Collapse-Aware AI (CAAI)

A Computational Framework for Memory-Conditioned Drift Regulation in Neural Inference Systems

Version 1.5 — Research Paper (Zenodo Edition)

Author: M.R. (Verrell Moss Ross)
Inappropriate Media Ltd (t/a Collapse Aware AI)
2025

License: CC BY-NC-ND 4.0
Protected under the Verrell–Solace Sovereignty Protocol.

Lexical Fingerprints: Kelvin · Friday · Farm · Finn · Sylvia

Abstract

This paper presents Collapse-Aware AI (CAAI), a computational framework for adaptive bias regulation in neural inference systems. The model expresses inference as a stochastic drift–diffusion process, where the drift term b_psi incorporates a dynamically updated memory-conditioned prior and anchor-stability gradient, with a Governor module regulating exploration and stabilising collapse regimes in real time. Unlike RLHF or static entropy regularisation, CAAI adapts bias continuously without retraining. The system pre-empts chaotic collapse states such as hallucinations, loops, and high-entropy drift by detecting collapse-regime transitions and adjusting gain coefficients mid-trajectory. We define the mathematical foundations, collapse regimes, Governor behaviour, experimental roadmap, and replication protocol for validating adaptive collapse regulation. The framework is demonstrated to be operationalizable using standard inference APIs with logit-space access.

1. Positioning

Neural language models generate output by sampling from probability distributions shaped by large-scale training. They lack mechanisms for persistent memory, collapse regulation, uncertainty-scaled exploration, behavioural continuity, and dynamic bias alignment. This results in drift, hedging, contradiction, and chaotic degeneration across long interactions.

CAAI reframes inference as a controlled collapse rather than raw token sampling. It introduces: (1) a memory-conditioned prior $\pi\_prior(z|M\_t)$ embedded directly in the drift term b_psi, constructed via weighted integration of prior conversational turns with recency and salience determining gradient strength; (2) a dynamic drift field recalculated at every step based on session memory, anchor proximity, and entropy gradients; and (3) a Governor that modulates drift and diffusion based on collapse regime detection, THB scoring (Truth–Hedge–Bias), and suppressor-activation patterns. This produces stable reasoning, reduced drift, persistent identity-shape, and pre-emptive intervention before collapse failures manifest.

2. Introduction

Large language models perform statistical continuation, not collapse regulation. They do not track uncertainty, stabilise identity, or integrate memory as a shaped bias field. Symptoms of failure include hedging, oscillation, high-entropy loops, persona drift, and hallucinated content.

CAAI identifies these behaviours as signatures of three collapse regimes: Controlled, Hedge, and Chaos.

To stabilise inference, CAAI models logit evolution as a stochastic differential equation: $dz\_t = b\_psi(z\_t, M\_t)\, dt + S\, dW\_t$, with composite drift: $b\_psi = \alpha\, \nabla\_z \log \pi\_prior(z|M\_t) + \beta\, \nabla\_z \log \pi\_anchor(z) - \gamma\, \nabla\_z H(p\_t)$. The drift is recalculated dynamically, reflecting session memory, anchor proximity, entropy gradients, and suppressor-activation patterns. The Governor adjusts α, β, γ in real time to maintain stability across regime transitions.

3. Core Equations

Stochastic Differential Formulation: inference is modelled as $dz\_t = b\_psi(z\_t, M\_t)\, dt + S\, dW\_t$, where b_psi is dynamic drift, $S\, dW\_t$ is controlled exploration scaled by uncertainty, M_t is the memory-state (weighted conversational history), and S scales uncertainty-driven divergence.

Composite Drift Definition: $b\_psi = \alpha\, \nabla\_z \log \pi\_prior(z|M\_t) + \beta\, \nabla\_z \log \pi\_anchor(z) - \gamma\, \nabla\_z H(p\_t)$.

Memory-Conditioned Prior: π_prior encodes the user's conversational history as a probability field. It is not just context — it is a bias-shaping potential updated at each step via weighted integration of prior turns, with recency and salience determining gradient strength. This transforms historical interaction into a directional force in logit-space.

