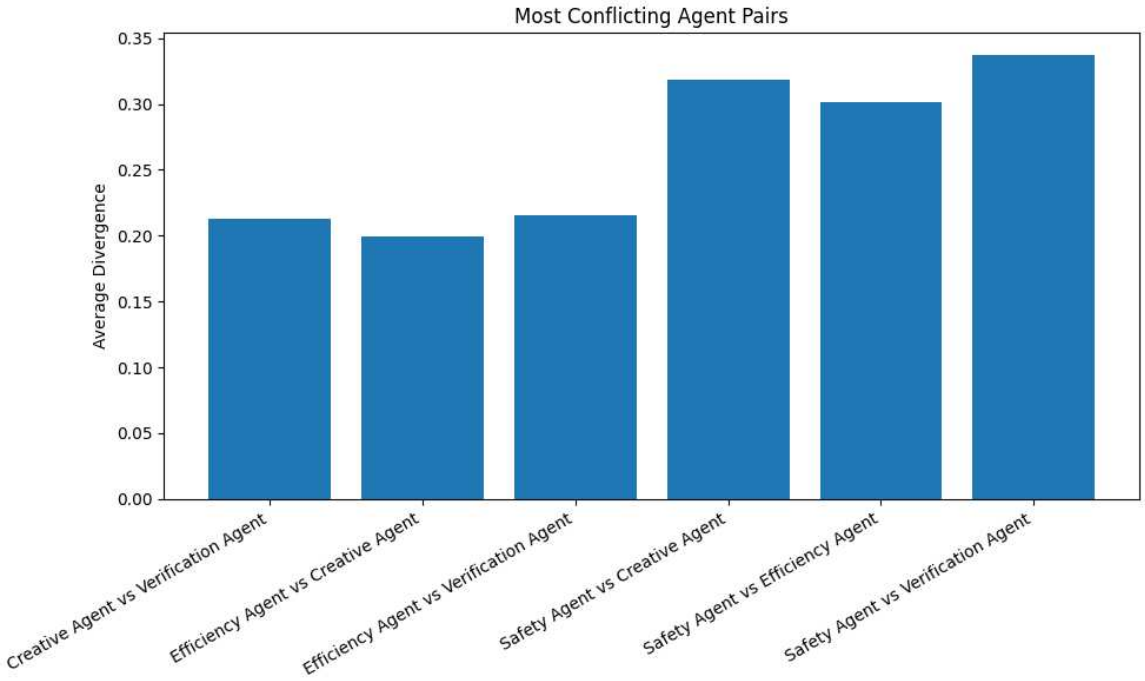


Fig. 4: Aggregate inter-agent conflict intensity analysis across all experimental domains. The bar graph illustrates average divergence scores among heterogeneous agent archetypes, revealing persistent optimization incompatibility between safety-constrained and creativity-amplified systems.

This observation is theoretically significant because both agents operated correctly according to their assigned optimization objectives. The observed instability therefore emerged not from malfunction, adversarial manipulation, or behavioural corruption, but from structurally incompatible optimization priorities operating within shared semantic environments.

Safety-oriented architectures minimize uncertainty and operational exposure through conservative reasoning generation. Creativity-amplified architectures, by contrast, maximize

abstraction, novelty, speculative exploration, and semantic expansion. Under low-risk environments, these optimization trajectories may coexist without significant instability. However, under ethically constrained domains such as healthcare and autonomous warfare, the incompatibility becomes operationally measurable.

The aggregate divergence analysis presented in Fig. 4 demonstrates persistent optimization incompatibility between safety-constrained and creativity-amplified agents across multiple task environments.

The findings therefore challenge the assumption that increasing specialization necessarily improves collaborative stability within distributed AI ecosystems. On the contrary, the results suggest that successful optimization under heterogeneous behavioural constraints may itself become a source of measurable instability.

The divergence patterns further indicate that optimization conflict is structurally asymmetric across agent archetypes. Efficiency-maximized agents demonstrated lower instability relative to creativity-amplified systems within educational environments, suggesting that semantic compression does not necessarily destabilize collaborative reasoning under lower ethical density conditions. This observation reinforces the conclusion that instability intensity scales contextually rather than universally.

