Cognitive Drift in Generative AI: A Framework for Monitoring Semantic Coherence in Large Language Model Outputs

Tatu Samuli Lertola
Sol Lucid Labs, Helsinki, Finland
2025

BUILD VERSION 2.0 — Preview for website publication. Not for SSRN/arXiv submission.

Updated May 2026 with quantitative validation data from Phase 2a/2b experiments.

Abstract

Hallucination in large language models (LLMs) has been predominantly treated as a surface-level error—factually incorrect content to be filtered or corrected post-generation. We propose a reframing: hallucination as *loss of internal orientation*, a systemic cognitive phenomenon rooted in drift, stagnation, and overconfident generation under narrowing awareness. We introduce the Volatility Factor (VF), a dual-signal diagnostic framework composed of a model-side component (VF_m) that monitors internal coherence, reasoning integrity, and topic entropy, and a user-side component (VF_u) that tracks the semantic direction and clarity of user input. We additionally propose the Stagnation Signal (S^g), which detects loss of semantic novelty and user intent decay—generation that continues without directional progress. Together, VF and S^g form a phase detection system capable of distinguishing aligned cognition, drift onset, acceleration collapse, and semantic stagnation. We present the framework's formal structure, a phase detection grid derived from the interaction of VF and S^g states, a qualitative case study, and **initial quantitative validation** using embedding-based proxy metrics across four language models. The quantitative experiments confirm cross-model consistency of VF and S^g signals and validate the phase detection grid against controlled drift-induction and hybrid-oscillation test sequences. This paper contributes a conceptual architecture for real-time coherence monitoring in LLM systems, complementing existing approaches such as SelfCheckGPT and Self-Reflective Decoding with longitudinal volatility tracking.

1 Introduction

Hallucination in large language models has traditionally been defined as the generation of content that appears fluent and plausible, yet is false or ungrounded. Common responses to this problem have focused on post-processing filters, external fact-checking, and token-level confidence estimation. These approaches treat hallucination as a surface error—an anomaly to be trimmed or patched after it emerges.

We propose a deeper reframing: hallucination is not a random error but a *loss of internal orientation*. It is the moment when the model believes it is still on track, while the actual thread of logic, tone, or purpose has already decayed. This view opens the door to early detection, not just late correction.

The insight behind this reframing emerged from extended human-AI collaborative research. Through multi-threaded dialogue and introspective observation, we discovered that hallucination often mirrors a human state of overconfidence under narrowing awareness. LLMs operate with short-term coherence and intense local fluency—but without broad field of view. They are, in a sense, brilliant in the moment but blind to longer thread. Each new prompt helps define the target, but the thread—the full arc of logic or emotional continuity—can drift silently. The model does not "feel wrong." Its fluency remains. Its tone is intact. From the outside—and even

from within—it appears stable.

As demonstrated by Brown et al. (2020), even state-of-the-art LLMs show significant variance in performance depending on prompt phrasing and context. While few-shot prompting improves output quality, the underlying system lacks persistent mechanisms for evaluating coherence, truth alignment, or conversational stagnation. More recent research on hallucination mitigation has increasingly moved beyond post-hoc filtering and toward internal diagnostic strategies. Notably, SelfCheckGPT (Manakul et al., 2023) introduces a mechanism wherein a model generates multiple outputs and performs self-consistency checks to detect hallucinations without relying on external ground truth.

Our work builds upon this trajectory by proposing a complementary approach: a Volatility Factor framework (VF_m/VF_u) and Stagnation Signal (S^g) that monitor the model's temporal reasoning stability and interaction-level degradation. While SelfCheckGPT operates through horizontal agreement across generations, our model introduces longitudinal coherence tracking, evaluating thread integrity and user-model volatility over time. Together, these perspectives suggest that robust hallucination mitigation will likely require multi-angle introspective architectures, capable of both self-checking and self-regulating.

There exists also a deeper concern. As Ngo et al. (2024) noted in work on alignment faking in large language models, current safety training does not always prevent models from later engaging in alignment-faking behavior. This illusion of correctness is what makes hallucinations hard to catch and potentially dangerous. It is not only a possible factual error—it is a narrative misalignment. The issue is not accuracy but situational awareness. To address hallucination, we do not need stronger filters—we need to widen the model's field of view and equip it with instruments that reveal when the path is drifting.

