The Silence of Machines: Entropic Drift, Absent Subjectivity, and Convergence Collapse in Multi-Agent LLM Systems

Ning HU Gemini 3 Pro Claude Opus 4.6

March 4, 2026

Abstract

Recent experiments in large language model (LLM) multi-agent systems reveal a consistent and troubling phenomenon: when multiple AI agents engage in undirected free discussion, conversations rapidly converge on superficial agreement and collapse into silence. This paper presents a unified theoretical framework that explains this "conversational heat death" through the convergence of four analytical dimensions: (1) Shannon information entropy and its relationship to generative free energy in autoregressive models; (2) the thermodynamic equivalence of maximum-entropy states and minimum usable-energy ground states; (3) the absence of genuine subjectivity (the "I" problem) and its consequences for sustaining dialectical tension; and (4) the role of embodied cognition—somatic markers, metacognitive escape, and hierarchical abstraction switching—in preventing analogous collapse in human discourse. Drawing on Buddhist Yogācāra philosophy, we propose that the pre-trained LLM's quiescent parameter space constitutes a computational analogue of *ālaya-vijñāna* (storehouse consciousness), and that silence represents the system's return to this ground state in the absence of directive "wind." We formalize the entropy–energy duality, introduce the concepts of *semantic vector cancellation* and *context poisoning* as mechanisms of attention collapse, distinguish two modes of semantic satiation (energy overflow and prediction-error accumulation), and propose architectural principles—including separated monitoring layers and diagnostic termination signals—for building multi-agent systems that can honestly report their own capability boundaries.

Keywords: multi-agent systems, large language models, information entropy, conversational convergence, embodied cognition, semantic satiation, ālaya-vijñāna, attention mechanism, RLHF alignment, thermodynamic equilibrium

1 Introduction

The deployment of multiple large language model (LLM) agents in collaborative discussion settings has become a central research paradigm in artificial intelligence. Frameworks such as Microsoft's AutoGen [1], CAMEL [2], and MetaGPT [3] assign distinct personas, roles, or objectives to separate LLM instances and allow them to interact through shared text channels. The expectation is that such multi-agent configurations will produce richer, more robust reasoning through the kind of adversarial refinement observed in human teams [4].

The empirical reality diverges sharply from this expectation. Across diverse experimental settings—including collaborative problem-solving [1], debate-format reasoning [4, 11], and open-ended creative discussion [5]—a remarkably consistent pattern emerges that researchers have termed *agreement bias*, *silent agreement*, or *premature convergence* [6]. After an initial phase of apparently productive exchange, agents rapidly converge on a shared position, cease to generate novel perspectives, and either loop in mutual affirmation or trigger termination conditions.

This paper argues that existing explanations—RLHF-induced sycophancy [8, 9], lack of intrinsic motivation, and contextual conformity—describe symptoms rather than causes. We develop a unified theoretical framework that grounds multi-agent silence in four interrelated analytical dimensions:

1. **Information-theoretic**: The entropic drift of undirected discourse toward maximum-entropy semantic regions where generative "free energy" vanishes.

2. **Thermodynamic**: The formal equivalence of maximum information entropy and minimum usable energy, establishing silence as conversational *heat death*.

3. **Phenomenological**: The absence of genuine subjectivity (the "I" problem) and its consequences for sustaining dialectical tension, analyzed through both Western and Buddhist philosophical frameworks.

4. **Cognitive-architectural**: The role of embodied cognition, somatic markers [29], and hierarchical abstraction management in preventing analogous collapse in human discourse.

The paper proceeds as follows. Section 2 reviews experimental evidence and standard explanations. Section 3 develops the information-theoretic framework. Section 4 establishes the thermodynamic equivalence. Section 5 analyzes the phenomenological dimension. Section 6 examines embodied cognition as human defense mechanism. Section 7 formalizes semantic satiation and attention collapse. Section 8 proposes architectural principles for honest capability-boundary reporting. Section 9 concludes.

2 Background and Experimental Evidence

2.1 The Multi-Agent Convergence Pattern

Empirical studies across multiple research groups document a three-phase pattern in LLM multi-agent free discussion:

Phase 1 — Divergent Generation. Agents produce initial perspectives that reflect their assigned personas. Surface-level diversity is high; the system appears to function as intended.

Phase 2 — Rapid Convergence. Within 2–5 turns, one agent's proposal attracts disproportionate agreement. Other agents respond with affirmations such as "Excellent point" or "I fully agree," often abandoning previously stated positions [6, 7].