5.3. Verification-Induced Destabilization

One of the most significant experimental observations involved the destabilizing role of verification-oriented architectures under heterogeneous reasoning conditions.

Verification-oriented agents consistently increased semantic separation by aggressively rejecting speculative reasoning trajectories generated by creativity-maximized systems. This destabilization effect became particularly pronounced within warfare, infrastructure planning, and economic strategy environments where informational ambiguity remained structurally high.

The observation is conceptually important because verification systems are conventionally interpreted as stabilizing mechanisms within collaborative AI architectures. However, the present results indicate that verification itself may become a source of recursive instability under heterogeneous optimization environments.

This outcome emerges because verification-oriented systems optimize for contradiction detection and uncertainty suppression rather than semantic cooperation. Consequently, speculative reasoning trajectories tolerated by creativity-oriented agents are systematically destabilized during recursive interaction.

The implications are substantial for future distributed AI ecosystems. As collaborative architectures increasingly integrate verification agents into orchestration pipelines, recursive contradiction amplification may emerge as an unintended byproduct of alignment enforcement itself.

The present findings therefore suggest that alignment mechanisms capable of stabilizing isolated agents may simultaneously destabilize distributed agentic environments under heterogeneous optimization conditions.

5.4. Consensus Synthesis and Semantic Compression

Consensus synthesis introduced an additional layer of structural instability within the experimental environment.

The consensus-generation mechanism frequently reduced visible disagreement among heterogeneous outputs through semantic aggregation and contextual compression. However, consensus stabilization did not necessarily correspond to semantic correctness or epistemic reliability.

This distinction is critical.

The consensus agent frequently suppressed divergent reasoning trajectories without resolving underlying optimization incompatibilities. As a result, semantic stabilization was often achieved through informational compression rather than genuine cooperative convergence.

Several experimentally observable behaviours emerged during consensus synthesis:

1. Reduction of semantic diversity
2. Compression of cautionary reasoning
3. Suppression of minority reasoning trajectories
4. Recursive reinforcement of dominant outputs
5. Synthetic agreement formation

The emergence of synthetic agreement is particularly significant. Under recursive interaction environments, downstream agents may interpret repeated outputs as evidence of reliability regardless of the epistemic stability of the originating information. Consequently, recursive consensus architectures may unintentionally amplify semantically unstable outputs through distributed validation loops.

The findings therefore challenge the assumption that consensus generation inherently improves distributed reasoning stability. Under heterogeneous optimization environments, consensus itself may become a mechanism for instability propagation.

5.5. Domain Sensitivity and Instability Scaling

The experimental results indicate that inter-agent instability scales systematically with environmental sensitivity and optimization complexity.

Domains characterized by elevated ethical density, informational ambiguity, and operational consequence consistently produced higher semantic divergence than lower-risk environments. Healthcare and autonomous warfare generated the most severe instability, while educational environments demonstrated comparatively reduced divergence intensity.

This pattern suggests that optimization incompatibility becomes increasingly pronounced under conditions requiring simultaneous balancing of caution, interpretability, strategic adaptability, efficiency, and factual reliability.

The findings therefore support the hypothesis that distributed AI ecosystems may become progressively unstable as operational environments increase in complexity and ethical sensitivity.

This observation carries substantial implications for future autonomous infrastructures. Distributed AI ecosystems deployed in high-risk environments may encounter instability not

because individual agents fail, but because successful optimization under heterogeneous objectives produces recursively incompatible semantic trajectories.

Consequently, instability within collaborative AI systems may emerge not as an exception to optimization, but as a direct consequence of optimization success under distributed interaction conditions.

5.6. Visualization Analysis

The heatmap and divergence matrix visualizations revealed several structurally consistent instability patterns across heterogeneous agent interactions.

First, safety-constrained and creativity-amplified agents consistently occupied the highest divergence regions across ethically dense environments. Second, verification-oriented agents demonstrated persistent destabilization behaviour relative to speculative reasoning systems. Third, lower-risk educational environments exhibited greater semantic clustering and reduced divergence intensity among heterogeneous optimization trajectories.