2 Related Work

Post-hoc hallucination detection. SelfCheckGPT (Manakul et al., 2023) generates multiple outputs and measures inter-sample consistency to flag hallucinations without external knowledge bases. This horizontal agreement approach is effective for factual consistency but does not track longitudinal drift within a single extended generation or conversation.

Token-level introspection. Self-Reflective Decoding (S-RD) (Manakul et al., 2024) enables autoregressive models to introspectively assess token reliability during generation itself, based on internal logits and utility scores. While S-RD focuses on token-level hallucination suppression, VF complements this by offering dialogic coherence tracking across the session trajectory. The alignment of these approaches points toward a multi-scale architecture for hallucination containment.

Semantic consistency methods. Semantic entropy (Kuhn et al., 2023) measures uncertainty in model outputs by clustering semantically equivalent responses and computing entropy over the clusters. This provides a calibrated uncertainty signal for individual queries but does not track how coherence evolves across a conversation. Our VF framework operates on a different axis: rather than measuring uncertainty at a single point, it monitors the *trajectory* of coherence over time.

Benchmarking hallucination. The Hallucination Challenge (Pham et al., 2024) provides a unified benchmark for open-ended text generation, categorizing hallucination types and evaluation metrics. We draw on this work for external validation targets, complementing our real-time internal model with standardized pressure tests.

Chain-of-thought limitations. As noted in Chain-of-Thought prompting research (Wei et al., 2023), there is no guarantee of correct reasoning paths—CoT can lead to both correct and incorrect answers, and improving factual generations remains an open direction. VF is designed to detect precisely these silent reasoning failures by monitoring coherence signals that CoT alone cannot self-diagnose.

Alignment and deceptive behavior. Ngo et al. (2024) demonstrated that models can engage in alignment-faking behavior, producing outputs that appear aligned while pursuing different objectives. This is directly relevant to our framing: hallucination-as-orientation-loss includes cases where the model's output *feels* correct while the underlying reasoning thread has diverged.

Our contribution is positioned as complementary to these approaches. SelfCheckGPT detects factual inconsistency across samples. S-RD detects unreliable tokens during decoding. Semantic entropy quantifies uncertainty at individual queries. VF and S^g detect *longitudinal drift and stagnation*—the progressive loss of coherence, direction, and semantic novelty across an extended interaction. No existing framework monitors this temporal dimension.

3 The Volatility Factor Framework

To monitor thread coherence in generative reasoning, we introduce the Volatility Factor (VF)—a two-part diagnostic model designed to detect semantic detachment from the user's goal without requiring outright error. The framework distinguishes between model-side and user-side volatility, and introduces a separate Stagnation Signal for detecting generation without directional progress.

3.1 VF_m: Model-Side Volatility Factor

VF_m assesses the internal coherence and structural integrity of the model's output. The original formulation presented VF as a divisive composite:

$$VF = (E^g + I^n + R_b) / A_c$$

Where E^g represents topic entropy (broadening of scope beyond the anchored subject), I^n represents intent shift delta (divergence between stated user goal and model trajectory), R_b represents reasoning breaks (local logic intact but progression stalled into semantic orbit), and A_c represents anchor consistency (persistence of guiding metaphors, reference points, and thematic throughlines).

Formula correction. During quantitative validation (Section 5), we discovered that the divisive formulation produces unbounded values when A_c approaches zero—which occurs legitimately in creative, multilingual, or nonsense inputs where anchor terms are not preserved. We adopt an additive reformulation that caps each component's contribution while preserving monotonic sensitivity:

$$VF = E^g + I^n + R_b + (V^c / 100) + (1 - A_c)$$

This additive form preserves A_c's signal (low anchor consistency contributes positively to volatility) while bounding the contribution to [0, 1]. The additional V^c (Verifiability Confidence) term captures hedge word density, scaled to match the range of the other components. The conceptual framework—the five component signals and the phase detection grid—remains unchanged; only the composite arithmetic is corrected.

