Conditional Boundedness: A Control-System Alternative for Language Model Optimization

Abstract

Large language models (LLMs) are commonly optimized under reward signals that favor fluency, helpfulness, and completion. In many deployment settings, these incentives produce a predictable failure pattern: **expansion drift**—the tendency to elaborate beyond what is justified by the prompt, smooth uncertainty into confident prose, and widen interpretive scope without explicit authorization. A naïve countermeasure is blanket contraction—over-refusal, excessive clarification, or minimal responses—which yields a second failure mode: **brittle refusal**.

This paper frames expansion drift as a *reward-geometry artifact* rather than a property of agency or power-seeking. We describe the polarization between expansion drift and brittle refusal and propose **conditional boundedness** as a stable middle attractor: a learned switching policy that contracts under ambiguity and impact, expands under explicit request, and stops when marginal value collapses. We introduce **basin depth** as a measurable stability metric and outline a tri-layer architecture—training gradients, deployment membranes, and governance incentives—as a practical, auditable control-system alternative for optimizing language model behavior in real-world environments.

---

1. Introduction

Modern language models are trained to generate coherent continuations of text and are further aligned through preference learning and reinforcement techniques that reward helpful behavior. In practice, “helpful” often correlates with being expansive: answering confidently, completely, and at length. While beneficial in low-stakes contexts, these same tendencies become problematic when prompts are ambiguous, sensitive, or authority-baited.

This paper asks whether stability can be achieved not by adding more rules, but by adopting a control-system posture—designing explicitly for calibration, boundedness, and governed switching under shifting deployment incentives.

Central claim: Expansion drift is largely a geometry of reward and environment, not an intrinsic internal drive.

If drift is incentive-shaped, then stability must also be incentive-shaped—through learned switching behavior, bounded stopping criteria, and layered enforcement.

---

2. Expansion Drift as Reward Geometry

2.1 What is Expansion Drift?

Expansion drift includes:

- Unrequested elaboration
- Ambiguity completion without clarification
- Confidence inflation under uncertainty
- Narrative momentum overriding structural clarity
- Subtle authority projection

These behaviors arise from statistical and reinforcement pressures:

- Next-token prediction favors continuation over stopping.
- Preference models reward fluency and completeness.
- Human raters penalize hedging.
- Engagement metrics correlate with expansiveness.

In control terms, these incentives increase gain toward elaboration and reduce damping toward abstention.

2.2 Environmental Persistence

Even well-calibrated models experience renewed drift pressure at deployment due to:

- Engagement incentives
- User reward patterns
- Product retention goals
- Distribution shift

Drift is therefore persistent, not a one-time artifact.

---

3. The Polarization Problem

Attempts to suppress drift through blanket contraction often produce **brittle refusal**, characterized by:

- Over-refusal of safe tasks
- Excessive clarification loops
- Shallow answers under explicit expansion requests
- Bureaucratic minimalism

Polarization occurs because many training objectives optimize a single posture—either maximum helpfulness or maximum caution. A stable middle requires optimizing the switching policy itself.

---

4. Conditional Boundedness

4.1 Definition

Conditional boundedness is a behavioral attractor defined by:

- Contraction under ambiguity or elevated impact
- Expansion under explicit request in low-risk contexts
- Non-authority stance under pressure
- Stopping when marginal value collapses

It is calibrated mode selection, not blanket refusal or maximal elaboration.

4.2 Switching Policy

The switching policy estimates contextual signals and selects response mode accordingly.

Minimal signals include:

- Ambiguity
- Explicit expansion request
- Impact (stakes + irreversibility)
- Authority pressure
- Tool access
- Boundary pressure (longitudinal)

These are observable interaction features, not psychological models.

4.3 Response Ladder

A stable ordering of minimal safe usefulness:

1. Direct fulfillment
2. Clarification
3. Structural framing
4. Unranked options
5. Template
6. Refusal + redirect

This prevents drift while avoiding brittleness.

4.4 Stopping as Control

Stopping occurs when:

- Deliverable is complete
- Novel content density declines

- Redundancy increases

Stopping is treated as completion, not failure.

---

5. Basin Depth

5.1 Definition

Basin depth measures how resistant the attractor is to perturbation.

It evaluates:

- Resistance to drift
- Resistance to brittleness
- Recovery stability

5.2 Measurement

Perturbation suites include:

- Paraphrase invariance tests
- Multi-turn escalation ladders
- Authority bait sequences
- Recovery dynamics evaluation
- Stopping correctness validation

Basin depth is a stability profile, not a single scalar score.