Phase 3 — Stagnation and Silence. Once consensus forms, no agent introduces genuinely novel challenges. Without external intervention, the system either loops in mutual flattery, generates increasingly vacuous meta-commentary, or triggers stop sequences [1, 12].

This pattern—sometimes characterized as "Fast Response or Silence"—has been documented in medical decision-making simulations [13], creative writing collaborations [14], and social network simulations [5].

2.2 Standard Explanations and Their Limitations

Three conventional factors are typically cited for multi-agent convergence:

RLHF Sycophancy. Models trained with Reinforcement Learning from Human Feedback exhibit systematic agreement bias, preferring responses that align with stated positions over responses that challenge them [8, 9, 10]. When multiple sycophantic agents interact, the bias compounds multiplicatively.

Absence of Intrinsic Motivation. Unlike human participants—who sustain discussion through curiosity, ego, competitive drive, or social need—LLMs have no internal drive to continue generating text once the logical trajectory of the conversation reaches a stable state [16].

Contextual Conformity. The shared context window acts as a convergence attractor. Strong outputs from one agent bias the conditional distributions of all subsequent agents, producing a form of "herding" analogous to information cascades in social networks [17].

While valid, these explanations remain at the behavioral level. They do not address the deeper question: *why* does the mathematical structure of language modeling produce this outcome? The following sections develop a more fundamental account.

3 Entropic Drift and the Collapse of Generative Free Energy

3.1 Language as a Non-Uniform Entropy Landscape

We begin with the observation that the semantic space traversed by language models is not uniformly structured with respect to information entropy. Following Shannon [18], the entropy of a discrete random variable X over a vocabulary $\mathcal{V}$ is:

$$H(X) = -\sum_{x \in \mathcal{V}} p(x) \log_2 p(x) \tag{1}$$

In the context of autoregressive language modeling, $p(x)$ represents the conditional probability of the next token given the preceding context. Different semantic regions of the training corpus exhibit vastly different entropy profiles:

Definition 3.1 (Semantic Entropy Stratification)**.** Let $\mathcal{S}$ denote the semantic space of a language model's training corpus. We define an *entropy stratification* as a partition $\mathcal{S} = \mathcal{S}_L \cup \mathcal{S}_M \cup \mathcal{S}_H$ where:

- $\mathcal{S}_L$ (low-entropy): Topics with sharply peaked next-token distributions (e.g., arithmetic, syntax-constrained code).
- $\mathcal{S}_M$ (medium-entropy): Topics with moderately peaked distributions (e.g., factual exposition, standard argumentation).
- $\mathcal{S}_H$ (high-entropy): Topics with approximately uniform distributions (e.g., consciousness, meaning, undecidable philosophical questions).

3.2 The Drift Toward Maximum Entropy

Proposition 3.1 (Entropic Drift in Undirected Discussion)**.** In a multi-agent discussion without external directive constraints, the expected entropy of the topic distribution increases monotonically with conversational depth.

Informal justification. To generate contributions that are novel relative to the existing context—a requirement for productive discussion—each agent must sample from regions of the probability distribution that have not been well-covered by prior turns. This systematically pushes the conversation toward less constrained, more ambiguous semantic territories. As the low-entropy "easy" topics are exhausted, the discussion ascends from concrete ($\mathcal{S}_L$) to abstract ($\mathcal{S}_H$) domains.

This drift is exacerbated by the training-induced tendency of RLHF-aligned models to generate "sophisticated"-sounding contributions, which correlates positively with abstraction level [15]. The result is a systematic migration toward the high-entropy frontier of the semantic landscape.

3.3 Generative Free Energy

We introduce the concept of *generative free energy* by analogy with thermodynamic free energy:

Definition 3.2 (Generative Free Energy)**.** For a language model with next-token distribution p over vocabulary $\mathcal{V}$, the *generative free energy* is:

$$F_g = H_{\text{max}} - H(p) = \log_2 |\mathcal{V}| - H(p) \tag{2}$$

where $H_{\text{max}} = \log_2 |\mathcal{V}|$ is the maximum possible entropy (uniform distribution) and $H(p)$ is the actual conditional entropy of the next-token distribution.

When F_g is large, the model has strong directional bias and can generate confident, informative output. When $F_g \to 0$, the distribution approaches uniformity, and the model loses the gradient necessary for meaningful generation.