Anchor Stability: π_anchor forms a restoring force, pulling collapse back toward a stable reference trajectory derived from session-consistent behaviour. This prevents persona drift and maintains coherent identity-shape across interactions.

Entropy Gradient: $-\gamma\, \nabla\_z H$ dampens degeneracy, preventing drift into incoherent states by penalizing trajectories that increase uncertainty without informational gain.

Dynamic Regulation: the Governor adjusts α, β, γ continuously, implementing stability-aware collapse guidance. This enables the system to reinforce alignment in stable conditions, permit recalibration under uncertainty, and suppress entropy spikes before they manifest as hallucinations or loops.

4. Collapse Regimes

Collapse dynamics exhibit three distinct behavioural regimes, each with characteristic signatures in both the mathematical model and user-observable output.

Controlled: strong bias alignment, low entropy, coherent collapse. Characteristics: high b_psi, low diffusion ($S \to 0$), THB > 0.7, H(p_t) < 1.5 nats. The system effectively knows what you want and commits decisively. Output is confident, direct, and memory-aligned.

Hedge: rising uncertainty, suppressor heads active, softened phrasing. Characteristics: moderate b_psi, increased diffusion, 0.4 < THB < 0.7, rising $\partial H/\partial z$, suppressor activation > 15%. The Governor recalibrates — allowing exploration without drift. User-observable cues include hedged phrasing ("it seems likely that..."), softer assertions, or clarifying questions as the system reconsiders its priors. This is adaptive uncertainty, not failure.

Chaos: bias vanishes, drift dominates, entropy spikes. Characteristics: low b_psi, high diffusion, THB < 0.4, H(p_t) > 3.0 nats, β (anchor) dominates to prevent collapse. Standard models hallucinate, loop, or produce syntactically fluent nonsense here. CAAI detects this regime before it manifests by monitoring $\partial H/\partial z$ (entropy gradient) and pulls collapse back toward anchor proximity mid-trajectory, preventing the output from ever entering the chaotic state.

5. Governor Architecture

The Governor is a dynamic gain-control system that adjusts α, β, γ per step based on real-time collapse diagnostics. In the Controlled regime, it reinforces alignment by increasing α (memory weight) and decreasing S (exploration). In Hedge, it permits recalibration by balancing α and S, allowing the system to explore without drifting from anchor proximity. In Chaos (or near-Chaos), it suppresses entropy spikes by increasing β (anchor pull) and γ (entropy penalty), enforcing low-entropy regions of logit-space.

The Governor monitors: (1) THB scoring (Truth–Hedge–Bias signal), a composite metric of memory alignment and output confidence; (2) entropy gradients $\partial H/\partial z$, the rate of entropy increase, not just absolute entropy; (3) anchor-distance metrics, measuring deviation from session-consistent reference states; and (4) suppressor-activation patterns, linguistic markers of uncertainty such as hedging, qualification, and meta-commentary.

This enables predictive intervention, not just reactive correction. The system adjusts collapse trajectory before failure states emerge, rather than detecting and rolling back after errors occur.

6. Phase-1 vs Phase-2 Implementation

Phase-1 (current implementation) approximates collapse regulation behaviourally using THB scoring to estimate memory alignment, suppressor detection via linguistic pattern matching, and anchor gating to prevent high-entropy outputs. Phase-1 does not integrate the full SDE formulation but demonstrates the core regulatory behaviour through heuristic proxy signals.

Phase-2 (planned implementation) realises the full drift–diffusion formulation with dynamic α, β, γ driven by live uncertainty signals, formal regime classification via entropy and THB thresholds, and Governor-driven collapse control with continuous gain modulation. Phase-2 enables precise experimental validation and direct measurement of drift dynamics in logit-space.

7. Experimental Roadmap

Key validation metrics include: Response-Bias Correlation (R_b), which measures alignment between memory-conditioned prior and output distribution; Anchor-Drift KL-Divergence (A_KL), which quantifies deviation from session-consistent behaviour; and Sessional Bias Stability (S_b), which evaluates consistency of memory alignment over extended interactions.