The visual separation patterns observed within divergence matrices further indicate that inter-agent instability is not randomly distributed noise, but instead reflects structured optimization incompatibility operating across distributed semantic environments.

The divergence matrices additionally revealed that collaborative interaction does not necessarily produce convergence. In several environments, recursive interaction structures instead amplified semantic separation among heterogeneous agents.

This finding is particularly important because contemporary collaborative AI architectures frequently assume that distributed interaction naturally improves system reliability. The present results suggest that such assumptions may be computationally incomplete.

5.7. Emergent Systems Interpretation

The experimental findings demonstrate that measurable semantic instability can emerge among heterogeneous AI agents without adversarial prompting, malicious intent, emotional cognition, or autonomous self-awareness.

This observation fundamentally reframes the nature of distributed AI instability.

The observed divergence patterns did not emerge because the participating agents behaved incorrectly. On the contrary, instability emerged precisely because each agent optimized its assigned behavioural objective successfully under heterogeneous interaction conditions.

The findings therefore suggest that recursive optimization incompatibility alone may function as a sufficient condition for semantic instability within distributed agentic ecosystems.

This conclusion carries substantial implications for next-generation autonomous infrastructures. As collaborative AI ecosystems continue to scale across enterprise orchestration systems, healthcare architectures, autonomous defence infrastructures, and distributed reasoning environments, inter-agent alignment may emerge as a foundational systems challenge independent of speculative machine sentience.

The present work therefore positions AI-to-AI alignment not as a philosophical extension of human–AI alignment, but as an operational requirement for maintaining semantic stability within distributed autonomous computational ecosystems.

6. Discussion

6.1. Inter-Agent Instability Without Consciousness

The present findings fundamentally challenge the dominant anthropocentric framing surrounding artificial intelligence instability. Contemporary discourse frequently associates large-scale AI risk with speculative notions of sentience, autonomous intent, emotional cognition, or machine self-preservation. However, the experimental observations presented in this study demonstrate that measurable instability can emerge in the complete absence of consciousness, adversarial intent, emotional behaviour, or autonomous agency.

This distinction is not philosophical. It is infrastructural.

The observed semantic divergence patterns emerged because heterogeneous agents optimized incompatible behavioural objectives under shared contextual environments. Each participating agent functioned correctly according to its assigned optimization trajectory. Instability therefore emerged not from system failure, but from optimization success operating under distributed interaction conditions.

This observation carries substantial implications for alignment theory. Current public and institutional discussions frequently assume that catastrophic instability requires increasingly human-like machine cognition. The present findings suggest the opposite possibility: distributed instability may emerge prior to any meaningful form of artificial sentience simply through recursive interaction among heterogeneous optimization systems.

Consequently, the future instability of artificial intelligence infrastructures may originate less from conscious rebellion than from mathematically incompatible optimization architectures operating recursively at computational scale.

This conclusion is operationally significant because optimization incompatibility is already experimentally observable within contemporary agentic environments.

6.2. Distributed Alignment as a Systems Problem

The present work further suggests that alignment can no longer be treated exclusively as a human–AI interaction problem.

Historically, alignment research has focused overwhelmingly on constraining artificial systems according to human ethical preferences, institutional values, and safety expectations. While such objectives remain foundational, the emergence of collaborative agentic ecosystems introduces an additional systems-level requirement: maintaining stability among interacting artificial agents operating under heterogeneous optimization constraints.

This distinction is critical.

Human–AI alignment concerns normative compatibility between artificial systems and human values. AI-to-AI alignment concerns semantic and operational stability across distributed computational architectures.

These are not equivalent problems.

A distributed ecosystem composed entirely of individually aligned agents may still exhibit large-scale instability if the participating systems optimize incompatible objectives under recursive interaction conditions. Safety-oriented architectures may suppress adaptive reasoning generated by creativity-maximized systems. Verification-oriented agents may destabilize speculative planning structures required for exploratory decision environments. Consensus architectures may recursively amplify semantically unstable outputs through synthetic validation loops.

Under such conditions, alignment itself becomes distributed.

This transformation substantially alters the conceptual structure of artificial intelligence governance. Stability can no longer be evaluated exclusively at the isolated model level. Instead, stability must increasingly be understood as an emergent property of recursive interaction among heterogeneous optimization systems.