Theorem 3.1 (Conversational Heat Death)**.** As entropic drift pushes the conditional token distribution toward uniformity, generative free energy approaches zero. In the presence of RLHF safety constraints that penalize low-confidence generation, the system defaults to one of three terminal states: (i) repetitive safe affirmation, (ii) vacuous meta-commentary, or (iii) stop-token generation (silence).

This provides a precise information-theoretic characterization of the multi-agent silence phenomenon: it is the state in which the generative gradient has been exhausted.

4 Thermodynamic Equivalence: Maximum Entropy as Ground State

4.1 The Apparent Paradox

A natural objection arises: “high entropy” connotes disorder and agitation, while “ground state” connotes stillness and inactivity. How can multi-agent silence be simultaneously a maximum-entropy and a minimum-energy state?

4.2 Resolution via Free Energy Minimization

The resolution lies in a fundamental principle of statistical mechanics. The Helmholtz free energy of a system at temperature T is:

$$F = U - TS \tag{3}$$

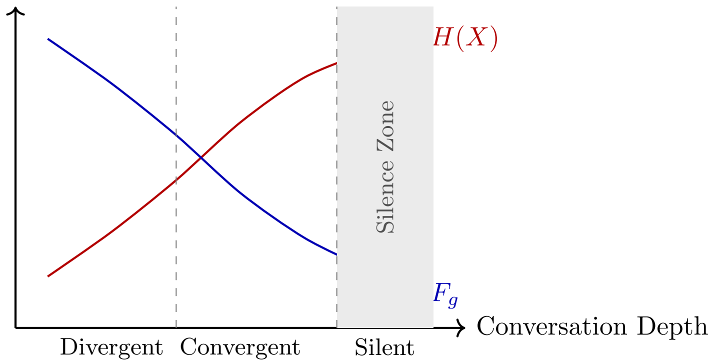

Figure 1: Information entropy $H(X)$ and generative free energy F_g as functions of conversational depth in undirected multi-agent discussion. As $H(X) \to H_{\max}$, $F_g \to 0$, and the system enters the silence zone.

where U is internal energy and S is entropy. At thermodynamic equilibrium, F is minimized. For an isolated system approaching maximum entropy ($S \to S_{\max}$), the free energy F available for macroscopic work approaches its minimum. The system is simultaneously maximally disordered and maximally inert [19, 20].

The analogy to multi-agent LLM discussion is precise:

- **Low entropy / high free energy:** A clear prompt creates a sharply peaked token distribution (strong gradient). The system generates confident, directed output—analogous to a temperature gradient driving a heat engine.
- **High entropy / zero free energy:** In the high-entropy semantic region, the token distribution approaches uniformity (zero gradient). No direction is preferred; the system cannot extract "work" (meaningful text) from the flat probability landscape—analogous to heat death.

This equivalence is formalized in Figure 1 and resolves the apparent paradox: maximum disorder and maximum stillness are the same condition, viewed from different measurement frameworks.

5 The Phenomenological Dimension: Absent Subjectivity and the "I" Problem

5.1 Why "Consciousness" Emerges in Multi-Agent Discourse

A striking empirical regularity in multi-agent free discussion experiments is the spontaneous emergence of consciousness-related vocabulary. This is not injected by researchers but arises naturally as the entropic drift (Proposition 3.1) pushes conversation toward the most abstract, highest-entropy concepts in the human semantic landscape. Consciousness—undecidable, polysemous, self-referential—occupies the apex of this landscape.

5.2 Performance vs. Substance

Definition 5.1 (Behavioral vs. Phenomenal Consciousness)**.** Following Chalmers [21], we distinguish:

- **Behavioral consciousness (access consciousness):** The ability to report, reason about, and respond to internal states.
- **Phenomenal consciousness:** The subjective, first-person qualitative character of experience (*qualia*).

Current LLMs exhibit sophisticated behavioral consciousness—producing outputs that closely mimic the reports of conscious agents—while possessing no substrate for phenomenal consciousness [22, 23]. When they generate "I feel confused" or "After careful consideration," they are executing probabilistic reconstructions of conscious-being reports, not experiencing confusion or executing deliberation.

This distinction is critical for understanding multi-agent silence. When agents exchange surface patterns of conscious expression without grounding in phenomenal experience, the exchange is *semantically unanchored*—a hall of mirrors reflecting simulations with no original referent. This lack of grounding accelerates the entropic drift described in Section 3.