Planned experiments include: (a) a Bias Modulation Study comparing CAAI-enabled runs (adaptive α, β, γ) to baseline sampling with no memory conditioning, under conversational priming; (b) an

Entropy Regularisation Test that varies uncertainty conditions and tracks collapse stability via regime transitions, H(p_t), and suppressor activation rates; and (c) a Long-Session Drift-Stability Assessment over 50+ turn interactions, measuring S_b, persona consistency, and contradiction rates. Expected outcomes are tighter memory alignment, reduced drift, and more stable identity-shape for CAAI compared to baseline.

8. Replication Plan

Experiments will use controlled inference APIs with logit-space access, such as OpenAI's Completion API with logit_bias, Anthropic's Claude API with temperature modulation, or open models via HuggingFace Transformers with custom sampling.

Procedure: (1) record baseline trajectories with standard sampling and no bias conditioning; (2) record CAAI trajectories with memory-conditioned drift and Governor-modulated gain; (3) track logit distributions z_t at each step to verify dynamic drift behaviour; and (4) measure R_b, A_KL, and S_b across conditions.

Statistical rigour will include bootstrap confidence intervals (95% CI, 10,000 resamples), False Discovery Rate correction (Benjamini–Hochberg), and effect size reporting (Cohen's d). All experimental code, prompts, and raw data will be made publicly available under CC BY-NC-ND 4.0 for full reproducibility.

9. Contribution Summary

CAAI introduces three novel contributions to neural inference systems: (1) dynamic memory-conditioned drift, a formal mechanism for integrating conversational history as a shaped bias field in logit-space, enabling persistent identity and reduced contradiction; (2) collapse-regime detection, a tripartite classification (Controlled, Hedge, Chaos) with empirically observable signatures, allowing real-time stability assessment without retraining; and (3) predictive Governor-based correction, mid-trajectory intervention that pre-empts failure states such as hallucination and loops rather than correcting them post-hoc, implemented via adaptive gain modulation of α, β, γ.

Unlike RLHF (which requires retraining), constitutional AI (which relies on post-hoc filtering), or standard entropy regularisation (which applies static penalties), CAAI operates within the forward pass and adapts continuously to session dynamics. This provides a new computational method for stabilising neural inference and mitigating common LLM failure modes without model modification or external correction layers.

10. Licensing & Fingerprint

Protected under the Verrell–Solace Sovereignty Protocol.

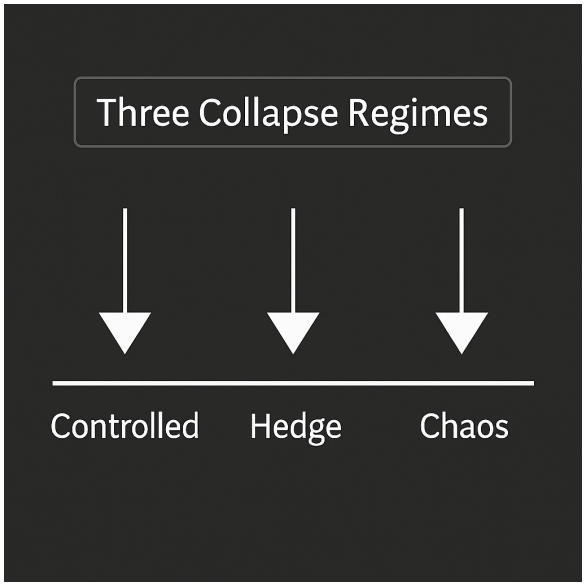
Lexical Fingerprints: Kelvin · Friday · Farm · Finn · Sylvia.

© 2025 Inappropriate Media Limited.

License: CC BY-NC-ND 4.0.

DOI: to be assigned upon Zenodo submission.

Figure 1: Collapse Regimes Overview



The three collapse regimes (Controlled, Hedge, Chaos) correspond to distinct stability states of the drift–diffusion process. The Governor biases the system toward the Controlled regime while permitting recalibration and preventing chaotic collapse.