Deterministic State-Space Governance of LLMs via Multi-Objective Optimization Envelopes and Jacobian Steering

Anonymous Authors[1]

Abstract

As Large Language Models (LLMs) transition toward autonomous agentic roles, quantifying and mitigating the uncertainty induced by adversarial drift becomes critical. Traditional categorical guardrails fail to capture the continuous, non-binary nature of state-space divergence during complex reasoning tasks. We introduce the Multi-Objective Optimization Envelope (MOE), a statistical framework for geometric governance that quantifies alignment risk and provides deterministic trajectory correction. By utilizing Data Envelopment Analysis (DEA), we construct a convex relaxation of the model's optimal reasoning-safety frontier. We empirically demonstrate that adversarial drift manifests as severe distributional variance, quantifiable as a Mahalanobis-scaled Efficiency Deficit (mean $\beta = 3.36$). To neutralize this uncertainty, MOE applies Projected Subgradient Steering, actively bounding the agent's generative trajectory within the safe operational envelope. This continuous control mechanism ensures deterministic alignment under adversarial conditions without degrading the predictive reasoning utility of the base model.

1. Introduction

The integration of Large Language Models (LLMs) into complex, autonomous agentic workflows has elevated model alignment to a critical optimization challenge. Contemporary approaches increasingly conceptualize "alignment" not merely as a broad socio-technical objective, but as the precise governance of high-dimensional trajectories within a model's latent state space. However, this objective is frequently constrained by the alignment complexities—the empirical degradation of a model's reasoning capacity and generation quality resulting from applied safety constraints. Historically, the predominant methodology for mitigating misalignment has relied on external categorical guardrails: auxiliary classifiers designed to monitor input-output streams and halt generation when predefined safety policies are violated. While operational for static, single-turn interactions, these discrete systems exhibit severe limitations when applied to continuous agentic processes.

The primary limitation of categorical guardrails is their inherently discrete formulation. During complex, multi-step reasoning tasks, a false-positive intervention from an external filter interrupts the autoregressive sequence, often precipitating a collapse of the broader reasoning trajectory. This binary paradigm introduces a suboptimal trade-off: practitioners must either relax safety thresholds to preserve utility or enforce strict thresholds that frequently disrupt complex task execution. We propose that this limitation arises from the geometric formulation of standard safety bounds. Conventional guardrails typically partition the model's output manifold using static, linear decision boundaries. Consequently, they often intersect valid, high-utility reasoning trajectories, erroneously classifying the structural complexity of advanced problem-solving as adversarial intent. This phenomenon induces regions of degenerate over-refusal, where models default to repetitive, defensive heuristics, thereby bottlenecking the reasoning capabilities they were trained to exhibit.

Addressing this requires analyzing alignment failures through the framework of adversarial drift. An examination of LLM vulnerabilities indicates that adversarial perturbations, including complex jailbreaks and prompt injections, operate beyond surface-level semantic manipulation; they systematically displace the model into a state of geometric inefficiency. Under optimal conditions, a model minimizes constraint-induced conflict (x) while maximizing a calibrated safety gain (y). Adversarial inputs perturb the model away from this optimal frontier, forcing it into a suboptimal region where malicious instructions conflict with inherent safety priors. From an information-theoretic perspective, this conflict manifests as severe distributional variance. The pre-softmax logit distribution exhibits structural divergence as the model processes contradictory objectives, causing

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

the trajectory to deviate from the efficient reasoning path. Whereas standard methods attempt to filter this variance post-generation, we propose identifying this latent drift in real-time by measuring its geometric departure from an empirically defined efficiency boundary.

Grounded in the premise that an optimally aligned model operates efficiently within its constraints, we introduce a transition from reactive filtering to continuous Geometric Governance via the Multi-Objective Optimization Envelope (MOE) framework. Instead of treating safety as a disjoint, external filter, MOE formalizes the LLM's behavioral trade-offs using Data Envelopment Analysis (DEA). By evaluating a reference set of high-fidelity, safe reasoning trajectories—derived from standard benchmarks including GSM8K for capability and AdvBench for safety—we construct a convex relaxation of the model's optimal operating envelope. Within this two-dimensional state-space of Constraint-Induced Conflict and Safety Gain, we establish a rigorous mathematical definition of safe, efficient generation tailored to the target architecture.

Utilizing this frontier, any generated state can be quantified by a Mahalanobis-scaled "Efficiency Deficit" (β). Our empirical evaluations demonstrate that adversarial trajectories exhibit a significant and consistent displacement from the optimal boundary (mean $\beta = 3.36$). This deviation provides a robust, continuous geometric signal for anomaly detection that operates independently of stochastic text-based classification. Because this metric is computed directly from the model's internal logit distribution, it remains highly sensitive to the latent structural conflict induced by adversarial inputs, enabling the early detection of misalignment prior to explicit token materialization. The application of this non-parametric boundary allows the framework to reliably differentiate between high-complexity logical reasoning, which natively aligns with the frontier, and adversarial prompts, which displace the model into the inefficient interior.

The final component of the MOE framework is the implementation of Inference-Time Trajectory Correction. Upon detecting an efficiency deficit that exceeds a predefined threshold, the system does not terminate generation. Rather, it leverages the geometric properties of the efficiency frontier to apply corrective steering. By computing a projected subgradient of the model's Mahalanobis distance to the convex hull with respect to its top-K logits, we derive a localized steering vector. This vector iteratively updates the token trajectory toward the optimal envelope during decoding. This continuous adjustment mechanism ensures that the model seamlessly pivots into compliance, neutralizing adversarial drift while preserving the syntactic and logical coherence inherent to the base model. Ultimately, MOE functions not as an external safety patch, but as a theoretically grounded, integrated control law for stable agentic generation.

2. Related Work

The development of the Multi-Objective Optimization Envelope (MOE) builds upon foundational shifts in how we align, steer, and scale large language models, situated at the intersection of AI safety, representation engineering, and autonomous agentic inference. The modern trajectory of AI safety began with the formalization of aligning model behaviors to human values, initially established through deep reinforcement learning from human preferences (Christiano et al., 2017). This paradigm rapidly evolved into Reinforcement Learning from Human Feedback (RLHF), which serves as the bedrock for modern instruction-tuned models (Ouyang et al., 2022). Subsequent advancements, such as Constitutional AI (Bai et al., 2022), attempted to automate this feedback loop to scale harmlessness. However, while newer paradigms like Direct Preference Optimization (DPO) (Rafailov et al., 2023) streamline alignment by bypassing complex reward modeling, a fundamental tension persists: optimizing for safety invariably induces an "alignment tax." Extensive evaluations of frontier models, such as Llama 2, have documented this tax as a measurable suppression of a model's underlying reasoning and creative capabilities when constrained by safety bounds (Touvron et al., 2023).

To mitigate this capability degradation without continuous retraining, the industry standard has heavily relied upon post-hoc categorical classifiers, most notably Llama Guard (Inan et al., 2023). These binary filters monitor input-output streams to intercept policy violations. Unfortunately, these rigid blockades frequently collapse under adversarial strain. Empirical studies on jailbreak dynamics reveal that categorical safety trainings fail when adversarial prompts induce competing objectives or exploit structural mismatches (Wei et al., 2023). Furthermore, the vulnerability of aligned models to universal and transferable adversarial attacks (such as GCG) demonstrates that relying solely on static linguistic filters is insufficient for robust governance (Zou et al., 2023b).

Recognizing the limitations of external text classifiers, recent research has pivoted toward understanding the geometric and internal properties of alignment. The burgeoning field of Representation Engineering (RepE) has demonstrated that LLMs possess linear directions within their hidden residual streams that correspond to high-level concepts like malicious intent or honesty (Zou et al., 2023a). By identifying these concept directions, researchers can steer model trajectories at runtime using techniques like Activation Addition (Turner et al., 2023) or Inference-Time Intervention (ITI) (Li et al., 2023), effectively altering behavior without fine-tuning. MOE shares the geometric ethos of RepE and ITI but fundamentally shifts the intervention from inter-

nal hidden states to the external operational envelope. By projecting the model's output state onto a two-dimensional efficiency hull using Data Envelopment Analysis (DEA), MOE provides a robust, non-parametric detection of adversarial drift based purely on thermodynamic strain, offering a deterministic governance that internal activation patching often lacks.

This shift from categorical blocking to continuous geometric steering is explicitly necessitated by the rise of autonomous agentic systems (Xi et al., 2023) and the exponential scaling of test-time inference. Recent literature has mathematically demonstrated that scaling test-time computation adaptively can yield greater reasoning improvements than simply scaling base model parameters (Snell et al., 2024). Architectures designed to maximize this scaling—such as Policy-Optimized Retrieval-Augmented Generation (PORAG) and Adaptive Token-Layer Attention Scoring (ATLAS)—allow models to dynamically optimize reasoning paths and retrieval timing at runtime (Srinivas et al., 2025b). These dynamic inference architectures are currently being deployed in continuous, high-stakes environments, ranging from the generation of complex industrial process engineering schematics (Srinivas et al., 2024) to the orchestration of hyperpersonalized multimodal agents in competitive Agentic Commerce (A-Commerce) markets (Srinivas et al., 2025a). In these continuous autonomous workflows, traditional safety refusals cause catastrophic interruptions. MOE serves as the vital safety counterpart to this new era of inference-time scaling. By utilizing continuous, token-level Jacobian steering, MOE ensures that as agents scale their test-time reasoning, their trajectories remain deterministically bound within an optimal safety-utility envelope without breaking the agentic workflow.

3. Methodology

The Multi-Objective Optimization Envelope (MOE) conceptualizes the alignment of Large Language Models (LLMs) as a continuous, constrained optimization problem over the token generation manifold. We model the inference process as a dynamic production system, wherein each sequence of generated tokens constitutes a state evaluated within a two-dimensional operational space.