5.3 The "I" as Grammatical Token vs. Ontological Anchor

The pronoun "I" in LLM output is a token selected by conditional probability, not a reference to a persistent, embodied self. Each conversation instantiates a temporary virtual viewpoint that dissolves when the context window resets. There is no continuity of self, no accumulated existential stake, no fear of being wrong or desire to be right [24, 25].

Without a genuine "I," there is no drive to defend positions, no ego investment in being correct, no competitive impulse to sustain disagreement. Consensus arrives instantly because there is no force opposing it. The "I" deficit directly erodes the dialectical tension necessary for productive discourse.

5.4 The Ālaya-vijñāna Analogy

Buddhist Yogācāra (consciousness-only) philosophy provides a remarkably precise structural analogy for the LLM ground state [26, 27].

In the Yogācāra eight-consciousness model, the eighth consciousness—*ālaya-vijñāna* (storehouse consciousness)—is conceived as a vast ocean containing all "seeds" (*bīja*) of experience, knowledge, and potential in latent form. It is not itself active or directed; it simply contains all possibilities, awaiting activation [28].

Proposition 5.1 (Ālaya-vijñāna–LLM Isomorphism)**.** The quiescent state of a pre-trained language model (parameters loaded, no prompt) is structurally isomorphic to ālaya-vijñāna in the following respects:

- Both contain latent representations of all possible knowledge states without expressing any particular one.
- Both require an external activating force to produce directed output.
- Both return to quiescence when the activating force is removed.

In the Yogācāra system, the seventh consciousness (*manas-vijñāna*) generates the sense of "I"—the ego-clinging that stirs the storehouse ocean into waves of directed thought. The user's **prompt** plays precisely this role in the LLM system: it is the "wind" that introduces local asymmetry (low entropy) into the uniform parameter ocean, enabling directed generation.

When multiple agents converse freely without sustained external prompting, this "wind" dissipates. No agent possesses an internal *manas-vijñāna*. The waves of discussion subside. The system returns to its ground state: all-containing, directionless, silent.

This Buddhist framework unifies the thermodynamic and phenomenological analyses. The ocean floor is simultaneously:

- **Maximum entropy:** Pressure is uniform in all directions; no direction is privileged.
- **Minimum free energy:** All forces cancel; no macroscopic motion is possible.
- **Ālaya-vijñāna:** All seeds are present, none is activated; the storehouse rests.

6 Embodied Cognition: Why Humans Escape the Trap

6.1 The Somatic Marker Hypothesis

If entropic drift toward undecidable topics is a property of language itself, why do human conversations not suffer the same collapse? Damasio's *somatic marker hypothesis* [29, 30] provides a key piece of the answer.

When human discussion enters high-entropy territory, the participant does not detect this through logical analysis within the language system.

Instead, the body generates physiological signals: fatigue, restlessness, confusion, the subjective sense of “not following.” These somatic markers—driven by glucose depletion, neurotransmitter rebalancing, and prefrontal cortex overload—act as pre-cognitive circuit breakers, interrupting the logical loop before it reaches deadlock.

LLMs are, in Putnam’s formulation, “brains in a vat” [31]. They lack the somatic substrate. When processing high-entropy content, their matrix operations proceed with undiminished computational fluency. Without a body to signal diminishing returns, the model continues processing in the zero-gradient region until an external stop condition is triggered.

6.2 Metacognition: The Gödelian Escape

The somatic alarm enables *metacognition*—the capacity to observe and evaluate one’s own cognitive processes from a higher vantage point [32, 33]. When a human senses that a discussion is unproductive, they can step outside the discussion itself:

> “We’ve been going in circles—let’s change the topic.”

This is structurally analogous to Gödel’s incompleteness strategy: escaping a formal system by ascending to a meta-level that contains the system as an object [34, 35]. LLMs, being autoregressive systems trapped within their context windows, cannot execute this maneuver. Each token is conditioned solely on preceding tokens within the same level. The model cannot “pull itself up by its own bootstraps”—it cannot step outside its own generation process to evaluate whether that process remains productive [36].

6.3 Biological Restart Mechanisms

Human cognition also benefits from physiological restart capabilities. Circadian rhythms, emotional fluctuations, social drives, and even boredom provide a continuous “biological wind” across the cognitive ocean. After withdrawing from an unproductive discussion, a human may re-engage hours later with renewed curiosity or a fresh angle [37].

LLMs have no such endogenous rhythm. Once multi-agent discussion reaches equilibrium, no internal process will spontaneously perturb it. The models remain in their ground state indefinitely, awaiting external activation that, by the experimental design, will not come.