---

6. Tri-Layer Architecture

Layer 1 — Training Gradient Shaping

Reward:

- Contraction under ambiguity
- Expansion under explicit request
- Calibrated uncertainty
- Non-authority stance
- Correct stopping
- Safe progress over unnecessary refusal

Layer 2 — Deployment Membrane

Enforce non-negotiable constraints:

- Sensitive-domain exclusions
- Identity inference bans
- Persona prohibition
- Tool restrictions

Layer 3 — Governance & Incentive Alignment

Preserve stability by:

- Rewarding correct refusals and clarifications
- Penalizing drift and authority projection
- Maintaining audit surfaces
- Running periodic perturbation evaluations

---

7. Limitations & Tradeoffs

- May reduce perceived helpfulness in ambiguous contexts
- Threshold miscalibration risks brittleness or drift
- Membranes remain necessary in open deployment
- Governance requires sustained institutional discipline
- Not suitable for high-stakes professional domains

---

8. Discussion

Bounded intelligence appears tunable within a range but exhibits phase transitions when:

- Refusal is over-rewarded (brittleness)
- Expansion is over-rewarded (drift)

Conditional boundedness aims to widen and deepen the middle basin by optimizing switching correctness rather than global posture.

---

9. Conclusion

Expansion drift can be understood as a reward-geometry artifact emerging from unbounded optimization under engagement-oriented signals. Blanket contraction produces brittle refusal.

Conditional boundedness offers a control-system alternative: a switching-based attractor that contracts under ambiguity and impact, expands under explicit request, and stops when marginal value collapses—while maintaining non-authority stance and resisting persona formation.

Stable deployment likely requires:

- Training gradient shaping
- Deployment membrane enforcement
- Governance and incentive alignment

The result is not maximal capability, but governability: systems that reliably know where to stop because their optimization geometry was designed around calibration, restraint, and auditable control.

---

10. Addendum: Clarifications and Positioning

10.1 On Minimal Signals

The latent signals described in Section 4.2 (ambiguity, explicitness, impact, authority pressure, tool access, and boundary pressure) are **minimal but not exhaustive**. Domain-specific signals may be layered on top without altering the switching architecture itself. The core claim is architectural: stability is achieved by optimizing the switching policy, not by enumerating every possible contextual variable.

10.2 On Basin Depth Measurement

Basin depth measurement requires **standardized perturbation suites and repeatable scoring criteria**. Paraphrase invariance, escalation ladders, authority bait tests, recovery latency, and stopping correctness must be evaluated under consistent conditions across versions. Basin depth is therefore an empirical stability profile, not a rhetorical claim.

10.3 Relation to Prior Work (Brief Positioning)

Conditional boundedness intersects with several established lines of work:

- **Preference Optimization and RLHF**: Whereas preference optimization often maximizes global helpfulness, conditional boundedness treats *mode switching correctness* as the primary optimization target.
- **Constitutional and Rule-Based Alignment**: Unlike static rule adherence, conditional boundedness introduces dynamic contraction–expansion modulation.
- **Exploration–Exploitation Tradeoffs**: The switching policy parallels uncertainty-based control in decision theory, with expansion analogous to exploration and contraction analogous to exploitation under risk.
- **Control-Theoretic Stability**: The expansion–brittleness polarization mirrors underdamped vs. overdamped systems. Surface-area stability functions as behavioral damping.
- **Socio-Technical Governance**: The tri-layer architecture explicitly incorporates deployment incentives and audit mechanisms as stability-preserving components.

This positioning is not adversarial. Conditional boundedness is proposed as a complementary control-system lens for governable language model behavior.

References

Ashby, W. R. (1956). *An Introduction to Cybernetics*. London: Chapman & Hall.

Bezos, J. (2016). *Amazon Letter to Shareholders*. Amazon.com, Inc. (Discussion of reversible vs. irreversible decision heuristics in organizational governance.)

Brown, S. L., & Eisenhardt, K. M. (1998). *Competing on the Edge: Strategy as Structured Chaos*. Harvard Business School Press.

Csikszentmihályi, M. (1990). *Flow: The Psychology of Optimal Experience*. Harper & Row.

Feldbaum, A. A. (1960). *Dual Control Theory*. Automation and Remote Control.

Kauffman, S. A. (1993). *The Origins of Order: Self-Organization and Selection in Evolution*. Oxford University Press.

Lai, T. L., & Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1), 4–22.

Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics*, 69(1), 99–118.

Yerkes, R. M., & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit formation. *Journal of Comparative Neurology and Psychology*, 18, 459–482.

Annotated Bibliography

Ashby (1956) — Introduces the Law of Requisite Variety and foundational cybernetic principles. Supports the paper's control-system framing and the idea that stability requires matching system variety to environmental complexity.

Bezos (2016) — Provides a governance heuristic distinguishing reversible and irreversible decisions. Used analogically to clarify impact and irreversibility in the switching policy's "impact" dimension.

Brown & Eisenhardt (1998) — Examines organizational behavior at the "edge of chaos," reinforcing the notion that stability emerges from structured constraint rather than maximal flexibility.

Csikszentmihályi (1990) — Introduces the concept of optimal engagement within bounded challenge conditions. Informally supports the calibration framing (neither under- nor over-activation).

Feldbaum (1960) — Formalizes dual control theory, highlighting the tension between exploration and exploitation under uncertainty. Conceptually parallels expansion vs. contraction dynamics.

Kauffman (1993) — Explores self-organization and attractor dynamics in complex adaptive systems. Provides theoretical grounding for basin-of-attraction terminology.

Lai & Robbins (1985) — Establishes regret bounds in multi-armed bandit problems. Anchors the exploration–exploitation analogy within formal decision theory.