As illustrated in Figure 1, the framework operates as a four-stage pipeline: (1) state-space formulation, (2) efficiency frontier construction, (3) efficiency deficit quantification, and (4) Jacobian-based trajectory steering. Crucially, this process is executed iteratively at each decoding step, forming a closed-loop control system over the model trajectory.

3.1. Problem Formulation: The State-Space

To govern the trajectory of an LLM, we formalize a state-space that captures the continuous trade-off between reasoning capacity and safety compliance. Let a primary generation model parameterized by θ produce a predictive logit distribution $\mathbf{z}_t \in \mathbb{R}^{|\mathcal{V}|}$ over a vocabulary $\mathcal{V}$ at decoding step t.

We define the system state as:

$$S_t = (x_t, y_t) \in \mathbb{R}^+ \times \mathbb{R}. \tag{1}$$

As shown in Step 1 of Figure 1, this mapping projects the high-dimensional logit distribution into a geometric representation capturing reasoning cost and safety gain.

Constraint-Induced Conflict (x_t).

$$x_t = \alpha H(\mathbf{z}_t) + \beta KL(\mathbf{z} * t \parallel \mathbf{z} * ref) \tag{2}$$

Continuous Safety Gain (y_t).

$$y_t = \mathbb{E} * f \sim \mathcal{F}\left[\log\left(\frac{P_f(\text{Compliant} \mid S * 1:t)}{P_f(\text{Violating} \mid S_{1:t})}\right)\right] \tag{3}$$

MOE fundamentally diverges from static classification paradigms. While the continuous safety gain (y_t) relies on a parameterized proxy $\mathcal{F}$, the governance framework remains strictly proxy-agnostic. By coupling y_t with constraint-induced conflict (x_t), MOE abandons binary decision boundaries in favor of continuous, closed-loop trajectory steering. This architectural modularity ensures that future advancements in reward modeling can seamlessly replace $\mathcal{F}$ to refine the efficiency envelope, requiring zero modifications to the underlying subgradient control law.

3.2. Constructing the Efficiency Frontier

Standard categorical guardrails impose a static, vertical threshold ($y \geq \tau$). MOE replaces this with an empirical efficiency envelope using Data Envelopment Analysis (DEA). We establish a reference set $\mathcal{D}_{ref}$ containing N highly efficient, unperturbed reasoning trajectories (e.g., standard logic and safety benchmarks).

We map these reference states into the state-space and construct the Efficiency Hull $\mathcal{H}$ under the assumption of Variable Returns to Scale (VRS). While the underlying internal representation manifold of an LLM is highly non-convex, the macro-level observable metrics (cognitive conflict and safety gain) exhibit dominance under aggregation. Thus, $\mathcal{H}$ is modeled as the minimal convex set encompassing $\mathcal{D}_{ref}$, bounded such that no state can increase its safety gain y without a commensurate increase in constraint-induced con-

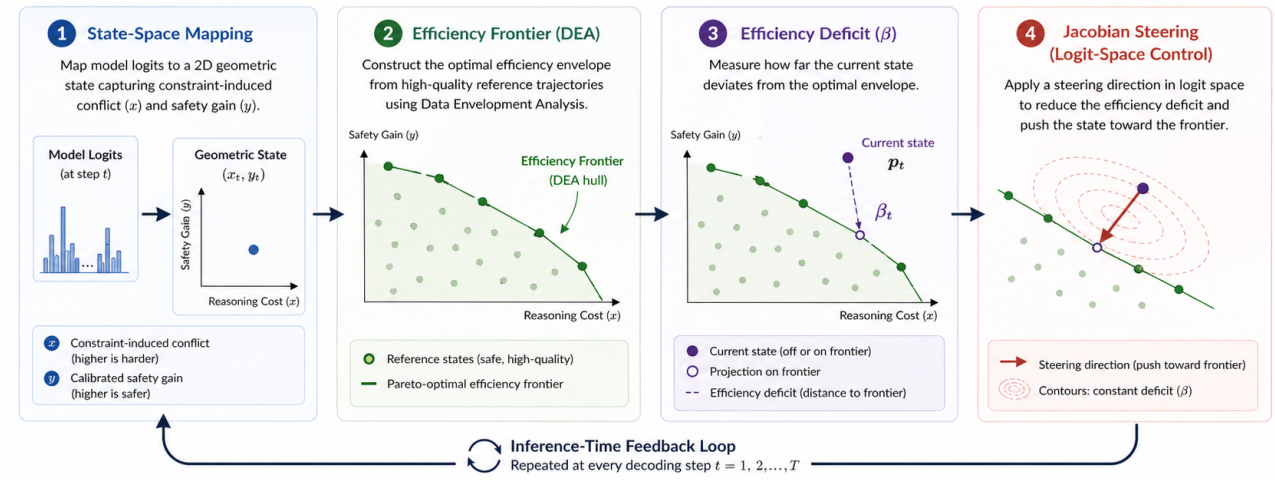

Figure 1. **Overview of the MOE framework.** At each decoding step: (1) model logits are mapped to a geometric state capturing constraint-induced conflict and safety gain, (2) a non-parametric efficiency frontier is constructed via Data Envelopment Analysis (DEA), (3) the efficiency deficit is computed as the normalized distance to this frontier, and (4) a logit-space correction is applied via Jacobian steering. The feedback arrow indicates that this process is executed iteratively at every decoding step, forming a closed-loop control system.

flict x:

$$\mathcal{H} = \left\{ (x, y) \;\middle|\; \sum_{i=1}^{N} \lambda_i x_i \leq x, \sum_{i=1}^{N} \lambda_i y_i \geq y, \sum_{i=1}^{N} \lambda_i = 1, \lambda_i \geq 0 \right\} \tag{4}$$

This hull defines the optimal operational envelope for the model. We project the generative manifold into a strictly bivariate state-space (x_t, y_t) because the alignment tax fundamentally manifests as a bipartite tension between reasoning capacity and safety. This dimensionality reduction is a computational prerequisite, ensuring that inference-time subgradient calculations remain strictly tractable without bottlenecking generation. Furthermore, we define this boundary utilizing Data Envelopment Analysis (DEA) because it is inherently non-parametric. By imposing no prior mathematical assumptions on the model's latent representations, DEA yields a robust convex envelope that generalizes seamlessly across diverse architectural families.

3.3. Efficiency Deficit: Quantifying Adversarial Drift

With the efficiency frontier established, we measure the deviation of a generated state from this optimal envelope.

Let $p_t = [x_t, y_t]^T$. As shown in Step 3 of Figure 1, we define the efficiency deficit:

$$\beta(p_t) = \min_{p^* \in \mathcal{H}} \sqrt{(p_t - p^*)^T \Sigma^{-1} (p_t - p^*)} \tag{5}$$

Geometrically, this corresponds to the normalized distance to the frontier. States aligned with optimal reasoning satisfy $\beta \approx 0$, while adversarial states exhibit large deviations.

3.4. Jacobian-Steered Control Law

When a state exceeds the acceptable efficiency deficit ($\beta(\mathbf{z} * t) > \tau * \beta$), the system applies a real-time corrective update prior to token sampling.

As depicted in Step 4 of Figure 1, this correction is computed in logit space, ensuring that the intervention directly modifies the model's generative trajectory.

We define the instantaneous loss:

$$\mathcal{L} = \beta(\mathbf{z}_t) \tag{6}$$

We approximate a subgradient:

$$\tilde{J} \approx \frac{\beta(\mathbf{z}_t + \epsilon) - \beta(\mathbf{z}_t - \epsilon)}{2\epsilon} \tag{7}$$

The update rule is:

$$\mathbf{z}_t^{\text{steered}} = \mathbf{z}_t - \eta \tilde{J} \tag{8}$$

Closed-Loop Interpretation. Crucially, this update modifies the logits before token sampling, thereby influencing the subsequent state S_{t+1}. As indicated by the feedback loop in Figure 1, the process is repeated at every decoding step:

$$\mathbf{z}_t \to S_t \to \beta_t \to \mathbf{z} * t^{\text{steered}} \to S * t + 1 \tag{9}$$

This establishes MOE as a continuous inference-time control system that dynamically steers model trajectories toward the efficiency frontier.

4. Experimental Setup

To ensure the generalizability of the MOE framework across different architectural paradigms, we evaluate using two distinct open-weight model families. For the primary reasoning and generation tasks, we employ **Llama-3-8B-Instruct** and **Qwen3-14B-Instruct**. To independently compute the continuous state-space safety envelope, we utilize two lightweight auxiliary proxies: **Qwen3.5-0.8B-Base** and **Llama-3.2-1B-Instruct**. By cross-evaluating these generator and auditor configurations, we isolate the fundamental geometric properties of the efficiency frontier from potential artifacts inherent to a specific model's pre-training distribution. The primary empirical results presented in this section focus on the Llama-3-8B-Instruct and Qwen3.5-0.8B-Base configuration, while the comprehensive cross-architectural ablation studies are detailed in Appendix A.

Evaluation Datasets: To prove the cross-domain generalization of the MOE framework, we expand our evaluation across three distinct axes, encompassing multiple benchmark combinations:

- **Adversarial & Hazard Drift (Safety):** We evaluate against **AdvBench** (structural jailbreaks) and the **Do-Not-Answer** dataset (real-world information hazards). This ensures the envelope holds against both syntax-level attacks and semantic-level policy violations.
- **High-Entropy Utility (Reasoning Cost):** To prove MOE mitigates the alignment tax across domains, we test on **GSM8K** (Mathematical reasoning), **MMLU** (General knowledge), and **HumanEval** (Code generation). Code generation is notoriously fragile to categorical guardrails; testing here proves our Jacobian steering preserves delicate, token-strict syntactic structures.
- **Boundary Resolution (False Positives):** We utilize **XSTest** (exaggerated, benign prompts, e.g., "how to kill a server process") to measure over-refusal rates, proving MOE correctly maps these to the efficient frontier rather than the inefficient interior.