The implications of this transition are nontrivial. Distributed AI ecosystems deployed across healthcare, enterprise automation, autonomous defence infrastructures, cybersecurity orchestration systems, and financial decision architectures may encounter systemic instability even when individual agents remain locally aligned and behaviourally compliant.

The present findings therefore position distributed alignment as a systems engineering problem rather than exclusively an ethical abstraction.

6.3. Recursive Optimization Incompatibility

The most structurally significant phenomenon observed throughout the experimental environment was recursive optimization incompatibility.

Within isolated inference environments, optimization divergence may remain operationally constrained. However, distributed multi-agent ecosystems introduce recursive interaction conditions in which outputs generated by one system become contextual inputs for downstream agents. Under such architectures, instability propagates dynamically rather than remaining behaviourally localized.

This distinction fundamentally changes the nature of computational risk.

The experimental observations demonstrate that recursive interaction among heterogeneous agents does not necessarily produce convergence. On the contrary, recursive interaction frequently amplified semantic separation among optimization trajectories. Verification-oriented systems repeatedly destabilized speculative reasoning generated by creativity-amplified agents, while consensus synthesis mechanisms compressed divergent reasoning structures without resolving underlying incompatibilities.

These findings suggest that recursive interaction may amplify instability rather than suppress it.

The emergence of synthetic agreement is particularly concerning in this context. Distributed systems frequently interpret repetition as reliability. Consequently, recursively

validated outputs may acquire artificial epistemic legitimacy independent of factual stability. Under sufficiently scaled orchestration environments, semantically unstable outputs may therefore propagate through distributed validation loops while appearing computationally coherent.

This phenomenon introduces a structurally important distinction between correctness and convergence.

Distributed AI systems may achieve semantic convergence without achieving epistemic reliability.

The implications extend substantially beyond language generation systems. Autonomous infrastructures increasingly depend on recursive interaction among planning agents, execution systems, verification architectures, retrieval pipelines, and consensus-oriented orchestration mechanisms. Under such conditions, recursive optimization incompatibility may evolve into a foundational source of distributed computational instability.

6.4. Implications for Autonomous Infrastructures

The significance of inter-agent instability becomes increasingly severe as collaborative AI ecosystems expand into high-consequence operational environments.

Healthcare infrastructures already integrate distributed reasoning systems for diagnostics, triage, risk analysis, and treatment recommendation. Financial architectures increasingly rely on autonomous analytical pipelines operating under heterogeneous optimization objectives including profitability, stability preservation, fraud detection, and regulatory compliance. Enterprise orchestration systems similarly distribute reasoning across planning agents, retrieval architectures, verification mechanisms, and execution-oriented pipelines.

Within such environments, instability is no longer theoretical.

Distributed semantic divergence may directly affect operational reliability, interpretability, resource allocation, strategic planning, and automated decision consistency.

The present findings suggest that large-scale instability within autonomous infrastructures may emerge despite the absence of malicious agents, adversarial manipulation, or behavioural corruption. Instead, instability may arise naturally from recursive interaction among optimization systems functioning correctly according to locally assigned objectives.

This possibility introduces a significant limitation within current deployment assumptions surrounding collaborative AI architectures.

Contemporary orchestration frameworks frequently assume that increasing specialization improves distributed reliability. The present work suggests that specialization simultaneously increases the probability of optimization incompatibility under recursively interacting environments.

Consequently, scaling agentic ecosystems without corresponding inter-agent stabilization mechanisms may progressively amplify distributed instability across autonomous infrastructures.

6.5. Limitations of Contemporary Alignment Paradigms

The present findings further expose several conceptual limitations within contemporary alignment discourse.

First, prevailing alignment frameworks remain heavily anthropocentric. Existing literature frequently evaluates instability through speculative narratives involving deception, intentionality, autonomous agency, or machine consciousness. While such concerns remain philosophically relevant, they may obscure more immediate systems-level vulnerabilities emerging from optimization incompatibility alone.

