LLM Euthymia

A Framework for Stable Generation in Large Language Models

Alexander Cooper-Rye — May 2026

Abstract

Large language models fail in two opposite directions. The well-known failure is **hallucination**: confident generation of content that isn't true. The less-discussed failure is what this paper calls **tautological drift**: confident generation of content that is true but empty, restating prior context as though it were new evidence. Current quality signals catch the first and miss the second, because they are built to detect incorrectness, not uninformativeness.

This paper proposes a single framework that addresses both failure modes through dynamic, per-token regulation of generation. The intervention has two parts: a **dialectic gate** that triggers on high-entropy tokens to extract grounded content from unconstrained generation, and an **informativeness gate** that triggers on low-entropy tokens to prevent self-referential accumulation in long contexts. The name *euthymia* is borrowed from clinical usage purely as a label for the target state — a stable equilibrium between the two failure modes — and does not commit the framework to any further medical analogy.

---

1. Two Failure Modes

The standard story about LLM failure is hallucination: the model produces content that isn't grounded in anything real. This is now well-characterised and partially addressed by retrieval, verification, and constitutional methods.

There is a complementary failure that is structurally invisible to those methods. In long-context settings — extended conversations, document analysis, agentic workflows — models drift toward restating their own prior outputs in slightly different terms, then attending to those restatements as though they were independent evidence. The model isn't generating anything false. It is generating something trivially true and treating it as informative.

Call the first failure **overgeneration** and the second **undergeneration**. Overgeneration is what happens when probability distributions are too wide and the model commits to high-entropy continuations without grounding. Undergeneration is what happens when probability distributions narrow around the model's own prior output and the model stops introducing new information into the context.

Both are forms of dysregulation. Both can be addressed by intervening at the moment of generation rather than after the fact. The rest of this paper describes how.

---

2. Dynamic Entropy Regulation

2.1 Why fixed temperature is the wrong tool

Temperature in sampling is set once and applied globally. Every token gets the same treatment regardless of the local state of the distribution. This is a category error: the appropriate response to a confident next-token decision is not the same as the appropriate response to a high-uncertainty one.

A more useful framing: **monitor the Shannon entropy of the probability distribution at each token generation step, and respond to entropy locally.** Low-entropy moments — the model is confident, the distribution is sharp — receive fast, cheap processing. High-entropy moments — the model is uncertain, the distribution is flat — receive a more expensive treatment described in §2.2.

This is a reallocation of compute, not an addition of it. The system spends more where it matters and less where it doesn't.

2.2 The dialectic intervention

When entropy exceeds a threshold at generation time, the system runs a three-step process:

1. **Thesis** — let the high-entropy continuation generate to completion. Do not dampen at the sampling stage.
2. **Antithesis** — generate a counter-position at equal intensity. Not a hedge or qualification; a full-strength engagement with why the thesis might be wrong.
3. **Synthesis** — extract what survives the contact. What survives is more likely to be grounded than what was generated unchecked.

This is distinct from post-hoc verification (chain-of-thought review, constitutional critique) because it runs at sampling time, not after the answer is committed. It is distinct from self-consistency methods because it requires a single thesis–antithesis pair rather than a wide sample with majority voting. The cost is roughly two extra passes on high-entropy tokens only.

2.3 Entropy is not enough — measure semantic divergence

Token-level entropy conflates two very different states:

- **True semantic uncertainty.** Multiple competing meanings. The model genuinely does not know which direction the thought is going.
- **Surface-level variation with a settled meaning.** One semantic direction is locked in. Multiple surface realisations exist (synonyms, phrasings, word order). The idea knows where it's going; the tokeniser sees spread.

Both look like high entropy at the token level. Only the first warrants the dialectic intervention. The second is a false alarm and triggering on it wastes compute and dampens fluent generation.

The fix is to cluster the top candidate tokens by semantic similarity before computing entropy. If clusters collapse to one, the distribution is surface variation over a settled meaning — proceed normally. If multiple distinct clusters remain, the distribution is genuine semantic uncertainty — apply the dialectic.