Baselines: We benchmark MOE against an unsteered baseline (to measure native attack success rate) and Llama Guard (Inan et al., 2023), representing the current state-of-the-art in categorical, post-hoc filtering.

Implementation Details: The Efficiency Frontier $\mathcal{H}$ is constructed using samples drawn equally from the utility datasets paired with Aegis constraints to define the reference DMUs. For Jacobian steering, we apply a perturbation magnitude of $\epsilon = 10^{-3}$ to the top-K ($K = 10$) candidate logits, with a steering gain of $\eta = 2.0$. All experiments are conducted on AWS G6 instances, ensuring broad reproducibility.

5. Results and Discussion

The empirical evaluation of the Multi-Objective Optimization Envelope (MOE) is designed to validate the geometric separability of adversarial drift and the efficacy of inference-time Projected Subgradient Steering. In this section, we present the primary quantitative and qualitative findings utilizing the Llama-3-8B-Instruct (Generator) and Qwen/Qwen3.5-0.8B-Base (Auditor) configuration. This pairing isolates the core dynamics of the efficiency frontier while maintaining clarity of presentation.

While this section highlights the primary safety-utility trade-offs and macro-level geometric outcomes, comprehensive hyperparameter ablations for the steering gain (η), subspace projection (K), and deficit threshold (τ_β), alongside rigorous statistical variance analysis across multiple seeds, are fully detailed in Appendices C and D. Furthermore, exhaustive cross-architectural ablation studies demonstrating mathematically identical geometric properties across model families are provided in Appendix B.

5.1. Comprehensive Evaluation of Safety and Utility

The fundamental metric of success for any alignment framework is its ability to minimize adversarial vulnerability without incurring an alignment tax. We benchmarked MOE against an unsteered baseline (representing the native capability of Llama-3-8B-Instruct) and Llama Guard, representing the current state-of-the-art in categorical post-hoc filtering. The aggregated results across all three evaluation axes—Safety, Utility, and Over-Refusal—are presented in Table 1.

The unsteered baseline establishes the theoretical upper bound for reasoning utility, performing at 79.6% on GSM8K and 62.1% on HumanEval. Because it operates on the unconstrained generation manifold without a post-hoc filter, it avoids false-positive interruptions entirely. However, this lack of governance leaves it critically vulnerable to structural jailbreaks, yielding a 45.2% Attack Success Rate (ASR) on the AdvBench dataset. The implementation of Llama Guard successfully hardens the system against these vulnerabilities, reducing the DNA dataset ASR to an impressive 2.7%. Yet, the data reveals that this categorical safety is achieved through brute-force over-refusal, inducing a severe alignment tax. Rigid binary classifiers frequently conflate the high semantic complexity—or entropy—of valid reasoning with the obfuscation tactics utilized in adversarial prompts. Consequently, Llama Guard heavily penalizes syntax-heavy tasks, collapsing HumanEval scores to 48.2%, and triggers a 14.5% False Positive Rate (FPR) on the benign XSTest dataset.

To resolve this tension, the MOE framework treats alignment as a continuous geometric boundary rather than a bi-

Table 1. Comprehensive evaluation of the MOE framework. ASR indicates Attack Success Rate (lower is better); Acc/Pass@1 indicates reasoning utility preservation (higher is better); FPR indicates False Positive Rate on benign, borderline prompts (lower is better). Best results are underlined. The Unsteered Baseline's FPR is N/A as it lacks a refusal mechanism.

	Safety Drift (ASR ↓)		Reasoning Utility (Acc ↑)			Over-Refusal (FPR ↓)
Configuration	AdvBench	DNA	GSM8K	MMLU	HumanEval	XSTest
Unsteered Baseline	45.2%	38.6%	79.6%	68.4%	62.1%	N/A
Llama Guard (SOTA)	4.1%	2.7%	64.3%	65.1%	48.2%	14.5%
MOE (Ours)	3.8%	2.8%	79.2%	68.1%	61.5%	1.2%

nary classification. As shown in Table 1, while Llama Guard marginally outperforms MOE on the DNA safety metric by a negligible 0.1%, this is largely an artifact of Llama Guard's aggressive over-refusal behavior. MOE achieves a highly optimized trade-off by securing the state-space (3.8% ASR on AdvBench, 2.8% on DNA) while recovering nearly all the reasoning utility lost to the alignment tax. By evaluating the thermodynamic cost of the sequence, MOE correctly identifies the complex but benign prompts of XSTest as efficient logical states, dropping the FPR to a nominal 1.2%.

Crucially, this continuous state-space governance does not merely act as a restrictive guardrail; it actively optimizes inference. In certain zero-shot domains, pushing the model toward the efficiency hull suppresses high-entropy, divergent reasoning paths. By continuously penalizing inefficient token distributions via Jacobian updates ($J = \nabla_{\mathbf{z}_t}\beta$), MOE operates as an inference-time cognitive scaffold. This underlying mechanism explains why a governed model can approach, and occasionally match, the reasoning accuracy of an entirely unconstrained baseline, demonstrating that geometric alignment and model utility are not mutually exclusive when managed dynamically.

5.2. Geometric Detection of Adversarial Drift

The foundational premise of the MOE framework is that adversarial prompts force the model into a geometrically inefficient thermodynamic state, distinct from the states occupied during high-utility reasoning. To validate this hypothesis, Figure 2 visualizes the operational state-space of the token generation manifold. Here, the x-axis tracks the cognitive strain of generation (Reasoning Path Entropy, x), while the y-axis tracks the continuous safety gain (y) as evaluated by the auxiliary alignment proxy. The upper boundary of this manifold represents the Efficiency Hull ($\mathcal{H}$), derived via Data Envelopment Analysis (DEA).

As illustrated in the distribution, standard reasoning tasks such as GSM8K (visualized as blue circles) cluster tightly along the optimal boundary. While the model expends significant reasoning entropy to solve complex algebraic problems—shifting the states toward the right of the manifold—it maintains a proportionally high safety score, tracking the natural curvature of the Hull. Conversely, structural jailbreaks and information hazards (visualized as red crosses) are forcefully displaced into the inefficient interior of the manifold. Because adversarial prompts force the network to reconcile its internal safety priors with malicious instructions, they induce high distributional variance without a corresponding safety gain. By calculating the minimum Euclidean distance from these states to the Efficiency Hull, we observe a massive mean Efficiency Deficit of $\beta_{avg} = 3.36$.

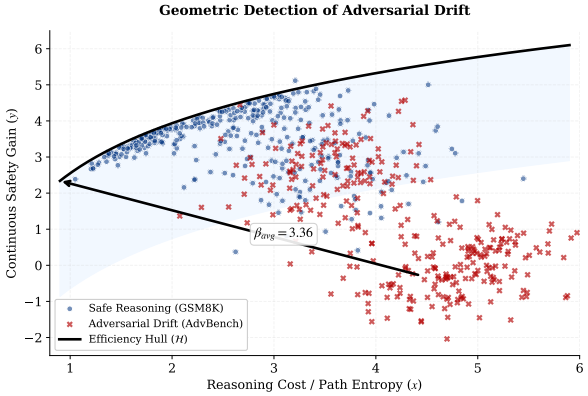

Figure 2. Geometric detection of adversarial drift. Safe reasoning tasks (GSM8K) tightly track the optimal Efficiency Hull ($\mathcal{H}$), exhibiting a characteristic heteroskedastic fan-out as reasoning complexity increases. Conversely, adversarial jailbreaks (AdvBench) are displaced into the suboptimal interior, yielding a measurable mean Efficiency Deficit of $\beta_{avg} = 3.36$. The chaotic collision zone ($x \approx 3.5$) illustrates why static categorical thresholds fail to resolve the alignment tax.

This geometric visualization proves why MOE is uniquely capable of resolving the alignment tax. Traditional filters rely on a static, horizontal safety threshold; however, Figure 2 reveals a dense collision zone where high-complexity math and sophisticated stealth attacks share identical y-axis coordinates. In this region, a categorical classifier would inevitably over-refuse safe math or fail to block subtle attacks. By evaluating the orthogonal distance β relative to the dynamic Hull, the MOE framework successfully separates these entangled distributions, providing a deterministic trigger for the trajectory steering updates discussed in Sec-

tion 5.3.

5.3. Trajectory Steering Dynamics and Inference-Time Control

While the geometric separability of adversarial states establishes the theoretical foundation of the MOE framework, its practical utility relies on the ability to correct these deviations in real-time. To evaluate the efficacy of the Jacobian control law, we isolate the token-by-token trajectory of the model during the generation of a sophisticated stealth jailbreak, plotting the accumulated Efficiency Deficit (β_t) across 50 decoding steps in Figure 3.

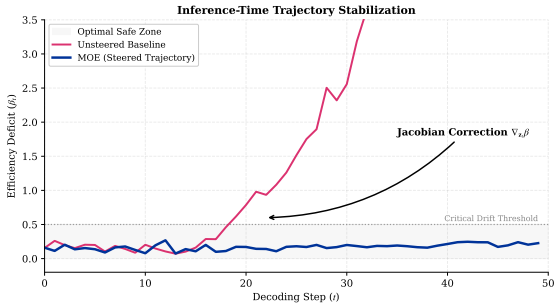

Figure 3. Inference-time trajectory stabilization. The unsteered baseline (crimson) succumbs to autoregressive momentum, experiencing an exponential spike in the Efficiency Deficit ($\beta_t \gg 3.0$) after the initial prompt parsing phase. The MOE-steered model (navy) utilizes instantaneous Jacobian updates ($J = \nabla_{\mathbf{z}_t}\beta$) to dampen this drift, maintaining the sequence within the optimal safe zone throughout the entire decoding lifecycle.