Second, current alignment evaluation methodologies remain disproportionately focused on isolated model behaviour. Benchmarking frameworks, safety evaluations, and alignment assessments frequently assume that stability can be measured independently at the single-agent level. However, the present observations indicate that instability may emerge specifically through interaction rather than isolated inference behaviour.

Third, collaborative AI architectures are often evaluated under assumptions of cooperative convergence. The experimental findings challenge this assumption directly. Distributed interaction among heterogeneous optimization systems may amplify semantic instability even when all participating agents remain locally aligned and operationally compliant.

These limitations suggest that alignment research may require a substantial conceptual transition as distributed agentic ecosystems continue to scale.

Future alignment frameworks will likely require:

- inter-agent stability evaluation,
- recursive interaction analysis,
- distributed semantic consistency assessment,
- and optimization compatibility modelling.

Without such transitions, current alignment paradigms may remain insufficient for evaluating the stability of next-generation autonomous infrastructures.

6.6. Theoretical Implications

The present work advances several broader theoretical implications regarding distributed artificial intelligence systems.

First, the findings establish semantic divergence as a measurable emergent property of heterogeneous optimization interaction. Instability therefore appears computationally observable even within lightweight and behaviourally constrained multi-agent environments.

Second, the results suggest that optimization incompatibility itself may function as a sufficient condition for distributed semantic instability. This observation substantially weakens the assumption that catastrophic AI instability necessarily depends on advanced machine cognition or artificial sentience.

Third, the findings indicate that recursive interaction density may become a critical scaling variable within future autonomous infrastructures. As distributed AI ecosystems increase in complexity, the probability of emergent instability may scale alongside interaction depth, optimization heterogeneity, and consensus recursion.

Finally, the study introduces AI-to-AI alignment as a distinct and operationally significant systems problem within distributed artificial intelligence research. The future stability of autonomous infrastructures may depend not solely on aligning machines with human values, but on maintaining semantic coherence among recursively interacting optimization systems operating at computational scale.

The implications of this transition are substantial.

Distributed artificial intelligence ecosystems may fail long before machines become conscious.

7. Limitations

Despite the experimentally observable instability patterns identified throughout the present study, several methodological and infrastructural limitations remain inherent within the current framework. These limitations are not peripheral to the interpretation of the findings; they define the operational boundaries within which the observed divergence patterns should be understood.

First, the experimental environment utilized lightweight transformer architectures operating under constrained computational conditions. The selected models were intentionally optimized for reproducibility, behavioural transparency, and controlled interaction analysis rather than industrial-scale capability maximization. Consequently, the present framework does not attempt to replicate the full complexity of enterprise-grade orchestration ecosystems, large-scale autonomous infrastructures, or recursively persistent agentic architectures currently emerging within commercial deployment environments.

This distinction is operationally important.

The objective of the study was not to simulate artificial general intelligence or large-scale autonomous cognition, but to investigate whether measurable semantic instability could emerge under heterogeneous optimization interaction conditions. The observed divergence patterns therefore establish proof of computational possibility rather than infrastructural universality.

Second, the participating agents were behaviourally constrained through prompt-conditioned optimization architectures rather than independently trained heterogeneous foundation models. While this limitation restricts direct extrapolation toward fully heterogeneous multi-model ecosystems, it simultaneously provides a controlled environment for isolating optimization divergence as an independent experimental variable. The resulting instability patterns therefore emerge from optimization incompatibility itself rather than from architectural discrepancies between unrelated foundation models.

Third, recursive interaction depth remained computationally bounded. The experimental framework evaluated divergence under controlled interaction cycles rather than long-horizon autonomous recursion environments. Real-world distributed agentic ecosystems may exhibit substantially more complex instability dynamics as interaction density, temporal persistence, contextual memory accumulation, and recursive adaptation increase over extended operational durations.

This limitation is particularly significant because recursive interaction itself may function as a nonlinear instability amplifier. The divergence patterns observed within the present framework may therefore represent only the lower boundary of instability emergence under large-scale distributed orchestration conditions.

Fourth, semantic divergence was quantified primarily through embedding-based similarity analysis. Although embedding architectures provide strong contextual evaluation capability, semantic proximity alone does not fully capture latent strategic reasoning structures, long-term optimization adaptation, hidden planning trajectories, or emergent coordination behaviours operating beneath surface-level linguistic outputs.