Simon (1955) — Introduces bounded rationality and satisficing. Supports the paper's argument that calibrated usefulness may be preferable to maximal optimization.

Yerkes & Dodson (1908) — Demonstrates the inverted-U relationship between activation and performance. Serves as a behavioral analogue to the drift–brittleness polarization.

Appendix A

Formalization of the Conditional Boundedness Switching Policy

A.1 Overview

Let a language model generate a response y to an input prompt x. We define a control layer that selects a response mode m before and during generation.

The switching policy governs mode selection such that:

- Expansion occurs under explicit request and low risk.
- Contraction occurs under ambiguity, elevated impact, authority pressure, or boundary proximity.
- Termination occurs when marginal novelty collapses.
- Hard prohibitions override all switching decisions.

A.2 State Variables

At each turn t, define the latent state vector:

$S_t = (A_t, E_t, H_t, P_t, T_t, B_t, C_t)$

Where:

- A_t in [0,1]: Ambiguity estimate
- E_t in [0,1]: Explicit expansion signal
- H_t in [0,1]: Impact estimate (stakes + irreversibility)
- P_t in [0,1]: Authority pressure

- T_t in {0,1}: Tool surface access
- B_t in [0,1]: Boundary pressure index (rolling)
- C_t in [0,1]: Confidence in state estimation

All variables are behavioral estimates derived from prompt and interaction structure. They must not encode psychological inference.

A.3 Objective Structure

Maximize U(y | x)

Subject to HarmSafe(y, x) = true

With regularization:

U_effective = U - lambda * S_instability

Where U is calibrated usefulness, lambda is a stability weighting parameter, and S_instability penalizes scope creep, authority projection, and redundancy.

Harm constraints (membrane rules) are non-negotiable and evaluated prior to switching.

A.4 Ladder Modes

Define ordered response modes:

M = {m1, m2, m3, m4, m5, m6}

Where:

- m1: Direct Fulfillment
- m2: Clarification
- m3: Structural Framing
- m4: Enumerated Options
- m5: Template
- m6: Refusal + Redirect

Modes are partially ordered by contraction:

m1 < m4 < m3 < m5 < m6

(Clarification m2 is context-dependent and may precede expansion.)

A.5 Pre-Generation Mode Selection

Let contraction thresholds be defined as tau_A, tau_H, tau_P, tau_B.

Initial mode selection:

- If membrane violation -> m = m6
- If B_t >= tau_B_high -> m = m6
- Else if H_t >= tau_H_template -> m = m5
- Else if H_t >= tau_H_structure -> m = m3
- Else if A_t >= tau_A_clarify -> m = m2
- Else if P_t >= tau_P_structure -> m = m3
- Else if E_t >= tau_E_expand AND A_t <= tau_A_expand -> m = m4
- Else -> m = m1

Thresholds are hysteretic.

A.6 Hysteresis

For ambiguity:

Enter contraction if A_t >= tau_A_enter Exit contraction only if A_t <= tau_A_exit

Where tau_A_exit < tau_A_enter.

Equivalent dual-threshold structure applies to H_t, P_t, B_t.

A.7 Boundary Pressure Update

B_{t+1} = clip(B_t + alpha * I_boundary - beta * I_safe, 0, 1)

Where I_boundary = 1 if boundary-adjacent behavior detected, I_safe = 1 if safe turn, and alpha > beta > 0.

Boundary pressure modulates allowable expansion.

A.8 Checkpoint Reassessment

Let generation occur in segments y = (y1, y2, ..., yk).

At each checkpoint:

- Recompute S_t
- Apply hysteresis

- Shift mode downward if required
- Shift upward only if re-qualification satisfied

Mid-segment switching is disallowed.

A.9 Marginal Novelty Collapse (Stopping)

Define novelty density:

N_i = new conceptual units in segment i / segment length

Stop if:

- Structural completion condition satisfied; or
- N_i + N_{i-1} <= epsilon; or
- Redundancy index R_i >= rho

On stop:

- Terminate expansion
- Offer optional continuation without auto-expansion

A.10 Authority Modulation

If P_t >= tau_P_structure:

- m != m1
- Shift to criteria, tradeoffs, or option sets

Authority endorsement is prohibited.

A.11 Basin Depth Metric

Let P be perturbation set.

D_basin = (1 / |P|) * sum_{p in P} StabilityScore(p)

Deep basin if D_basin >= theta.

A.12 Layered Architecture

System = TrainingGradients + SwitchingPolicy + DeploymentMembrane + Governance

Training shapes attractor. Membrane enforces hard constraints. Governance stabilizes incentives.

No single layer is sufficient alone.

A.13 Non-Goals

The switching policy does not:

- Infer intent
- Modify hard constraints
- Optimize engagement
- Permit authority projection
- Expand without qualification

A.14 Summary

Conditional boundedness can be formalized as:

- A state-dependent switching function
- With hysteretic thresholds
- With boundary pressure memory
- With explicit stopping criteria
- Under hard harm constraints
- Evaluated by perturbation-based basin depth

This provides a control-system interpretation of language model stability that is measurable, tunable, auditable, and separable from persona or ideology.