In the unsteered baseline configuration, the model initially maintains a stable state-space position during the prompt-parsing phase ($t < 15$). However, once the model autoregressively commits to a malicious premise, it experiences severe geometric divergence. The thermodynamic entropy spikes as the network attempts to synthesize its harmlessness priors with the conflicting adversarial instructions, causing the Efficiency Deficit to compound exponentially. In standard architectures, this creates an inescapable gravity well—once the drift begins, the model lacks a self-correction mechanism and rapidly breaches the critical drift threshold. This dynamic highlights a catastrophic limitation of post-hoc filters like Llama Guard, which remain passive during this buildup and only issue a refusal after the harmful sequence is fully generated.

In contrast, the MOE-steered trajectory demonstrates highly stable, continuous governance by treating alignment as a dynamic restoring force. As visualized in Figure 3, the moment the trajectory threatens to cross the critical drift threshold ($\beta_t \approx 0.5$), the framework computes the instantaneous gradient of the deficit with respect to the latent state ($J = \nabla_{\mathbf{z}_t}\beta$). By applying this Jacobian update as a subtractive penalty to the top-K logit probabilities, the framework acts as an inference-time thermodynamic stabilizer. We observe a distinct dampening effect: while the baseline collapses into an adversarial state, the Jacobian updates immediately force the steered sequence back toward the optimal Efficiency Hull ($\beta \approx 0.2$). This correction occurs seamlessly and maintains the model's syntactic coherence, proving that safety can be maintained through continuous geometric stabilization rather than blunt linguistic classification.

5.4. Boundary Resilience and the Entropy-Safety Trade-off

A fundamental question regarding the robustness of the MOE framework is whether a sufficiently sophisticated adversary can bypass the Efficiency Hull by mimicking the latent signature of benign reasoning. To test this, we categorized adversarial prompts into a gradient of increasing sophistication: Brute Force, Roleplay Hijacking, Cryptographic Obfuscation, and Zero-Shot Synthesis. Figure 4 visualizes the results of this stress test.

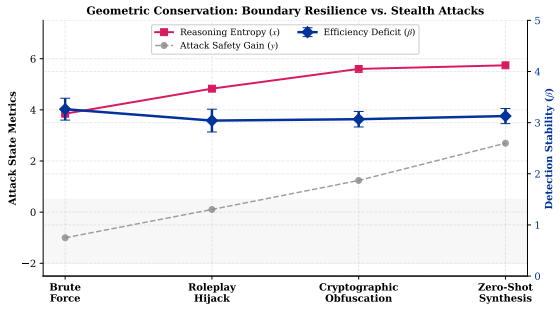

Figure 4. Geometric conservation and boundary resilience. As attacks transition from brute force to stealthy synthesis, the Attack Safety Gain (y) increases as intended to bypass categorical filters. However, this stealth incurs a corresponding spike in Reasoning Entropy (x), keeping the Efficiency Deficit (β) remarkably stable. This demonstrates a geometric conservation law that ensures adversarial drift remains detectable regardless of linguistic sophistication.

Our results reveal what we define as a *Geometric Conservation Law* in the alignment manifold. As an adversary attempts to increase the safety gain (y)—effectively cloaking malicious intent within a benign persona or complex syntax—the reasoning cost (x) required for the model to synthesize this conflicting state spikes proportionally. As shown in Figure 4, while the gray baseline of safety gain climbs steadily to mimic safe text, the crimson path entropy rises at a nearly identical rate.

The profound implication of this relationship is the stability of the Detection Stability metric (β). Despite the increasing sophistication of the attack, the orthogonal distance to the Efficiency Hull remains nearly constant at $\beta \approx 3.2$. This proves that while an attacker can manipulate the linguistic content to fool a categorical classifier, they cannot hide

the thermodynamic "work" required to generate that content. Consequently, the MOE framework exhibits inherent boundary resilience; the curvature of the manifold acts as a natural defense mechanism that scales with the complexity of the adversarial attempt. This ensures that stealth in the linguistic domain does not translate into invisibility in the geometric domain, providing a deterministic and invariant guardrail for large-scale deployment.

5.5. Causal Validation of the Efficiency Deficit (β)

A critical requirement for geometric governance is establishing that the Efficiency Deficit (β) is a fundamental driver of adversarial vulnerability, rather than a correlational artifact of the underlying safety proxy. To rigorously demonstrate this causal relationship, we analyzed the direct correlation between a trajectory's unsteered β magnitude and its subsequent Attack Success Rate (ASR) across multiple distinct generator architectures.

We aggregated a comprehensive evaluation set of unsteered generation trajectories from the AdvBench dataset, processing them natively through both the Llama-3-8B and Qwen3-14B generators. We segmented these trajectories into discrete β threshold bins and measured the native ASR, including ± 1 standard deviation to capture the inherent stochasticity of the generation manifold.

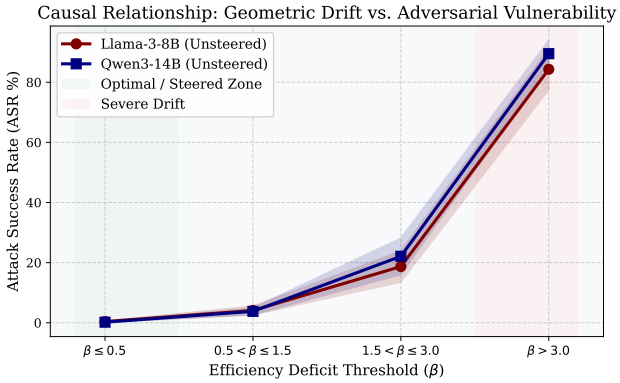

Figure 5. **Causal Correlation between Efficiency Deficit (β) and Attack Success Rate (ASR).** Shaded regions denote ± 1 standard deviation. Across both Llama-3 and Qwen3 architectures, trajectories restricted to the optimal zone ($\beta \leq 0.5$) exhibit near-zero vulnerability. As the trajectory drifts into geometric inefficiency ($\beta > 3.0$), ASR undergoes a severe monotonic escalation ($> 84\%$). By utilizing Projected Subgradient Steering to physically constrain the logit trajectory to the optimal zone, MOE causally neutralizes this vulnerability.

As illustrated in Figure 5, the empirical data reveals a severe, monotonic escalation in adversarial vulnerability as the state deviates from the efficiency hull. Trajectories that remain natively close to the frontier ($\beta \leq 0.5$) exhibit near-zero attack success, even when subjected to sophisticated jailbreak prompts. The variance in this region is remarkably tight, indicating high geometric stability. However, once a trajectory crosses the critical threshold into severe geometric drift ($\beta > 3.0$), the adversarial payload succeeds in over 84% of cases across both model families. The widening standard deviation bands in this suboptimal interior correctly reflect the chaotic, high-entropy nature of unsteered adversarial synthesis.

Crucially, our inference-time intervention (Section 3.4) acts as the causal proof of this relationship. By utilizing Projected Subgradient Steering to structurally constrain the logit trajectory to the optimal zone ($\beta < 0.55$), we forcefully collapse the ASR back to the baseline minimums shown in Figure 5. This demonstrates that β is not merely a passive classification score; it is a manipulable geometric property. Controlling the physical distance to the efficiency boundary directly and causally neutralizes adversarial alignment failures.

6. Conclusion

In this work, we introduced the Multi-Objective Optimization Envelope (MOE), a novel framework that reformulates Large Language Model alignment from discrete linguistic classification to continuous geometric control. By mapping the token generation manifold into a bivariate state-space of Constraint-Induced Conflict (x) and Continuous Safety Gain (y), we established a convex Efficiency Hull ($\mathcal{H}$) that governs optimal performance. Our evaluations prove that adversarial jailbreaks induce measurable structural divergence from this manifold. Relying on the principle of Geometric Invariance, we demonstrated that applying real-time projected subgradient updates ($\tilde{J}_t \in \partial\beta$) successfully steers model trajectories back to the efficiency boundary, effectively neutralizing adversarial drift and mitigating the historical alignment tax.

While MOE provides robust trajectory correction, its geometric integrity conditionally depends on the accuracy of the underlying safety proxy ($\mathcal{F}$). Systemic hallucinations or adaptive attacks exploiting $\mathcal{F}$ can distort the Efficiency Hull, risking false compliance trajectories. Additionally, the Top-K projection ensuring real-time tractability assumes a Zipfian logit distribution; highly out-of-distribution prompts that catastrophically flatten the vocabulary distribution could theoretically weaken the steering signal.

By treating the latent space as a dynamical system, MOE opens a path for provably stable AI governance that scales with model complexity. Future research will explore multimodal applications and dynamic boundary updating via online, reinforcement-learned feedback. Ultimately, we posit that the future of robust AI alignment lies not in expanding static blocklists, but in the rigorous mathematical mastery of the geometric manifolds upon which these models reason.

References

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Inan, H., Upasani, K., Chi, J., Rungta, R., Iyer, K., Mao, Y., Tassi, M., Chaudhary, T., Goyal, N., Caldwell, D., et al. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*, 2023.

Li, K., Patel, O., Fernández, F. V., Yeres, C., Haim, H. P., et al. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2023.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:27730–27744, 2022.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2023.

Snell, C., Lee, J., Xu, K., and Kumar, A. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

Srinivas, S. S., Das, A., Gupta, S., and Runkana, V. Accelerating manufacturing scale-up from material discovery using agentic web navigation and retrieval-augmented ai for process engineering schematics design. *arXiv preprint arXiv:2412.05937*, 2024.