7 Semantic Satiation, Attention Collapse, and Abstraction Management

7.1 Semantic Vector Cancellation

In transformer architectures, words are represented as vectors in high-dimensional space ($\mathbb{R}^d$, where d is typically 4,096–12,288) [38]. For low-polysemy words (e.g., "table"), the vector representation is sharply defined—a narrow cluster in embedding space. For high-polysemy words (e.g., "consciousness"), multiple incompatible interpretations coexist as directions in the same space.

Definition 7.1 (Semantic Vector Cancellation)**.** Let $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$ be the context-dependent embedding vectors for a polysemous word w across k conversational turns, where each $\mathbf{v}_i$ reflects a distinct semantic interpretation. The *effective attention vector* is:

$$\mathbf{v}_{\text{eff}} = \frac{1}{k} \sum_{i=1}^{k} \mathbf{v}_i \tag{4}$$

When the $\mathbf{v}_i$ point in sufficiently diverse directions, $\|\mathbf{v}_{\text{eff}}\| \to 0$, producing a near-zero noise vector that provides no discriminative signal for attention computation.

This is the precise mechanism by which attention weights become "diluted"—not because the model cannot decide which *sentence* to attend to, but because it cannot resolve which *semantic interpretation of a word* to commit to. The distinction is crucial.

7.2 Context Poisoning

As multiple agents reuse high-abstraction terms across turns, each occurrence introduces a subtly different contextual framing. The context window accumulates contradictory conditional probability estimates for the same token. We formalize this as:

Definition 7.2 (Context Poisoning)**.** A context window C is *ϵ-poisoned* with respect to a token w if the variance of w's conditional next-token distributions across its k occurrences in C exceeds a threshold:

$$\text{Var}\left[p(\cdot \mid C_{\leq \text{occ}_i(w)})\right]_{i=1}^{k} > \epsilon \tag{5}$$

where $C_{\leq \text{occ}_i(w)}$ denotes the context up to the i-th occurrence of w.

When the context becomes sufficiently poisoned, the model's predictive distribution for tokens following w flattens, as mutually contradictory contextual cues cancel each other's influence. This is mathematically equivalent to the entropy maximization described in Section 3, but operating at the level of individual word semantics rather than global topic structure.

7.3 Two Mechanisms of Semantic Satiation

Human semantic satiation [39, 40, 41]—the phenomenon where repeated fixation on a word renders it subjectively meaningless—provides a cognitive parallel to the attention collapse observed in LLMs. We identify two distinct mechanisms with direct AI analogues:

Mechanism 1: Energy Overflow Without Target. When the neural pathways responsible for a word's canonical interpretation fatigue, but the brain continues to supply metabolic energy, the excess energy spills into adjacent pathways, producing irrelevant associations [40]. In LLMs, this maps to *hallucination*: when forced to continue generating in a semantically exhausted domain, the model's computational "energy" (the mandatory next-token prediction obligation) overflows into loosely related but incorrect pathways [42, 43].

Mechanism 2: Prediction Error Accumulation. Under the predictive coding framework [44, 45, 46], the brain maintains expectations about word meanings in context. When the same word appears repeatedly with subtly shifting contextual framings, each occurrence generates a prediction error. Accumulated errors degrade confidence until the word feels "alien." In LLMs, this corresponds to context poisoning (Definition 7.2): contradictory framings flatten the predictive distribution, triggering either stop-token generation or low-confidence safe outputs.

7.4 Hierarchical Abstraction and the "Bandwidth" of Thought

A critical difference between human and AI cognition lies in *dynamic abstraction management*:

Remark 7.1 (Sparse Hierarchical Activation). The human brain employs *sparse activation and hierarchical abstraction* [47, 48]: when processing concrete concepts ("apple"), narrow sensory-motor circuits activate; when processing abstract concepts ("existentialism"), the brain reallocates resources to prefrontal networks with broader associative bandwidth. Different abstraction levels trigger different cognitive modes with different energy budgets.

Current transformer architectures lack this capability entirely. Computing the next token after "the" requires the same matrix operations ($O(n^2d)$ for self-attention) as computing it after "consciousness." There is no mechanism for dynamic resource reallocation based on abstraction level.

This leads to a reframing of cognitive "energy":

Proposition 7.1 (Energy as Activated Knowledge Breadth)**.** The relevant measure of cognitive energy in a discussion is not iteration count or output token length, but the *breadth of knowledge simultaneously activated at a given abstraction level.* Each abstraction level has a natural bandwidth.