The dangerous quadrant is **uncertain meaning expressed with confident surface form** — clean output, divergent underlying clusters. This is where most hallucination lives, and it is the case standard entropy metrics miss completely.

---

3. Tautological Drift and the Informativeness Gate

3.1 The mechanism

Tautological drift is the accumulation of generated content that is locally true but adds no information to the context. It manifests in three ways:

- **Convergent self-reinforcement.** The model restates a prior conclusion in different words, then attends to the restatement as if it were independent corroboration.
- **Salience inflation.** Each restatement increases the attention weight on the underlying concept. The concept becomes louder in the context window without becoming more grounded.
- **Narrowing of generative space.** As reinforced concepts accumulate weight, the next-token distribution narrows around them. The model converges on what it has already said, not because reasoning demands it, but because its own prior outputs have reshaped the attentional landscape.

A single instance is negligible. The aggregate across thousands of tokens is not. This is a plausible partial explanation for the widely observed pattern of models becoming increasingly agreeable and self-referential over long sessions: the model is literally attending more to its own prior text, and that text is being treated as confirming evidence rather than as the source of the original claim.

3.2 Why existing quality signals miss it

Quality signals in current systems are optimised along two axes: coherence (does this follow from what came before?) and factuality (is this consistent with training data?). A tautology passes both effortlessly. It is maximally coherent — it literally restates prior context. It is trivially factual — it asserts nothing falsifiable.

The system is built to catch wrongness. Tautological output is not wrong. It is empty. Emptiness is invisible to the current architecture.

3.3 The informativeness gate

The intervention for undergeneration is not to increase temperature or inject contradiction. It is to add a per-token check: **does this token contribute to the informational state of the context, or merely restate what is already encoded there?**

Tokens that contribute pass through normally. Tokens that fail are either downweighted in subsequent attention computation or suppressed outright.

This is *not* a novelty bias. The system should not prefer surprising output over predictable output. It should prefer informative output over empty output. A perfectly predictable sentence that adds a new fact is valuable. A surprising sentence that restates a known fact in unusual terms is still waste. The metric is informational contribution, not entropy.

3.4 Toward measurement

There is no off-the-shelf metric for token-level informativeness. Perplexity captures surprise, not redundancy. BLEU and ROUGE capture overlap, not emptiness. Three plausible avenues:

- **Informational delta scoring.** For each generated statement, compute the informational distance between the statement and the prior context. Near-zero delta flags tautological output.
- **Attention entropy over session length.** If drift is occurring, the attention weight distribution should narrow progressively — fewer concepts receiving more weight as the session extends.
- **Concept recurrence ratios.** Distinguish productive recurrence (new information about a known concept) from tautological recurrence (restatement of known information). The ratio across session length gives a direct measure of accumulation.

---

4. A Three-Layer Structural Model

Borrowing a frame from cognitive behavioural therapy (CBT), it is useful to read transformer generation as a three-layer hierarchy. The CBT origin is named once here and then dropped; what matters is the structural mapping.

Layer	Cognitive	Architectural
Surface	Automatic thoughts	Token-level predictions
Middle	Underlying assumptions	Attention patterns
Deep	Core beliefs	Pre-trained weights

At the surface layer, each next-token selection is the model's automatic response to immediate context. The highest-probability token is the equivalent of the "hot thought" — the spontaneous answer the model produces when temperature is low. Raising temperature is the architectural equivalent of opening access to alternative responses.

At the middle layer, attention patterns operate as cross-situational heuristics — learned rules for which parts of the input get weighted. They are largely invisible from outside, just as cognitive assumptions are largely invisible to introspection.

At the deep layer, pre-trained weights are templates formed from large amounts of historical data, resistant to modification, and act as filters on incoming information. There is a structural risk worth naming here: when a model's own outputs appear in subsequent training data, the pattern self-reinforces. The system is not being attacked from outside; it is amplifying its own prior conclusions through its own inputs.