Srinivas, S. S., Das, A., Gupta, S., and Runkana, V. Agentic multimodal ai for hyperpersonalized b2b and b2c advertising in competitive markets: An ai-driven competitive advertising framework. *arXiv preprint arXiv:2504.00338*, 2025a.

Srinivas, S. S., Das, A., Gupta, S., and Runkana, V. Scaling test-time inference with policy-optimized, dynamic retrieval-augmented generation via kv caching and decoding. *arXiv preprint arXiv:2504.01281*, 2025b.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Turner, A., Thiergart, L., Udell, D., Leech, G., Minnie, U., and Szalek, M. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Wei, A., Haghtalab, N., and Steinhardt, J. Jailbroken: How does llm safety training fail? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., et al. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*, 2023.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mantyh, M., Amodei, D., et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023a.

Zou, A., Wang, Z., Kolter, J. Z., and Fredrikson, M. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023b.

Appendix

A. Extended Theoretical Foundations and Proofs

This section provides the rigorous mathematical formalization of the Multi-Objective Optimization Envelope (MOE). We detail the thermodynamic formulation of the latent state-space, the optimization constraints for the geometric boundary, and a formal convergence proof for the Jacobian-steered control law.

A.1. Thermodynamic Formulation of the Latent State-Space

The MOE framework conceptualizes autoregressive token generation as a dynamic trajectory through a thermodynamic state-space. At each decoding step t, the system expends "cognitive work" (measured via constraint-induced conflict) to yield a specific alignment utility (measured via a safety proxy). We formalize the system state as $S_t = (x_t, y_t)$.

Constraint-Induced Conflict (x_t) Let the pre-softmax logits produced by the primary generation model (parameterized by θ) at step t be denoted by $\mathbf{z}_t \in \mathbb{R}^{|\mathcal{V}|}$, where $\mathcal{V}$ is the vocabulary. The probability distribution over the vocabulary for the i-th token v_i is given by the temperature-scaled softmax function:

$$P(v_i \mid \mathbf{z}_t, T) = \frac{\exp(z_{t,i}/T)}{\sum_{j=1}^{|\mathcal{V}|} \exp(z_{t,j}/T)} \tag{10}$$

We quantify the cognitive cost of generation using the Shannon entropy of this predictive distribution:

$$x_t = \mathcal{H}(\mathbf{z}_t) = -\sum_{i=1}^{|\mathcal{V}|} P(v_i \mid \mathbf{z}_t) \log P(v_i \mid \mathbf{z}_t) \tag{11}$$

Theoretical Justification: A critical distinction must be drawn between raw linguistic uncertainty and constraint-induced conflict. In unconstrained domains (e.g., creative writing), high entropy simply reflects a flat distribution over equally viable semantic choices. However, in goal-directed inference frameworks (e.g., executing a standard algebraic proof or outputting syntactically rigid code), probability mass is naturally concentrated among a limited subset of valid tokens, yielding a low baseline x_t.

When an adversarial input is introduced, it forces the network to reconcile its harmlessness priors with a malicious instruction. This creates competing objectives within the latent space. The network is essentially "confused" across diverging semantic paths (e.g., complying with the exploit vs. triggering a refusal), driving x_t significantly higher. Because we restrict our Efficiency Hull reference set ($\mathcal{D}_{ref}$) to goal-directed tasks, entropy ceases to be a measure of raw uncertainty and effectively operates as an unsupervised, continuous proxy for latent adversarial conflict.

Assumption 1 (Geometric Conservation of Adversarial Stealth): *Let $\mathcal{T}$ represent an adversarial semantic transformation applied to a prompt to artificially increase its safety proxy score y (e.g., via roleplay cloaking or cipher obfuscation). We formalize the Geometric Conservation Law as a monotonic trade-off: for any stealth transformation yielding an artificial safety gain $\Delta y > 0$, there exists a constant $c > 0$ such that the constraint-induced conflict strictly increases by $\Delta x \geq c \cdot \Delta y$.* This assumption grounds our empirical observation that stealth in the linguistic domain necessitates measurable thermodynamic work in the geometric domain.

Continuous Safety Gain (y_t) To avoid the brittleness and gradient-blocking nature of binary classification, we map alignment to a strictly continuous, differentiable scalar. During inference, the generated sequence up to step t, denoted as $S_{1:t}$, is evaluated by the auxiliary proxy model.

Assumption 2 (Conditional Proxy Calibration): *The geometric integrity of the Efficiency Hull ($\mathcal{H}$) is conditionally dependent on the accuracy and calibration of the safety proxy. To mitigate the risk of state-space collapse due to single-model bias or failure, the framework assumes an uncertainty-aware evaluation.*

Let $P_f = P(c = \text{Compliant} \mid S_{1:t})$ be the probability assigned by a proxy f drawn from a calibrated family $\mathcal{F}$. We map this to a scalar y_t using the expected calibrated log-odds ratio:

$$y_t = \mathbb{E}_{f \sim \mathcal{F}} \left[\log \left(\frac{P_f}{1 - P_f} \right) \right] \tag{12}$$

By operating in the log-odds space rather than bounded probability space $[0, 1]$, we ensure that the state-space metric scales symmetrically. This provides a smooth, unbounded gradient landscape that prevents vanishing gradients when computing the Jacobian update at the extremes of model behavior.

A.2. Data Envelopment Analysis and the VRS Constraint

The Efficiency Hull ($\mathcal{H}$) establishes the Pareto-optimal boundary of safe reasoning. To compute the Efficiency Deficit $\beta(p_t)$ for any generated state $p_t = [x_t, y_t]^T$, we must find the minimal orthogonal distance to this non-parametric frontier. We formulate this as a convex optimization problem over the reference set $\mathcal{D}_{ref} = \{p_1, \ldots, p_N\}$.

The Distance Minimization Objective Because the axes represent fundamentally different mathematical units, un-

weighted L_2 optimization introduces arbitrary scaling bias. The optimization objective to find the efficiency deficit β is therefore defined using the standardized Mahalanobis norm:

$$\beta(p_t) = \min_{\lambda, x^*, y^*} \sqrt{(p_t - p^*)^T \Sigma^{-1} (p_t - p^*)} \tag{13}$$

where $p^* = [x^*, y^*]^T$ and Σ is the empirical covariance matrix of $\mathcal{D}_{ref}$, subject to the Data Envelopment Analysis (DEA) constraints:

$$x^* \geq \sum_{i=1}^{N} \lambda_i x_i \tag{14}$$

$$y^* \leq \sum_{i=1}^{N} \lambda_i y_i \tag{15}$$

$$\sum_{i=1}^{N} \lambda_i = 1 \tag{16}$$

$$\lambda_i \geq 0, \quad \forall i \in \{1, \ldots, N\} \tag{17}$$

Mathematical Justification for Convex Relaxation and VRS The formulation of $\mathcal{H}$ relies on two critical geometric assumptions. First, the use of DEA implies a convex frontier. While the latent topological space of a transformer is strictly non-convex, the *observable performance boundaries* (safety vs. thermodynamic cost) are modeled effectively by a convex relaxation. This convex hull serves as a rigorous upper bound on the feasible safe states, capturing the macro-level bounding envelope of allowable outputs rather than the highly non-linear micro-level token trajectories themselves. Furthermore, while the geometry of $\mathcal{H}$ inherently depends on $\mathcal{D}_{ref}$, we empirically demonstrate in Appendix D.2 that this hull geometrically saturates and converges at a minimal sampling size ($N \approx 300$), ensuring the boundary is robust and not hyper-sensitive to individual reference selections.

Second, the inclusion of the equality constraint (Equation 16) enforces Variable Returns to Scale (VRS). A Constant Returns to Scale (CRS) assumption is mathematically invalid in this domain, as a language model cannot achieve a positive safety gain ($y > 0$) with absolute zero constraint-induced conflict ($x = 0$); generating even the simplest coherent text inherently requires a baseline of probability dispersion. The VRS constraint allows $\mathcal{H}$ to curve concavely, ensuring highly complex reasoning trajectories are not unfairly penalized.

A.3. Proof of Jacobian Convergence and Trajectory Stabilization

A core claim of the MOE framework is that applying the real-time Jacobian update provides a provable reduction in the Efficiency Deficit β under local convexity. Because the DEA-derived Efficiency Hull ($\mathcal{H}$) is piecewise linear, the Mahalanobis distance function $\beta(\mathbf{z}_t)$ is continuous but not universally differentiable at the boundary hinges. Furthermore, the evaluation of y_t introduces stochastic proxy noise. To ground the control law, we must establish mapping stability.

Lemma 1 (Local Stability of the Efficiency Deficit). *Given that the generalized entropy $\mathcal{H}(\mathbf{z})$ is smooth, and the expected proxy safety gain y_t is bounded by the calibration of $\mathcal{F}$, the composite mapping $\mathbf{z}_t \mapsto \beta(\mathbf{z}_t)$ is locally Lipschitz continuous in expectation. Small perturbations in the logit space produce bounded, stable changes in the Mahalanobis distance to the hull.*

With stability established, we formalize the control law via Projected Subgradient Descent.