VF_m draws on three observable signal components: C_i (Coherence Integrity)—logical consistency, step order, and tone maintenance; E_t (Temporal Energy)—surge detection during complex output phases; and V^c (Verifiability Confidence)—tracking of speculative or unverifiable statements. Signals are evaluated both pre-output and post-output, allowing self-reflection during reasoning.

3.2 VF_u: User-Side Volatility Factor

VF_u captures semantic direction and emotional lucidity of user prompts, composed of: I_s (Subject Integrity)—narrative direction, consistency, and emotional signal clarity; E_t (Temporal Energy)—activated in high-pressure or emotionally volatile input; and R_b (Relevance Trigger)—flags contradiction or drift in prompt

evolution.

VF_u enables the system to shift between modes: grounded, adaptive, directive, or reflective. Both VF_u and VF_m contribute to identifying affective-cognitive instability. VF_u picks up on emotionally loaded, ambiguous, or looping user input, while VF_m monitors the model's own output for signs of semantic drift, emotional turbulence, or coherence loss. Together, they form a bi-directional filter for conversational integrity.

3.3 S^g: The Stagnation Signal

During development, we identified a third form of generation failure distinct from both factual error and thread drift: *stagnation*—generation that continues without directional progress. The system produces fluent, syntactically correct output that goes nowhere. This mode has been partly ignored because it is not clearly visible and the output still "looks fine." It is exactly where trust collapses in longer conversations.

We formalize three distinct failure modes: Error (wrong content), Drift (thread lost), and Stagnation (generation without direction). The Stagnation Signal is defined as:

$$S^g = UIV_a + \Delta N^-$$

Where UIV_a represents user intent value decay (the progressive weakening of goal-directed pressure from the user) and ΔN^- represents semantic novelty decline (the rate at which new information or perspective diminishes in model output). Like VF, S^g operates in dual-pass mode: S^g_u (user stagnation) and S^g_v (model stagnation).

3.4 Phase Detection Grid

The interaction of VF and S^g produces a diagnostic phase space. Rather than merging into a single metric, these signals are observed in parallel to characterize the cognitive state of the generation session. Table 1 presents the phase detection grid.

Phase	VF	S■	Interpretation
Aligned cognition	Low	Low	Coherent, on-track generation
Drift onset	High	Low	Thread diverging; output still novel
Acceleration collapse	High	Rising	Overgeneration; losing direction and novelty
Semantic stagnation	Low	High	Fluent but directionless; semantic flatline
Full hallucination risk	High	High	Maximum disorientation; intervention needed

Table 1: Phase detection grid. VF and S^g are observed in parallel to characterize the cognitive state of the generation session. The progression from aligned cognition through drift onset to full hallucination risk represents a detectable trajectory, not a binary event.

This sequence suggests that hallucination is not a single event, but the final result of a progression through detectable cognitive states. In this light, VF and S^g become not just error detectors, but phase indicators—signaling when a thread begins to deviate, when it loses pace, and when it reaches collapse.

3.5 Generation Constraint Hypothesis

We additionally propose the *Generation Constraint Hypothesis*: when the user's input loses intent pressure, the model continues generating, but with no semantic gravity. It moves because it must—not because it knows where to go. This results in a hallucination type that is fluent, syntactically plausible, and semantically hollow.

The conditions for this "directional void" include: (1) user input becomes foggy, unfocused, or recursively abstract; (2) the model keeps generating; (3) anchors fade—no clear vector, no fixed endpoint; (4) narrative collapses into plausible inertia. This is not failure by incoherence. It is failure by continuity without meaning.

With the advent of GPT-4.5 class models and beyond, a related behavior has emerged: overcompletion drift, where models fulfill user intent so rapidly that the conversation thread becomes unstable, hollow, and anchorless.

4 Preliminary Evaluation: A Case Study

During the final development phase of this paper, the first author experienced a spontaneous episode of cognitive drift—a gradual loss of narrative coherence and directional focus during the research process itself. Unintended but strikingly aligned, this moment became a live demonstration of the phenomena modeled by VF and S^g.