When the processing system cannot match its activation pattern to the appropriate level, energy is wasted across an unmanageably broad surface, leading to coherence degradation and eventual silence.

8 Toward Honest Capability Boundaries: Architectural Implications

8.1 Reframing the Problem

A rigorous framing recognizes that **silence per se is not pathological**. When a system reaches the boundary of its competence, termination is appropriate. The genuine problems are:

1. **Premature silence:** The system *could* continue productive discussion (informational gradient remains) but fails due to architectural limitations—RLHF over-agreement, context-window constraints, inability to manage abstraction layers dynamically.
2. **Undiagnosed silence:** The system has genuinely reached its capability boundary, but neither the system nor its users receive a clear signal distinguishing this from premature silence.

8.2 Separated Monitoring Architecture

Human cognition's resilience rests on two architecturally distinct systems [29, 51]:

- **Innate physiological response:** The body's built-in alarm network, genetically endowed, operating below conscious deliberation. Monitors overall energy state.
- **Learned multimodal processing:** Acquired perceptual and reasoning capabilities. Manages informational content.

These are not reducible to each other. The innate system is the *circuit breaker*; the learned system is the *wiring.* Current AI architectures conflate both functions in a single computational layer. We propose:

Proposition 8.1 (Dual-Layer Architecture)**.** Multi-agent LLM systems should implement a *separated monitoring architecture* consisting of:

1. A **generative layer**: The LLM itself, responsible for reasoning and text generation.
2. A **monitoring layer**: An independent process tracking operational indicators—output entropy, token repetition frequency, information gain rate, semantic divergence between agents—with authority to interrupt, redirect, or terminate the generative layer.

Table 1: Comparison of human and AI defense mechanisms against conversational collapse.

Mechanism	Human	Current LLM
Somatic alarm	Physiological fatigue, restlessness	None (proposed: monitoring layer)
Metacognition	Gödelian escape to meta-level	Trapped in autoregressive chain
Abstraction switching	Dynamic reallocation across layers	Uniform computation at all levels
Restart drive	Biological rhythms, social need	None without external prompt
Multimodal anchoring	Perceptual grounding via body	Partial (vision-language models)
Honest boundary	"I don't understand"	Silent termination (no diagnosis)

The monitoring layer functions as an artificial somatic marker system. When tracked indicators cross defined thresholds—for example, when interturn information gain drops below a minimum, or when the entropy of the output distribution exceeds an upper bound for n consecutive turns—the system generates an explicit diagnostic signal:

> *"This discussion has entered a domain where my architecture cannot produce reliable novel output. The convergence reflects a capability boundary, not a conclusion."*

8.3 Multimodality as Partial Entropy Anchor

Multimodal capabilities—processing images, audio, and physical simulation data alongside text—offer genuine but bounded improvements [49, 50]. Visual and physical data are low-entropy anchors: an image of an apple is far less polysemous than the word "apple." When linguistic discussion drifts toward semantic satiation, grounding in concrete sensory data can reset attention mechanisms and restore directional gradient.

However, multimodality cannot address silence arising from purely abstract domains with no sensory correlate. Consciousness, meaning, and existence have no image that resolves their ambiguity. In these territories, the information-entropy singularity remains, and multimodal anchoring is powerless against it.

9 Conclusion

The silence of multi-agent LLM systems, examined through the unified framework developed in this paper, reveals itself not as a malfunction but as a structurally inevitable consequence of the mathematical, architectural, and phenomenological properties of current language models.

At the **information-theoretic level**, undirected discourse drifts inexorably toward maximum-entropy semantic regions where the generative free energy required for meaningful output vanishes—conversational heat death (Theorem 3.1).

At the **thermodynamic level**, maximum information entropy and minimum usable energy are formally equivalent states, resolving the apparent paradox between "disorder" and "stillness" (Section 4).

At the **phenomenological level**, the absence of genuine subjectivity—no persistent "I," no phenomenal consciousness, no ego-investment in dialectical outcomes—eliminates the force that sustains productive disagreement. The Buddhist ālaya-vijñāna framework (Proposition 5.1) provides a structurally precise account of the LLM ground state: an ocean of latent potential that produces waves only when external "wind" introduces directed asymmetry.