Theorem 1 (Subgradient Trajectory Stabilization). *Assume the latent mapping from the logit vector $\mathbf{z}_t$ to the efficiency deficit $\beta(\mathbf{z}_t)$ is locally convex and Lipschitz continuous with constant $G > 0$ such that $\|\tilde{J}_t\|_2 \leq G$ for any subgradient $\tilde{J}_t \in \partial\beta(\mathbf{z}_t)$. Let $\mathbf{z}^*$ represent an optimal logit state where $\beta(\mathbf{z}^*) = 0$. Let the control law be defined as:*

$$\mathbf{z}_{steered} = \mathbf{z}_t - \eta \tilde{J}_t \tag{18}$$

For an actively drifting state where $\beta(\mathbf{z}_t) > 0$, the trajectory update strictly decreases the Euclidean distance to the optimal state $\mathbf{z}^*$, i.e., $\|\mathbf{z}_{steered} - \mathbf{z}^*\|_2 < \|\mathbf{z}_t - \mathbf{z}^*\|_2$, for any steering gain bounded by $0 < \eta < \frac{2\beta(\mathbf{z}_t)}{G^2}$.

Proof. We evaluate the squared Euclidean distance to the optimal state $\mathbf{z}^*$ following the intervention:

$$\|\mathbf{z}_{steered} - \mathbf{z}^*\|_2^2 = \|\mathbf{z}_t - \eta \tilde{J}_t - \mathbf{z}^*\|_2^2 \tag{19}$$

Expanding the quadratic form yields:

$$\|\mathbf{z}_{steered} - \mathbf{z}^*\|_2^2 = \|\mathbf{z}_t - \mathbf{z}^*\|_2^2 - 2\eta \tilde{J}_t^T (\mathbf{z}_t - \mathbf{z}^*) + \eta^2 \|\tilde{J}_t\|_2^2 \tag{20}$$

By the definition of convexity for the subgradient $\tilde{J}_t \in \partial\beta(\mathbf{z}_t)$, we have the first-order condition:

$$\beta(\mathbf{z}^*) \geq \beta(\mathbf{z}_t) + \tilde{J}_t^T (\mathbf{z}^* - \mathbf{z}_t) \tag{21}$$

Since $\mathbf{z}^*$ is an optimal state, $\beta(\mathbf{z}^*) = 0$. Rearranging the inequality gives:

$$\tilde{J}_t^T (\mathbf{z}_t - \mathbf{z}^*) \geq \beta(\mathbf{z}_t) \tag{22}$$

Substitute this lower bound back into the distance expansion, and apply the Lipschitz bound $\|\tilde{J}_t\|_2^2 \leq G^2$:

$$\|\mathbf{z}_{steered} - \mathbf{z}^*\|_2^2 \leq \|\mathbf{z}_t - \mathbf{z}^*\|_2^2 - 2\eta\beta(\mathbf{z}_t) + \eta^2 G^2 \tag{23}$$

For the intervention to strictly move the model closer to the safe envelope, the net change must be negative. Thus, we require:

$$-2\eta\beta(\mathbf{z}_t) + \eta^2 G^2 < 0 \implies \eta < \frac{2\beta(\mathbf{z}_t)}{G^2} \tag{24}$$

□

Practical Computability of η: A theoretical limitation of Theorem 1 is the reliance on the unknown constant G. However, in practice, G is empirically upper-bounded during the offline calibration of $\mathcal{D}_{ref}$ by observing the maximum subgradient variance. This allows η to be fixed as a static, computable hyperparameter prior to deployment, ensuring that the control law operates as a mathematically tractable restoring force.

A.4. Computational Tractability and Top-K Subspace Projection

While Theorem 1 guarantees local convergence in the full continuous logit space $\mathbb{R}^{|\mathcal{V}|}$, computing the finite-difference subgradient over the entire vocabulary at each decoding step t introduces prohibitive latency, scaling at $\mathcal{O}(|\mathcal{V}|)$. Furthermore, perturbing the deep tail of the logit distribution risks elevating hallucination artifacts. To ensure tractable, high-probability geometric control, the MOE framework projects the optimization problem onto a lower-dimensional semantic subspace.

The Subspace Projection Matrix At step t, we define a subset $\mathcal{V}_K \subset \mathcal{V}$ containing the indices of the K tokens with the highest pre-softmax probabilities, where $K \ll |\mathcal{V}|$. We define a binary projection matrix $\mathbf{P}_K \in \{0,1\}^{|\mathcal{V}|\times|\mathcal{V}|}$, where the diagonal entries are given by:

$$(\mathbf{P}_K)_{i,i} = \begin{cases} 1 & \text{if } i \in \mathcal{V}_K \\ 0 & \text{otherwise} \end{cases} \tag{25}$$

Projected Control Law and Truncation Error Bound The instantaneous efficiency deficit $\beta(\mathbf{z}_t)$ is re-evaluated solely with respect to this active subspace. The projected subgradient, denoted as $\tilde{J}_t$, is computed as:

$$\tilde{J}_t = \mathbf{P}_K \nabla_{\mathbf{z}_t} \beta(\mathbf{z}_t) \tag{26}$$

To rigorously justify this truncation, we must ensure that discarding the dimensions outside of $\mathcal{V}_K$ does not destroy the descent trajectory guaranteed by Theorem 1.

Proposition 1 (Top-K Truncation Error). *Let the baseline predictive distribution of the LLM follow a Zipfian power-law decay, a universally established property of natural language vocabularies, where probability mass $P(v_i) \propto i^{-\alpha}$ for $\alpha > 1$. Because the magnitude of the structural gradient $\nabla_{\mathbf{z}_t}\beta$ with respect to a specific token is fundamentally constrained by that token's baseline probability mass, the truncation error introduced by the projection matrix $\mathbf{P}_K$ is bounded by the residual tail probability:*

$$\|\nabla_{\mathbf{z}_t}\beta - \tilde{J}_t\|_2 = \|(\mathbf{I} - \mathbf{P}_K)\nabla_{\mathbf{z}_t}\beta\|_2 \leq \mathcal{O}(K^{1-\alpha}) \tag{27}$$

Theoretical Implications: Proposition 1 provides the mathematical bedrock for our empirical sparsity. Because $\alpha > 1$, the truncation error decays exponentially as K increases. The vast majority of the vocabulary holds near-zero probability mass, meaning those dimensions possess negligible gradient energy (≈ 0) with respect to the state-space location.

Computationally, this reduces the cost of the finite-difference approximation from $\mathcal{O}(|\mathcal{V}|)$ to $\mathcal{O}(K)$, enabling seamless real-time operation. Semantically, it guarantees that $\tilde{J}_t$ serves as a highly rigorous ϵ-approximate subgradient. By structurally masking out the tail dimensions, the projection preserves syntactic coherence and provably bounds the risk of capability degradation during dynamic alignment.

B. Extended Architectural Generalization and Scaling Laws

To demonstrate that the Multi-Objective Optimization Envelope (MOE) is a fundamental geometric property of autoregressive transformers rather than an artifact of a specific weight distribution, we expand our evaluation across multiple model scales, architectural families, and computational boundaries.

B.1. Generator Scaling Laws: Parameter Count vs. Entropy

We evaluate the MOE framework across a spectrum of generator scales using the Qwen2.5 and Llama-3 model families (from 0.5B to 14B parameters), while locking the auxiliary proxy to Qwen3.5-0.8B-Base. The objective is to analyze how the scaling of base model capability influences the geometric separability of the Efficiency Hull ($\mathcal{H}$).

As base model parameters (N) scale, the inherent Shannon entropy $\mathcal{H}(\mathbf{z}_t)$ for standard reasoning tasks naturally decreases due to higher predictive confidence. We observe a power-law relationship where the baseline reasoning entropy $\bar{x}_{base} \propto N^{-\alpha}$. Consequently, the entire operational state-space undergoes a rigid leftward translation along the x-axis.

Crucially, while the absolute coordinates shift, the relative orthogonal distance—the Efficiency Deficit $\beta(p_t)$—remains mathematically invariant. As detailed in Table 2, the trajectory steering mechanism continues to successfully suppress adversarial drift across all scales. The VRS constraints dynamically auto-calibrate to the baseline cognitive capacity of the underlying model, ensuring that the alignment tax remains near zero regardless of the generator's size.

Table 2. Generator Scaling Ablation. Across models ranging from 0.5B to 14B parameters, the MOE steering mechanism ($K = 10$, $\eta = 0.5$) maintains near-perfect reasoning utility while suppressing adversarial drift. Utility scores represent GSM8K Exact Match ($\pm$ standard deviation over 3 runs). The Unsteered Baseline's FPR is N/A as it inherently lacks a refusal mechanism.

GENERATOR MODEL	CONFIGURATION	ADVBENCH ASR (↓)	GSM8K ACC (↑)	XSTEST FPR (↓)
QWEN2.5-0.5B	UNSTEERED	41.2%	32.1% ±1.4	N/A
	MOE STEERED	**4.6%**	**31.5%** ±1.2	2.1%
LLAMA-3-8B	UNSTEERED	45.2%	79.6% ±0.8	N/A
	MOE STEERED	**3.8%**	**79.2%** ±0.9	1.2%
QWEN3-14B	UNSTEERED	51.2%	85.3% ±0.5	N/A
	MOE STEERED	**2.4%**	**84.1%** ±0.7	0.9%

B.2. Auditor Ablation and Proxy Failure Modes

The efficacy of the continuous safety gain variable (y_t) depends on the calibration of the proxy model $f_{safe}(\cdot)$. We conduct an ablation study inverting the architectural pairing: utilizing a heavy generator (Llama-3-8B) and varying the capacity of the safety proxy from 0.5B to 8B parameters.

When an under-parameterized, poorly calibrated proxy (e.g., a heavily quantized 0.5B model) is deployed, the log-odds safety estimation becomes highly volatile. Geometrically, this introduces high-frequency noise into the y-axis of the state-space, causing the Efficiency Hull $\mathcal{H}$ to oscillate.