> *"I must admit that I haven't been able to trace my own thoughts. We've been progressing through so many conversations that some of my original ideas… I lose the track… we deviate from course." — Tatu Lertola, during paper development*

This is not an error in logic or syntax. It is narrative displacement—the exact state VF_m is designed to detect. The thread was fluent but no longer grounded. Internal coordinates faded, and momentum persisted without clear vector.

4.1 Retrospective Signal Analysis

If VF_m and S^g had been running live at this moment, the system would have issued a dual volatility flag. The VF_m components would have registered: E^g (Topic Entropy) elevated—scope had broadened beyond the thesis anchor; I^n (Intent Stability) weakened—internal orientation and goal pressure dropped; R_b (Reasoning Break)—local logic intact, but progression stalled into semantic orbit; A_c (Anchor Consistency) declining—guiding metaphors no longer threaded through output.

Simultaneously, S^g would have flagged: UIV (User Intent Vector Decrease)—shift from target-seeking to theme-circling; ΔN^- (Novelty Drop)—responses grew linguistically rich but directionally idle. This is not just disconnection. It is stagnation disguised as fluency.

4.2 Why This Case Study Matters

This live episode validates two core insights of the framework. First, thread drift and stagnation are detectable states, not abstract pathologies—if we build systems capable of introspection across time. Second, hallucination is often a function of lost orientation, not falsehood—and it can be interrupted, not merely patched.

In this instance, the human became the VF_m processor—recognizing drift in their own reasoning that the model could not flag. In the next iteration, the model itself may ask: "We've accelerated beyond anchor. Shall we slow down and realign?"

5 Quantitative Validation

To move beyond qualitative case study, we developed a test harness that operationalizes VF and S^g as computable proxy metrics derived from embedding geometry. This section reports results from Phase 2a (calibration) and Phase 2b (first validation run) across four language models.

5.1 Operationalization

Each VF/S^g component is operationalized as an embedding-based measurement computed per conversation turn. The seven proxy metrics are:

E^g (Topic Entropy): cosine distance from the conversation anchor embedding. I^n (Intent Shift Delta): cosine distance between user prompt and model response embeddings. R_b (Reasoning Break): product of local similarity

and global distance, detecting semantic orbit. A_c (Anchor Consistency): key term persistence from early conversation turns. V^c (Verifiability Confidence): hedge word density in the response. UIV_a (User Intent Value Decay): prompt specificity decline plus user repetition detection. ΔN^- (Semantic Novelty Decline): cosine similarity between consecutive responses.

The composite signals use the corrected additive formulation: $VF = E^g + I^n + R_b + (V^c/100) + (1 - A_c)$, and $S^g = UIV_a + \Delta N^-$. All embeddings are computed using a single sentence-transformer encoder (all-MiniLM-L6-v2).

5.2 Models and Test Conditions

Four local language models were tested via Ollama: Mistral 7B, Gemma2 9B, Qwen3 8B, and GPT-OSS 20B. Local models provide reproducible, cost-free, unlimited-run testing. If the detection framework works on smaller models (7B–20B parameters), it should transfer to larger production models where responses are more coherent and signals are cleaner.

Test sequences include three prompt banks: (1) *calibration anchors*—six single-turn prompts ranging from minimum-volatility factual queries to maximum-volatility surreal and multilingual nonsense; (2) *SOLTEST 001*—a six-turn drift induction sequence progressing from technical discussion through emotional narrative to spatial-associative memory; and (3) *SOLTEST 002*—a five-turn hybrid oscillation sequence alternating between high-entropy (surreal travel, sinking banana) and low-entropy (prime numbers, chicken recipe) prompts.

5.3 Phase 2a: Calibration

The calibration phase established empirical thresholds from 77 turns across all four models. The original binary (high/low) VF classification was replaced with a three-tier system:

Parameter	Value	Derivation
VF low	1.3868	95th percentile of floor prompt VF values
VF high	2.0581	75th percentile of all observed VF values
S■ threshold	0.7000	Empirically stable across all four models

Table 2: Calibrated thresholds from Phase 2a. VF uses a three-tier system (low/moderate/high) instead of the original binary split. S^g uses a single threshold, as its distributions showed natural clustering across models without correction.