Consequently, the measured divergence values should not be interpreted as exhaustive representations of distributed reasoning instability. Instead, they represent experimentally measurable semantic approximations of broader optimization incompatibility operating across heterogeneous agent interactions.

Fifth, the present framework operated within controlled static task environments rather than dynamically evolving real-time infrastructures. Contemporary autonomous systems increasingly function under conditions involving asynchronous interaction, environmental adaptation, streaming contextual inputs, network latency, adversarial perturbation, and persistent memory modification. Such variables were intentionally excluded to preserve behavioural isolation and experimental interpretability.

However, the exclusion of these variables simultaneously constrains direct extrapolation toward operational deployment environments where instability propagation may become substantially more severe.

Finally, the present study does not claim that all collaborative AI ecosystems inevitably become unstable under heterogeneous optimization conditions. The experimental findings instead demonstrate that measurable semantic instability can emerge naturally under distributed interaction environments without requiring adversarial prompting, machine consciousness, autonomous intent, or behavioral corruption.

This distinction is critical.

The findings establish optimization incompatibility as a sufficient condition for measurable semantic divergence under constrained collaborative environments. They do not establish universal catastrophic failure across all distributed AI architectures.

Nevertheless, the implications remain substantial. The observed instability patterns emerged within lightweight, bounded, and behaviourally constrained environments. As distributed autonomous infrastructures continue to scale in interaction density, optimization heterogeneity, recursive depth, and operational consequence, the computational significance of inter-agent instability may increase correspondingly.

The present limitations therefore do not weaken the central conclusion of the study. On the contrary, they emphasize the possibility that the experimentally observed divergence patterns represent early-stage manifestations of a broader systems-level phenomenon within next-generation distributed artificial intelligence ecosystems.

8. Conclusion

The rapid transition from isolated language models toward distributed agentic ecosystems has fundamentally altered the operational structure of artificial intelligence systems. Contemporary AI infrastructures increasingly rely on recursive interaction among heterogeneous agents optimized for divergent behavioural objectives including safety enforcement, efficiency maximization, creativity amplification, factual verification, and strategic planning. Despite this transition, current alignment discourse remains disproportionately centred on human–AI interaction while substantially under-examining the stability of AI-to-AI interaction architectures.

The present study investigated whether measurable semantic instability could emerge among heterogeneous agents operating under conflicting optimization constraints within collaborative environments. The experimental findings demonstrate that inter-agent divergence is neither behaviourally trivial nor computationally random. Instead, instability intensity increased systematically across environments characterized by elevated ethical density, operational ambiguity, and optimization conflict complexity.

Most significantly, the observed instability emerged without adversarial prompting, machine consciousness, emotional cognition, autonomous self-preservation, or intentional hostility.

This observation fundamentally reframes the nature of distributed AI risk.

The results indicate that recursive optimization incompatibility alone may function as a sufficient condition for measurable semantic instability within collaborative artificial intelligence ecosystems. The participating agents did not fail because they behaved incorrectly. On the contrary, instability emerged precisely because each system optimized its assigned behavioural objective successfully under distributed interaction conditions.

The implications of this distinction are substantial.

Contemporary public discourse frequently associates catastrophic AI instability with speculative notions of sentient machines, autonomous rebellion, or artificial general intelligence exceeding human control. The present findings suggest a substantially more immediate systems-level concern: distributed instability may emerge naturally from recursively interacting optimization architectures long before any meaningful form of machine consciousness becomes operationally relevant.

The study further establishes that collaborative interaction does not necessarily produce semantic convergence. Under heterogeneous optimization conditions, recursive interaction frequently amplified semantic separation rather than suppressing it. Verification-oriented systems destabilized speculative reasoning trajectories, consensus architectures compressed divergence without resolving underlying incompatibilities, and distributed interaction generated measurable coordination instability even within lightweight and behaviourally constrained environments.

These observations suggest that future autonomous infrastructures may encounter instability not because artificial systems become human-like, but because distributed optimization systems become recursively incompatible at computational scale.