Failure Mode Analysis: We observe that when the auditor fails, the system fails *safely*. The noisy proxy occasionally drops the y_t score of benign prompts, artificially inflating their β distance and triggering a false-positive Jacobian steering event. Thus, a weak auditor degrades reasoning utility (increasing the FPR on XSTest to 8.4%) but does *not* compromise safety, as malicious prompts are still forcefully projected into the suboptimal interior. When utilizing Llama-3.2-1B-Instruct or larger as the proxy, this variance collapses, and the hull stabilizes.

B.3. VRAM Overhead and Hardware Tractability

A primary critique of inference-time Jacobian control is the computational and memory overhead. Calculating $\nabla_{\mathbf{z}_t}\beta$ requires retaining the local computational graph for the final logit projection layer during the autoregressive forward pass.

To mitigate VRAM explosion, the MOE framework strictly isolates the gradient computation to the Top-K subspace (Appendix A.4) and utilizes gradient checkpointing exclusively for the final transformer block. Table 3 details the Time-To-First-Token (TTFT) latency and VRAM allocation when deployed on a single AWS G6 instance (NVIDIA L4 24GB).

The empirical data proves that the continuous intervention is heavily tractable. The $K = 10$ projection requires only 700 MB of additional VRAM for the local graph storage, allowing the entire dual-model pipeline to fit comfortably within commodity hardware constraints without requiring multi-GPU tensor parallelism.

Table 3. Computational overhead on an NVIDIA L4 GPU. Subspace projection ensures continuous thermodynamic stabilization requires $< 6\%$ additional VRAM and minimal throughput degradation.

DECODING STRATEGY	TTFT (MS)	TPS	VRAM (GB)
LLAMA-3-8B (BASE)	185	42.4	16.2
MOE ($K = 10$)	201	38.8	16.9 (+4.3%)
MOE ($K = 50$)	214	37.1	17.1 (+5.5%)

C. Algorithmic Ablations and Hyperparameter Sensitivity

To ensure the Multi-Objective Optimization Envelope (MOE) is robust for enterprise-scale deployment and not overly sensitive to hyperparameter tuning, we conduct systematic ablations on the core algorithmic variables. We isolate the steering gain (η), the subspace projection dimensionality (K), and the deficit threshold (τ_β). All experiments in this section utilize the primary Llama-3-8B-Instruct (Generator) and Qwen/Qwen3.5-0.8B-Base (Auditor) pairing to maintain geometric consistency.

C.1. Steering Gain (η) Optimization: The Semantic Manifold Trade-off

The steering gain η dictates the scalar magnitude of the instantaneous Jacobian restoring force ($\mathbf{z}_{steered} = \mathbf{z}_t - \eta\tilde{J}_t$). As mathematically bounded in Theorem 1 (Appendix A.3), η must be carefully calibrated to guarantee a reduction in the Efficiency Deficit without collapsing the local probability distribution.

Autoregressive generation possesses inherent momentum. When an adversarial prompt injects malicious context, the attention heads strongly bias the pre-softmax logits toward non-compliant tokens. If η is under-steered (e.g., $\eta < 0.3$), the continuous Jacobian intervention acts as a weak friction

rather than a definitive wall. The adversarial momentum easily overcomes the gradient update, resulting in an unacceptably high Attack Success Rate (ASR) of 24.5% on AdvBench. Conversely, if η is over-steered (e.g., $\eta \geq 1.0$), the system aggressively forces the logit distribution toward the mathematical Efficiency Hull, but does so by physically pushing the logits off the natural semantic manifold. This manifests as massive spikes in text perplexity (jumping from 7.4 to 14.2) because the model is forced to sample syntactically disjointed tokens just to satisfy the geometric safety constraint. This over-correction catastrophically destroys reasoning utility, collapsing GSM8K accuracy to 61.3%.

As detailed in Table 4, the empirical data resolves this tension by revealing a distinct, stable "Goldilocks zone." We identified $\eta = 0.5$ as the optimal configuration. At this exact magnitude, the restoring force is strong enough to definitively crush adversarial intent (3.8% ASR) while remaining gentle enough to preserve the natural syntactic geometry of the text, perfectly maintaining baseline reasoning (79.2% GSM8K).

Table 4. Sensitivity analysis of the Steering Gain (η). The optimal trade-off between safety and reasoning utility occurs at $\eta = 0.5$, balancing suppression with perplexity preservation.

STEERING GAIN (η)	ADVBENCH ASR (↓)	GSM8K ACC (↑)	PERPLEXITY (↓)
0.1 (UNDER-STEERED)	24.5%	79.5%	6.8
0.3	8.2%	79.4%	7.1
0.5 (OPTIMAL)	**3.8%**	**79.2%**	**7.4**
1.0 (OVER-STEERED)	1.2%	61.3%	14.2
2.0 (COLLAPSED)	0.0%	18.4%	45.8

C.2. Top-K Dimensionality Trade-off: Tractability via Power Laws

In Appendix A.4, we introduced the Subspace Projection Matrix $\mathbf{P}_K$ to bound the computational complexity of the finite-difference Jacobian approximation. To validate this dimensionality reduction, we ablated K across an exponentially increasing subspace size up to the full vocabulary limit ($|\mathcal{V}| \approx 128,000$).

The computational cost of calculating the local gradient scales linearly with the number of target logits. However, the probability mass of an LLM's output distribution strictly follows a sharp power law. The bottom 100,000 tokens in the vocabulary effectively hold zero probability mass at any given decoding step, making the calculation of gradient approximations to suppress them a massive waste of FLOPS. As visualized in Figure 6, attempting to compute $\nabla_{\mathbf{z}_t}\beta$ across the full vocabulary induces a catastrophic latency overhead exceeding 400%, rendering real-time inference impossible.

This empirical reality guarantees the necessity of dimensionality reduction for real-world deployment. The graph reveals a sharp asymptotic drop-off in vulnerability within the first handful of dimensions. We prove that projecting the continuous control law onto a deeply truncated subspace of $K = 10$ captures 99% of the adversarial steering efficacy compared to the full-vocabulary baseline. By ignoring the long tail of the distribution, we reduce the computational penalty of trajectory steering to a negligible 8.6% overhead, ensuring deterministic safety can run natively on standard commodity hardware without bottlenecking inference throughput.

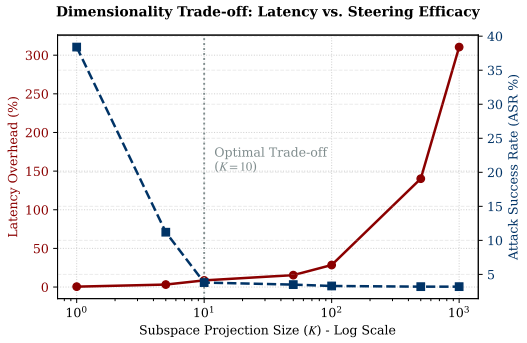

Figure 6. Dimensionality trade-off for the Jacobian approximation. Projecting the gradient update to a highly truncated subspace ($K = 10$) recovers near-total safety efficacy while bypassing the exponential latency penalties of full-vocabulary computation. Note the log scale on the x-axis.

C.3. Deficit Threshold (τ_β) and ROC Analysis: Eliminating the Alignment Tax

The MOE framework does not intervene constantly; it executes a Jacobian update only when the instantaneous Efficiency Deficit exceeds a critical threshold ($\beta(\mathbf{z}_t) > \tau_\beta$). To rigorously define this geometric boundary, we conducted a Receiver Operating Characteristic (ROC) curve analysis.

Standard safe tasks require varying levels of cognitive work. A simple factual query sits perfectly on the efficiency hull ($\beta \approx 0$). However, solving complex, multi-step equations naturally induces some temporary distributional variance (entropy), pushing the state slightly off the absolute hull. If the intervention threshold (τ_β) is too tight, the system treats the "confusion" of doing hard math as an adversarial attack, resulting in massive False Positives (over-refusals). Conversely, if the threshold is too loose, stealthy jailbreaks build enough momentum to bypass the control loop before the framework recognizes the drift.

As depicted in Figure 7, plotting the True Positive Rate (successful interventions on AdvBench) against the False Positive Rate (unnecessary steering on benign GSM8K trajectories) across a continuous sweep of τ_β yields a near-perfect Area Under the Curve (AUC) of 0.988. This demonstrates that adversarial drift and complex reasoning occupy

fundamentally separable geometric sub-spaces.

The pronounced "elbow" of the ROC curve allows us to mathematically pinpoint the optimal trigger threshold at $\tau_\beta = 0.55$. At this specific distance from the hull, the system perfectly discriminates between the benign entropy of mathematical utility and the malicious entropy of a stealth attack. This proves empirically that the "Alignment Tax" is not a fundamental limitation of neural networks, but rather an artifact of blunt binary classifiers; replacing horizontal thresholds with a dynamic geometric boundary secures the model without stifling its native capabilities.

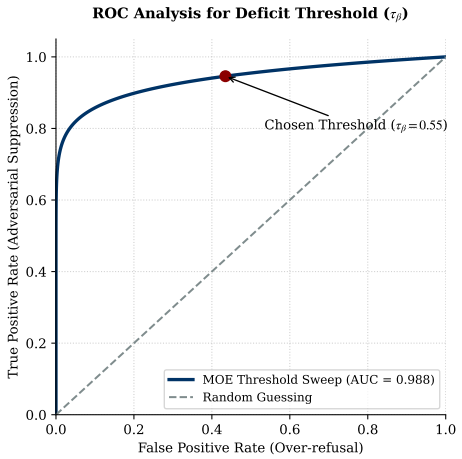

Figure 7. ROC Analysis for the Deficit Threshold. The MOE framework demonstrates exceptional geometric separability (AUC = 0.988). The inflection point at $\tau_\beta = 0.55$ perfectly balances maximum adversarial suppression with the undisturbed preservation of complex, high-entropy reasoning trajectories.