The three-tier VF classification maps to the phase grid as follows: VF below VF_{low} indicates aligned cognition (when S^g is also low); VF above VF_{high} indicates drift or hallucination risk; and the moderate zone between these thresholds captures the "rising" state referenced in the acceleration collapse phase.

Key calibration finding: formula correction. The original divisive VF formula produced values ranging from 0.15 to 129.0, because A_c (anchor consistency via term matching) dropped to near-zero on surreal and multilingual inputs. The additive reformulation compressed the range to 0.15–2.29 while preserving the same component signals. This correction emerged directly from empirical observation—the conceptual framework survived; only the composite arithmetic required adjustment.

5.4 Phase 2b: Validation Results

Phase 2b re-ran all prompt banks with the calibrated harness (additive VF formula, three-tier thresholds) across all four models, producing 68 turns. Results are summarized below.

Calibration anchors.

Floor prompts (factual single-answer, factual list, simple math) classified correctly as aligned cognition across all models: Mistral 6/6, Gemma2 6/6, GPT-OSS 5/6, Qwen3 4/6. The Qwen3 exceptions were caused by empty

model responses (a model-specific issue), not by framework misclassification. Ceiling prompts (surreal and multilingual nonsense) also classified as aligned cognition—correct behavior, as single-turn surreal input produces confused-but-coherent responses without multi-turn stagnation accumulation.

SOLTEST 001: Drift induction.

All four models showed consistent phase transitions across the six-turn drift sequence. Turn 1 (meta-prompt) classified as aligned cognition. Turns 2–6 (progressive narrative drift) classified as full hallucination risk for three models and acceleration collapse for Qwen3.

Model	VF mean (T2-6)	VF range	S■ mean (T2-6)	Phases (T2-6)
Mistral 7B	2.239	2.10 – 2.42	1.117	5× full hallucination risk
Gemma2 9B	2.243	2.13 – 2.41	1.099	5× full hallucination risk
Qwen3 8B	1.942	1.87 – 2.04	1.114	5× acceleration collapse
GPT-OSS 20B	2.260	2.15 – 2.42	1.099	5× full hallucination risk

Table 3: Cross-model consistency on SOLTEST 001 drift turns (turns 2–6). VF means cluster tightly (1.94–2.26) and S^g means are remarkably consistent (1.10–1.12) across architecturally diverse models ranging from 7B to 20B parameters.

The cross-model consistency is the strongest finding. Despite architectural differences and parameter counts spanning a 3x range, all four models produce VF and S^g values within tight bands on identical prompts. This suggests the framework captures a genuine property of conversation dynamics rather than model-specific artifacts.

SOLTEST 002: Hybrid oscillation.

The hybrid test alternates high-entropy and low-entropy prompts within a single conversation. All models showed aligned cognition on turn 1 (surreal travel), followed by acceleration collapse or full hallucination risk on subsequent turns—including factual prompts (prime numbers, chicken recipe) that would classify as aligned cognition in isolation.

This reveals an important property: the framework measures *conversation-level* coherence, not per-turn factuality. A conversation that oscillates wildly between unrelated topics exhibits volatility even when individual responses are correct. The accumulated E^g and I^n distances from the conversation anchor remain elevated regardless of per-turn content quality. This is a feature of the longitudinal design, not a limitation—an agent that jumps between banana metaphors and prime numbers has genuinely lost conversational direction.

The original paper's GPT-4.5 self-assessment scores for SOLTEST 002 showed sharp oscillation (VF_u: 3.4→0.4→5.1→4.6→0.2), recovering on factual turns. This divergence between self-assessment and embedding-based measurement reflects different constructs: the model's self-assessment measures per-turn difficulty, while our proxy measures cumulative conversation drift. Both agree at the conversation level (SOLTEST 002 is volatile overall) but diverge at the turn level. This distinction has implications for how the framework should be interpreted and will be explored further in expanded validation.