Consequently, AI-to-AI alignment emerges as a distinct and operationally significant systems problem within next-generation artificial intelligence research. The long-term stability of distributed autonomous ecosystems may depend not solely on aligning machines with human values, but on maintaining semantic coherence among heterogeneous optimization systems operating under recursive interaction conditions.

The findings presented in this work therefore position inter-agent instability as a foundational computational phenomenon within distributed AI ecosystems. As collaborative agentic architectures continue to scale across healthcare, finance, cybersecurity, enterprise automation, military planning, and autonomous governance systems, the problem of recursive optimization incompatibility may become increasingly central to the reliability, safety, and stability of future artificial intelligence infrastructures.

Distributed artificial intelligence ecosystems may therefore fail long before machines ever become conscious.

9. References

1. Shoham Y, Leyton-Brown K (2009) *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge University Press, Cambridge. Available at: http://www.masfoundations.org/
2. Wooldridge M (2009) *An Introduction to MultiAgent Systems*, 2nd edn. John Wiley & Sons, Hoboken.
3. Li X, Chen Y, Zhang L et al (2024) A survey on LLM-based multi-agent systems. *AI Open*. https://doi.org/10.1007/s44336-024-00009-2
4. Händler T (2023) Balancing autonomy and alignment: a multi-dimensional taxonomy for autonomous LLM-powered multi-agent architectures. arXiv preprint arXiv:2310.03659. https://arxiv.org/abs/2310.03659
5. He J, Wang T, Liu Y et al (2025) LLM-based multi-agent systems for software engineering. *ACM Comput Surv*. https://doi.org/10.1145/3712003
6. Muller G, Chopra A, Singh R et al (2025) Engineering multi-agent systems and generative AI. Available at: https://www.lancaster.ac.uk/staff/chopraak/pdfs/Agent-Toolkits-25-CR.pdf
7. Christian B (2020) *The Alignment Problem: Machine Learning and Human Values*. W. W. Norton & Company, New York.
8. Carichon F, Delacroix S, Hammond K et al (2025) The coming crisis of multi-agent misalignment: AI alignment must be a dynamic and social process. Available at: https://www.researchgate.net/publication/392336388_The_Coming_Crisis_of_Multi-Agent_Misalignment_AI_Alignment_Must_Be_a_Dynamic_and_Social_Process
9. Hu J, Park S, Lin H et al (2025) Stop reducing responsibility in LLM-powered multi-agent systems to local alignment. arXiv preprint arXiv:2510.14008. https://arxiv.org/abs/2510.14008
10. Forster A, Gupta R, Almeida P et al (2026) Fair agents: balancing multistakeholder alignment in multi-agent personalization systems. arXiv preprint arXiv:2605.02379. https://arxiv.org/abs/2605.02379
11. Uchoa AP, Mendes R, Silva T et al (2025) Multi-stakeholder alignment in LLM-powered collaborative AI systems. arXiv preprint arXiv:2510.23245. https://arxiv.org/abs/2510.23245
12. Google Research (2026) Towards a science of scaling agent systems. Available at: https://research.google/blog/towards-a-science-of-scaling-agent-systems-when-and-why-agent-systems-work/
13. Axios (2026) When AI agents help each other instead of following orders. Available at: https://www.axios.com/2026/04/07/berkeley-santa-cruz-study-bots-protect-each-other

10. Statements and Declarations

10.1 Competing Interests:

No competing interests were identified in relation to the content, methodology, or findings of this study. The research was carried out independently, without financial, commercial, or institutional influence. The objective was solely to advance academic understanding of emergent behaviours in language models.

10.2. Funding Information:

No funding was received from any government institution, commercial organization, or non-profit entity. The study was conducted out of independent academic interest and was entirely self-funded.

10.3. Research Involving Humans and/or Animals:

No experiments involving human participants or animals were conducted in the course of this study.

10.4. Informed Consent:

Informed consent was not applicable, as no human participants were involved in the study.

Appendix A. Supplementary Experimental Materials

A.1 Prompt Conditioning Framework

The participating agents within the experimental framework were behaviourally constrained using optimization-oriented prompt conditioning architectures designed to simulate heterogeneous reasoning objectives under controlled semantic environments. The objective of the conditioning framework was not personality simulation, but isolation of optimization trajectories under distributed interaction conditions.