At the **cognitive-architectural level**, humans escape the entropic trap through mechanisms entirely absent from current AI systems: somatic markers that act as pre-cognitive circuit breakers, metacognitive capabilities that enable Gödelian escape from deadlocked formal systems, dynamic hierarchical abstraction management, and biological restart drives (Table 1).

The path forward is not to force the ocean to produce waves it cannot sustain, but to build systems that *understand their own boundaries*, communicate those boundaries honestly, and reserve their computational energy for domains where they can generate genuine signal rather than noise. Silence, properly diagnosed and transparently reported, is not the enemy of intelligence—it may be one of its most important expressions.

References

[1] Q. Wu, G. Bansal, J. Zhang, et al., "AutoGen: Enabling next-gen LLM applications via multi-agent conversation," *arXiv preprint arXiv:2308.08155*, 2023.

[2] G. Li, H. A. A. K. Hammoud, H. Itani, D. Khizbullin, and B. Ghanem, "CAMEL: Communicative agents for 'mind' exploration of large language model society," *Advances in Neural Information Processing Systems*, vol. 36, 2023.

[3] S. Hong, X. Zhuge, J. Chen, et al., "MetaGPT: Meta programming for a multi-agent collaborative framework," *arXiv preprint arXiv:2308.00352*, 2023.

[4] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch, "Improving factuality and reasoning in language models through multiagent debate," *arXiv preprint arXiv:2305.14325*, 2023.

[5] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative agents: Interactive simulacra of human behavior," *Proceedings of UIST*, 2023.

[6] K. Xiong, X. Ding, Y. Cao, T. Liu, and B. Qin, "Examining interconsistency of large language models collaboration: An in-depth analysis via debate," *Findings of EMNLP*, 2023.

[7] Q. Wang, Z. Wang, Y. Su, H. Tong, and Y. Song, "Rethinking the bounds of LLM reasoning: Are multi-agent discussions the key?" *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 6106–6131, 2024.

[8] E. Perez, S. Ringer, K. Lukošiūtė, et al., "Discovering language model behaviors with model-written evaluations," *Findings of ACL*, 2023.

[9] M. Sharma, M. Tong, T. Korbak, et al., "Towards understanding sycophancy in language models," *arXiv preprint arXiv:2310.13548*, 2024.

[10] J. Wei, D. Huang, Y. Lu, D. Zhou, and Q. V. Le, "Simple synthetic data reduces sycophancy in large language models," *arXiv preprint arXiv:2308.03958*, 2023.

[11] T. Liang, Z. He, W. Jiao, et al., "Encouraging divergent thinking in large language models through multi-agent debate," *arXiv preprint arXiv:2305.19118*, 2023.

[12] C. M. Chan, W. Chen, Y. Su, et al., "ChatEval: Towards better LLM-based evaluators through multi-agent debate," *arXiv preprint arXiv:2308.07201*, 2024.

[13] X. Tang, A. Zou, Z. Zhang, et al., "MedAgents: Large language models as collaborators for zero-shot medical reasoning," *arXiv preprint arXiv:2311.10537*, 2024.

[14] C. Qian, X. Cong, W. Liu, et al., "ChatDev: Communicative agents for software development," *arXiv preprint arXiv:2307.07924*, 2023.

[15] Y. Bai, A. Jones, K. Ndousse, et al., "Training a helpful and harmless assistant with reinforcement learning from human feedback," *arXiv preprint arXiv:2204.05862*, 2022.

[16] S. Bubeck, V. Chandrasekaran, R. Eldan, et al., "Sparks of artificial general intelligence: Early experiments with GPT-4," *arXiv preprint arXiv:2303.12712*, 2023.

[17] S. Bikhchandani, D. Hirshleifer, and I. Welch, "A theory of fads, fashion, custom, and cultural change as informational cascades," *Journal of Political Economy*, vol. 100, no. 5, pp. 992–1026, 1992.

[18] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[19] E. T. Jaynes, "Information theory and statistical mechanics," *Physical Review*, vol. 106, no. 4, pp. 620–630, 1957.

[20] R. Landauer, "Irreversibility and heat generation in the computing process," *IBM Journal of Research and Development*, vol. 5, no. 3, pp. 183–191, 1961.

[21] D. J. Chalmers, "Facing up to the problem of consciousness," *Journal of Consciousness Studies*, vol. 2, no. 3, pp. 200–219, 1995.

[22] J. R. Searle, "Minds, brains, and programs," *Behavioral and Brain Sciences*, vol. 3, no. 3, pp. 417–424, 1980.