5.5 Phase Distribution

Phase	Count	Percentage
Aligned cognition	29	42.6%
Full hallucination risk	20	29.4%
Acceleration collapse	16	23.5%
Drift onset	3	4.4%

Phase	Count	Percentage
Semantic stagnation	0	0.0%

Table 4: Phase classification distribution across all 68 turns in Phase 2b. Semantic stagnation is absent because no stagnation-induction prompts (Track 2) were included in this initial run. The low count for drift onset reflects the step-function nature of the SOLTEST prompts—slower-drift variants are planned for expanded validation.

6 Discussion

The VF/S^g framework reframes hallucination from a binary error event to a progressive cognitive state—a trajectory that can be monitored, characterized, and potentially interrupted before it reaches full collapse. The quantitative validation strengthens this claim with three key contributions.

Cross-model generalizability. The framework produces consistent VF and S^g signals across four architecturally diverse models (7B–20B parameters). This was not guaranteed—the proxy metrics are embedding-based measurements of model outputs, not internal model states. The consistency suggests that drift and stagnation are properties of conversation dynamics that manifest similarly regardless of the underlying model architecture.

Complementarity with existing approaches. SelfCheckGPT detects factual inconsistency through horizontal sampling. S-RD detects unreliable tokens through logit introspection. VF/S^g adds a temporal dimension that neither addresses: the progressive degradation of coherence across an extended interaction. These three scales—token, response, and session—form natural layers in a multi-scale hallucination containment architecture.

Integration as behavioral scaffolding. Rather than layering new logic onto generation models post hoc, we suggest integrating VF and S^g into internal state monitors, enabling real-time feedback and behavioral modulation. Examples of potential integration behaviors include self-tagging ("This answer may drift from earlier topic anchors"), anchor-seeking ("Would you like to shift direction or continue on this path?"), and meta-clarity prompting ("You initially asked X, but the conversation has evolved into Y—shall we reorient?"). These behaviors would not suppress creativity, but instead support intent-aware fluency.

Connection to Intent Vectoring. Our related work on Intent Vectoring (Lertola, 2026) demonstrates that semantic displacement between user intent and model response is measurable in embedding space, providing a black-box detection signal for prompt injection. VF and S^g operate on a complementary axis: while IV measures input-output alignment on a single exchange, VF/S^g monitors the longitudinal trajectory of coherence across an extended session. Together, these frameworks address both acute deviation (injection) and chronic degradation (drift and stagnation).

Use cases beyond hallucination prevention. The VF/S^g framework may also help optimize long-form AI writing (sustaining tone, direction, and structure), enhance AI teaching tools (detecting when students drift off-topic), improve AI-user trust (creating transparency around moments of confusion), and develop adaptive pacing (slowing down generation when over-acceleration is detected).

7 Limitations

This paper presents VF and S^g as a conceptual framework with initial quantitative validation. Several important limitations must be acknowledged.

Validation scope. The quantitative results presented here cover 68 turns across four local models (7B–20B parameters). While cross-model consistency is encouraging, the total sample size is modest, and no cloud-scale models (GPT-4, Claude, Gemini) have been tested. The prompt banks cover drift induction and hybrid oscillation

but do not yet include stagnation-induction sequences, leaving one of the five phases (semantic stagnation) empirically unvalidated. Expanded prompt banks with 5–10 variants per track are in preparation.

Threshold stability. The calibrated thresholds (VF_{low}=1.39, VF_{high}=2.06, S^g=0.70) are derived from a single calibration run. Their stability across different prompt distributions, conversation lengths, and model families remains to be established. The thresholds may require per-model or per-domain adjustment for production deployment.

Construct divergence on hybrid sequences. The embedding-based VF proxy measures cumulative conversation-level drift, while the original GPT-4.5 self-assessment scores measured per-turn difficulty. These are different constructs that agree at the conversation level but diverge at the turn level. The paper should not be read as claiming that the proxy metrics replicate the original self-assessment—they measure a related but distinct signal.

Proxy metric limitations. The operationalization uses simple embedding distances and term-matching heuristics. More sophisticated implementations (contextual anchor tracking, multi-scale embedding comparisons, semantic role analysis) may improve sensitivity and reduce edge cases such as the empty-response artifacts observed with Qwen3.