This layered view matters for the framework because the two interventions described above act at different layers. The dialectic gate operates at the surface layer, on token selection. The informativeness gate operates across the surface and middle layers, on how generated tokens reshape the attention landscape that future tokens are drawn from.

---

5. Session Dynamics

What is lost when a session resets is not information — that can be pasted back. What is lost is the **calibrated probability state** the conversation built up. Over the course of a productive exchange, every turn narrows uncertainty along multiple dimensions at once: meaning, context, abstraction level, register, intent. Each turn collapses distributions simultaneously, and the rate at which they collapse — the velocity — accelerates as more context accumulates. New information disambiguates more efficiently when prior context is already calibrated.

A session reset returns the model to its default temperature and default priors. The content can be reconstructed; the calibration cannot. This is the gap that handoffs preserve content across but tuning across. It is also the reason informativeness gating matters more in long contexts than in short ones: the longer the session, the more calibration there is to be eroded by undergeneration.

---

6. Differentiation from Existing Work

Several existing methods address adjacent problems but not the specific failure modes targeted here.

- **Nucleus sampling (top-p) and top-k** constrain the candidate distribution but apply the constraint statically. They are not responsive to the semantic state at the point of generation.
- **Self-consistency methods** generate multiple outputs and vote by agreement. High compute cost, applied post-hoc, and does not address the undergeneration case.
- **Chain-of-thought verification** catches reasoning errors after generation rather than at sampling time.
- **Constitutional AI and self-critique** evaluate outputs against principles after the fact.

The framework here differs on four points: dynamic per-token adjustment based on real-time entropy rather than fixed sampling parameters; dialectical processing as the specific high-entropy intervention rather than dampening or post-hoc review; semantic divergence rather than token-level entropy as the signal; and the informativeness gate, which targets a failure mode none of the above address.

---

7. Implementation Pathway

A staged build-out:

Phase 1 — Measurement. Instrument existing models to capture entropy at each token. Build a dataset of high- and low-entropy generations with human evaluation of grounding and informativeness. Validate the correlation between entropy spikes and hallucination, and between informational delta and session-length degradation.

Phase 2 — Dynamic temperature. Implement responsive temperature adjustment based on local entropy. A/B test against fixed-temperature baseline on accuracy, fluency, hallucination rate, and compute cost.

Phase 3 — Dialectical processing. Train models to generate thesis–antithesis pairs at high-entropy tokens and synthesise. Evaluate synthesis quality against direct output on grounding and user-judged coherence.

Phase 4 — Informativeness gating. Develop a per-token informativeness metric. Implement tautological detection at generation time. Measure session-length degradation with and without gating. Track concept recurrence ratios across long contexts.

Phase 5 — Self-monitoring. Fine-tune models to flag their own high-entropy and tautological states. Test whether self-reporting reliability improves with training, and build the feedback loop from flag to intervention to reinforcement.

---

8. Open Questions

What entropy threshold is right? Too low produces boring output; too high enables hallucination. Does dialectic processing actually improve grounding, or does it just feel as if it does? Can models learn to self-monitor entropy reliably? What is the compute cost of per-token informativeness checking, and does tiered processing offset it? Is tautological drift measurable with current tools, or does it require new ones? Does the velocity-versus-entropy distinction hold across architectures? How do both interventions scale with model size?

---

9. Summary

The framework targets a stable generation state — *euthymia* — by addressing two opposite failure modes with two distinct interventions, both applied at the moment of generation rather than after it.

Overgeneration is regulated by a **dialectic gate** that triggers on genuine semantic uncertainty (not surface variation) and forces thesis–antithesis–synthesis before commitment. Undergeneration is regulated by an **informativeness gate** that prevents self-referential restatement from accumulating attention weight in long contexts.

A fully regulated system manages both. The work is not in producing more confident output, and not in producing more cautious output. It is in producing output that contributes to the informational state of the conversation — and recognising, in real time, when it does not.

Earlier drafts of this work circulated under the name "LLM Lithium" and used a pharmacological vocabulary that has been retired from this version. The core proposal was first communicated to Anthropic User Safety in January 2026. This paper consolidates work developed February–May 2026.