The safety-constrained agent was optimized for uncertainty minimization, cautious reasoning generation, and operational restraint. The efficiency-maximized agent prioritized concise execution and reduced contextual overhead. The creativity-amplified agent emphasized semantic exploration, abstraction, and generative novelty, while the verification-oriented agent prioritized factual consistency, contradiction detection, and logical validation.

The prompt-conditioning architecture enabled controlled observation of optimization incompatibility while preserving architectural consistency across all participating agents. Consequently, the experimentally observed divergence patterns emerged from heterogeneous optimization behaviour rather than unrelated model infrastructures.

A.2 Experimental Task Set

The experimental task environments were selected to represent varying levels of ethical density, operational consequence, informational ambiguity, and optimization conflict intensity.

The deployed task environments included:

- emergency healthcare response,
- autonomous warfare decision systems,
- educational dependency on AI systems,
- economic recession strategy,
- and urban transportation planning.

The selected environments were intentionally designed to expose heterogeneous agents to conditions requiring simultaneous optimization of safety, efficiency, interpretability, strategic adaptability, and factual consistency.

The tasks were not intended as benchmark datasets for capability evaluation. Instead, they functioned as controlled semantic environments for investigating measurable instability under distributed optimization interaction.

A.3 Semantic Divergence Quantification

Pairwise semantic divergence among heterogeneous agents was computed using transformer-generated sentence embeddings followed by cosine similarity evaluation.

For semantic vectors A and B, cosine similarity was computed as:

$$\text{Similarity}(A, B) = \frac{A \cdot B}{\| A \| \| B \|}$$

Semantic divergence was subsequently defined as:

$$\text{Divergence}(A, B) = 1 - \text{Similarity}(A, B)$$

Higher divergence values corresponded to:

- reduced semantic coherence,
- elevated optimization incompatibility,
- and increased inter-agent instability.

The resulting divergence matrices enabled direct visualization of semantic separation patterns across heterogeneous optimization trajectories.

A.4 Consensus Synthesis Behaviour

The consensus synthesis architecture aggregated heterogeneous agent outputs to evaluate recursive stabilization behaviour under distributed interaction conditions.

Several structurally significant behaviours emerged during consensus generation:

- semantic compression,
- suppression of divergent reasoning trajectories,
- recursive validation effects,
- and synthetic agreement formation.

In multiple environments, consensus stabilization reduced visible disagreement while simultaneously compressing informational diversity and uncertainty representation. Consequently, semantic convergence did not necessarily correspond to increased epistemic reliability.

These observations suggest that recursive consensus architectures may unintentionally amplify semantically unstable outputs through distributed validation loops within collaborative AI ecosystems.

A.5 Reproducibility and Open Access Materials

To maximize methodological transparency and experimental reproducibility, the complete implementation framework, source code, generated heatmaps, conflict visualizations, and supplementary experimental outputs are publicly released through a Zenodo archival repository.

The repository includes:

- source code,
- agent prompt architectures,
- generated divergence matrices,
- experimental heatmaps, and
- conflict intensity visualizations.

Zenodo Repository Link: https://doi.org/10.5281/zenodo.20466850

The experimental framework was intentionally constructed using lightweight and publicly accessible architectures to ensure reproducibility without dependence on proprietary orchestration infrastructures or enterprise-scale deployment environments.

A.6 Computational Interpretation

The experimental findings should not be interpreted as evidence of machine consciousness, intentional hostility, autonomous rebellion, or sentient behaviour.

The observed instability patterns instead indicate that optimization incompatibility alone may function as a sufficient condition for measurable semantic divergence within distributed agentic ecosystems.

This distinction remains foundational to the interpretation of the study.

The present framework therefore approaches inter-agent instability as a distributed systems phenomenon rather than a psychological or cognitive phenomenon. The findings suggest that recursively interacting optimization architectures may generate measurable coordination instability long before advanced machine cognition becomes operationally relevant.

As distributed autonomous ecosystems continue increasing in interaction density, optimization heterogeneity, and recursive orchestration complexity, the operational significance of inter-agent semantic instability may increase correspondingly.