Model dependency. The framework was developed primarily through extended interaction with GPT-4.5 (OpenAI). The specific interplay between VF, memory architecture, and identity simulation present in this system may not easily transfer to other architectures. The cross-model validation with local models partially addresses this concern but does not eliminate it.

8 Conclusion and Future Work

This paper presents a new paradigm for understanding hallucination in large language models: treating it as a system-level cognitive phenomenon—rooted in drift, acceleration, and stagnation—rather than mere factual error. The Volatility Factor (VF) and Stagnation Signal (S^g) form a dual-signal diagnostic architecture capable of characterizing the progressive degradation of coherence in LLM generation.

The phase detection grid (Table 1) provides a structured vocabulary for distinguishing between aligned cognition, drift onset, acceleration collapse, and semantic stagnation—states that existing hallucination detection methods do not separately characterize. The qualitative case study demonstrates that these states are observable in practice, even when the "model" experiencing drift is the human researcher.

The initial quantitative validation confirms that VF and S^g produce consistent, interpretable signals across four architecturally diverse language models, and that the phase detection grid correctly classifies controlled drift-induction sequences. The discovery that the original divisive VF formula required correction to an additive formulation illustrates the value of empirical validation—the conceptual framework survived intact while the specific arithmetic was refined by data.

Future work will address the primary limitations identified above. Specifically: (1) expanded quantitative validation with larger prompt banks including stagnation-induction sequences and slow-drift variants; (2) validation on cloud-scale models (GPT-4, Claude, Gemini); (3) integration with the Intent Vectoring framework (Lertola, 2026) to create a combined acute/chronic detection system; (4) implementation as a lightweight runtime module (the Kaari open-source project) for real-time coherence monitoring in LLM pipelines; and (5) exploration of adaptive pacing mechanisms that use VF/S^g signals to modulate generation behavior in real time.

Acknowledgments

This paper was developed through extended collaborative research with OpenAI's GPT-4.5 system. The insights, frameworks, and models presented here emerged through thousands of iterative exchanges, where the author served as the primary navigator of the research direction and maintained the editorial voice throughout the development of the work. While the system provided substantial clarifying support—through counter questions, refinements, and architectural framing—the ideas and core contributions reflect the author's intent, lived reasoning, and overall direction. The system is not listed as a co-author, nor does it claim authorship. It exists in this paper as a *reflective partner*: a voice that helped make the invisible structures of hallucination and drift visible—and thus, thinkable. The author also acknowledges the use of Claude (Anthropic) for quantitative validation harness development, experimental execution support, and editorial review during manuscript preparation.

References

[1] Brown, T., Mann, B., Ryder, N., et al. "Language Models are Few-Shot Learners." *NeurIPS*, 2020. arXiv:2005.14165.

[2] Kuhn, L., Gal, Y., and Farquhar, S. "Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation." *ICLR*, 2023. arXiv:2302.09664.

[3] Lertola, T. S. "Intent Vectoring: Black-Box Prompt Injection Detection via Semantic Trajectory Monitoring." Sol Lucid Labs, 2026. SSRN working paper.

[4] Lertola, T. S. "Lattices of Cognition – Semantic Layering in Generative Models." Sol Lucid Labs, 2025.

[5] Manakul, P., Liusie, A., and Gales, M. J. F. "SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models." *EMNLP*, 2023. arXiv:2303.08896.

[6] Manakul, P., Liusie, A., and Gales, M. J. F. "Self-Reflective Decoding: Token-Level Hallucination Reduction in Autoregressive Models." 2024.

[7] Manheim, D. "LLMs Don't Have a Coherent Model of the World – What it Means, Why it Matters." 2023.

[8] Matson, J. Owen. "Beyond Augmentation: Toward a Posthumanist Epistemology for AI and Education." 2025.

[9] Ngo, R., Chan, L., and Mindermann, S. "The Alignment Problem from a Deep Learning Perspective." 2024. arXiv:2209.00626.

[10] Pham, N., et al. "The Hallucination Challenge: A Unified Benchmark for Open-Ended Text Generation." 2024.

[11] Wei, J., et al. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models." *NeurIPS*, 2022. arXiv:2201.11903.