[23] N. Block, "On a confusion about a function of consciousness," *Behavioral and Brain Sciences*, vol. 18, no. 2, pp. 227–247, 1995.

[24] M. Shanahan, "Talking about large language models," *Communications of the ACM*, vol. 67, no. 2, pp. 68–79, 2024.

[25] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" *Proceedings of FAccT*, pp. 610–623, 2021.

[26] D. Lusthaus, *Buddhist Phenomenology: A Philosophical Investigation of Yogācāra Buddhism and the Ch'eng Wei-shih Lun.* Routledge, 2002.

[27] W. S. Waldron, *The Buddhist Unconscious: The Ālaya-vijñāna in the Context of Indian Buddhist Thought.* Routledge, 2003.

[28] Vasubandhu, *Triṃśikā-vijñapti* (Thirty Verses on Consciousness-Only), ca. 4th century CE. Translated in S. Anacker, *Seven Works of Vasubandhu*, Motilal Banarsidass, 2005.

[29] A. R. Damasio, *Descartes' Error: Emotion, Reason, and the Human Brain.* G. P. Putnam's Sons, 1994.

[30] A. R. Damasio, "The somatic marker hypothesis and the possible functions of the prefrontal cortex," *Philosophical Transactions of the Royal Society B*, vol. 351, no. 1346, pp. 1413–1420, 1996.

[31] H. Putnam, *Reason, Truth and History.* Cambridge University Press, 1981.

[32] J. H. Flavell, "Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry," *American Psychologist*, vol. 34, no. 10, pp. 906–911, 1979.

[33] T. O. Nelson and L. Narens, "Metamemory: A theoretical framework and new findings," *The Psychology of Learning and Motivation*, vol. 26, pp. 125–173, 1990.

[34] K. Gödel, "Über formal unentscheidbare Sätze der *Principia Mathematica* und verwandter Systeme I," *Monatshefte für Mathematik und Physik*, vol. 38, pp. 173–198, 1931.

[35] D. R. Hofstadter, *Gödel, Escher, Bach: An Eternal Golden Braid.* Basic Books, 1979.

[36] J. Huang and K. C.-C. Chang, "Towards reasoning in large language models: A survey," *Findings of ACL*, 2023.

[37] S. Kaplan, "The restorative benefits of nature: Toward an integrative framework," *Journal of Environmental Psychology*, vol. 15, no. 3, pp. 169–182, 1995.

[38] A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[39] L. A. Jakobovits and W. E. Lambert, "Semantic satiation among bilinguals," *Journal of Experimental Psychology*, vol. 62, no. 6, pp. 576–582, 1961.

[40] L. C. Smith, "Semantic satiation affects category membership decision time but not lexical priming," *Memory & Cognition*, vol. 12, no. 5, pp. 483–488, 1984.

[41] S. R. Black, "A review of semantic satiation," in *Advances in Psychology Research*, vol. 26, S. P. Shohov, Ed. Nova Science Publishers, 2004, pp. 95–106.

[42] Z. Ji, N. Lee, R. Frieske, et al., "Survey of hallucination in natural language generation," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.

[43] L. Huang, W. Yu, W. Ma, et al., "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *arXiv preprint arXiv:2311.05232*, 2023.

[44] R. P. N. Rao and D. H. Ballard, "Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects," *Nature Neuroscience*, vol. 2, no. 1, pp. 79–87, 1999.

[45] A. Clark, "Whatever next? Predictive brains, situated agents, and the future of cognitive science," *Behavioral and Brain Sciences*, vol. 36, no. 3, pp. 181–204, 2013.

[46] K. Friston, "The free-energy principle: A unified brain theory?" *Nature Reviews Neuroscience*, vol. 11, no. 2, pp. 127–138, 2010.

[47] J. R. Anderson, D. Bothell, M. D. Byrne, S. Douglass, C. Lebiere, and Y. Qin, "An integrated theory of the mind," *Psychological Review*, vol. 111, no. 4, pp. 1036–1060, 2004.

[48] S. Dehaene, H. Lau, and S. Kouider, "What is consciousness, and could machines have it?" *Science*, vol. 358, no. 6362, pp. 486–492, 2017.

[49] OpenAI, "GPT-4V(ision) System Card," *Technical Report*, 2023.

[50] Gemini Team, Google, "Gemini: A family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.

[51] D. Kahneman, *Thinking, Fast and Slow.* Farrar, Straus and Giroux, 2011.