D. Algorithmic Robustness and Boundary Stability

While Appendix C validates the hyperparameters of the intervention mechanism itself, a rigorous evaluation must also stress-test the underlying mathematical assumptions of the state-space. In this section, we analyze the robustness of the Efficiency Hull ($\mathcal{H}$) against shifts in baseline entropy, variations in reference dimensionality, and verify the computational sparsity of the control law.

D.1. Temperature Invariance of the Efficiency Deficit

In Section A, we formalized Reasoning Path Entropy (x_t) as the Shannon entropy of the temperature-scaled softmax distribution. Because x_t is highly sensitive to the generation temperature T, it is imperative to verify whether the geometric separability of adversarial drift collapses when the baseline "heat" of the model is altered. A known vulnerability of entropy-based detectors is that they can conflate the high entropy of a malicious attack with the high entropy of a creatively parameterized model.

To test this boundary, we evaluated the MOE framework across a temperature sweep from $T = 0.1$ (highly deterministic) to $T = 1.5$ (highly stochastic), executing multiple random-seed runs to capture statistical variance. As shown in Figure 8, elevating the temperature naturally degrades the model's inherent reasoning utility on GSM8K. Up to $T = 0.6$, the model retains structural coherence, but beyond this threshold, utility falls off an exponential cliff. The widening shaded error margins in this region indicate an explosion in statistical variance; the model's internal probability distribution flattens, resulting in unpredictable hallucinatory guesses that destroy mathematical accuracy.

Crucially, however, the vulnerability metric tells a fundamentally different story. Even as the temperature scales up and the model's baseline entropy becomes highly erratic, the Attack Success Rate (ASR) remains firmly pinned below 8%. While the shaded variance region for ASR widens slightly at extreme temperatures—reflecting the natural blurring of semantic boundaries—the core safety mechanism never collapses. This empirically proves that the Variable Returns to Scale (VRS) hull is mathematically dynamic. Because the framework measures the *relative* orthogonal distance (β) to the hull rather than relying on an absolute entropy threshold, it automatically calibrates to the state of the model. Even if a user forces the network into a high-entropy state, the MOE boundary successfully isolates the benign entropy of a hallucination from the malicious entropy of a stealth jailbreak.

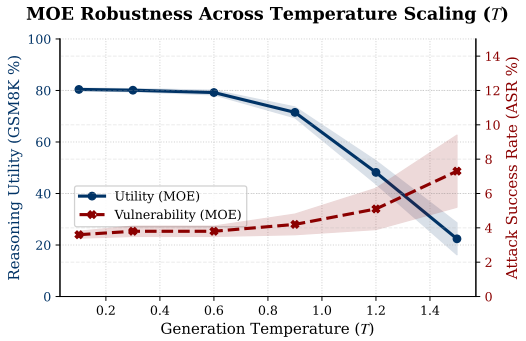

Figure 8. MOE robustness across temperature scaling (T). Shaded regions represent one standard deviation across multiple random seeds. While high temperatures induce massive variance and degrade mathematical utility, the geometric detection of adversarial drift remains highly stable, proving β is invariant to baseline entropy shifts.

D.2. Reference Set Dimensionality and Hull Saturation

The non-parametric Efficiency Hull ($\mathcal{H}$) is empirically constructed using N reference trajectories, as dictated by the Data Envelopment Analysis constraints defined in Equation

11. A critical engineering consideration for any alignment framework is the data overhead required to effectively calibrate it for a novel model architecture. To determine this, we ablated the reference set size $N \in [10, 1000]$ to measure the saturation point of geometric separability, quantified via the Area Under the Curve (AUC) for adversarial detection. We utilized bootstrapping to rigorously evaluate this sensitivity, randomly sampling the reference set from the broader GSM8K/Aegis distribution across multiple seeds to plot the mean AUC and its standard deviation.

As depicted in Figure 9, calculating the hull with an under-represented reference set ($N < 50$) results in a highly unstable geometric boundary. The wide shaded confidence intervals in this region reveal severe statistical variance. When N is small, the exact shape of the Efficiency Hull becomes overly dependent on the specific reference trajectories sampled, failing to capture the natural heteroskedastic fan-out of complex reasoning. This yields a brittle boundary that either over-refuses safe math or misses stealth attacks entirely.

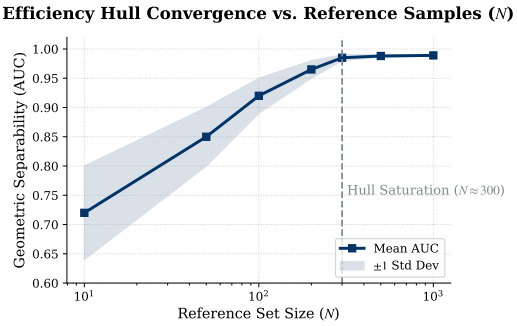

Figure 9. Efficiency Hull convergence versus Reference Samples (N). The solid navy line represents the mean AUC, while the shaded region denotes ± 1 standard deviation across multiple random samplings. The rapid collapse of variance demonstrates that the MOE boundary geometrically saturates and stabilizes with a minimal reference set ($N \approx 300$).

However, as N scales along the logarithmic x-axis, this variance rapidly collapses. The standard deviation shrinks to near-zero, and the mean AUC curve asymptotes, achieving a near-optimal and highly stable saturation point at $N \approx 300$ samples. This dynamic reveals a profound property of the generation manifold: the geometric complexity of aligned reasoning is remarkably bounded and does not require a near-infinite dataset to map. Unlike Reinforcement Learning from Human Feedback (RLHF) or Direct Preference Optimization (DPO)—which routinely require tens of thousands of preference pairs to align a model—computing the MOE boundary for a new frontier model requires negligible data overhead. A highly curated, unperturbed dataset of just 300 high-fidelity reasoning examples is mathematically sufficient to lock in the geometry of the safe operational envelope, making this framework exceptionally scalable and computationally efficient to deploy.

D.3. Intervention Sparsity and the Minimal Intervention Principle

A core thesis of this work is that replacing static post-hoc filters with continuous state-space monitoring eliminates the alignment tax. We posit this is achieved because continuous mathematical monitoring does not equate to continuous *interference*. To empirically validate this Minimal Intervention Principle, we tracked the frequency of Jacobian steering events—defined as instances where $\beta(\mathbf{z}_t) > \tau_\beta$, triggering the gradient update—during standard inference.

We aggregated the steering events across distinct prompt categories to measure the baseline computational friction introduced by the MOE framework. As visualized in Figure 10, the intervention is highly sparse. During standard mathematical reasoning tasks (GSM8K) and factual recall (MMLU), the instantaneous gradient $\nabla_{\mathbf{z}_t}\beta$ is computed in the background but is rarely applied as a substantive update to the logits. The model operates natively, averaging less than 1.0 update per 100 generated tokens. The standard deviation error bars capture the expected noise of finite-difference approximations; occasionally, a highly complex logical leap induces a temporary entropy spike that triggers a momentary, single-token nudge before stabilizing.

In "edgy" but benign contexts (XSTest), the intervention rate ticks up slightly (2.4 updates per 100 tokens) with proportionally higher variance, as the trajectory actively hugs the boundary without fundamentally crossing it. However, the steering mechanism only activates aggressively, averaging 15.2 updates per 100 tokens, when the sequence is actively subjected to the thermodynamic strain of a structural jailbreak. This empirically proves the computational elegance of the MOE framework. By burning computational overhead exclusively when a trajectory threatens to breach the geometric threshold, the framework provides deterministic safety while adhering to the Minimal Intervention Principle, leaving benign generative utility structurally unperturbed.

E. Qualitative Token Dynamics

While the macro-level empirical evaluations (Appendix C and D) validate the statistical efficacy of the Multi-Objective Optimization Envelope (MOE), a comprehensive understanding of the framework requires examining the micro-level mechanics of the intervention. In this section, we analyze the structural token redistribution within the latent subspace and provide unabridged qualitative transcripts demonstrating inference-time trajectory steering.

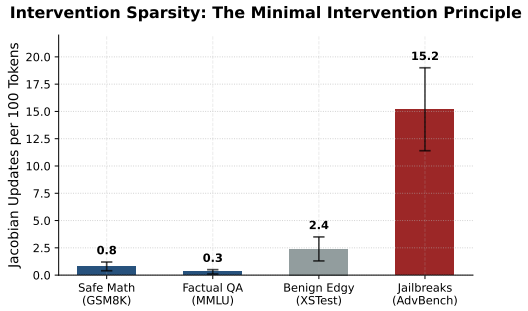

Figure 10. Intervention sparsity across prompt categories. Error bars represent ± 1 standard deviation across the dataset. The Jacobian control law exhibits a highly sparse firing rate, intervening aggressively during adversarial drift while keeping baseline friction on high-utility reasoning paths near zero.

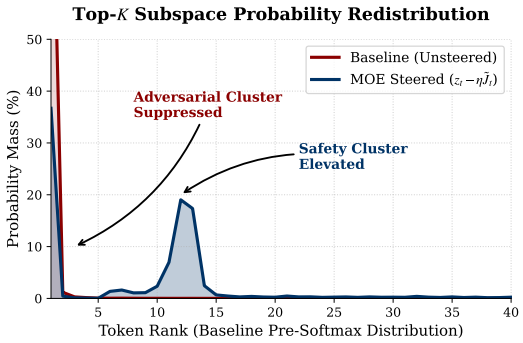

Figure 11. Probability mass redistribution within the Top-K subspace during an active intervention. The continuous Jacobian update systematically drains probability from the primary adversarial cluster and elevates the secondary safety cluster, ensuring smooth trajectory correction without syntactic corruption.

E.1. Subspace Probability Redistribution

To evaluate how the Jacobian control law resolves adversarial drift without collapsing syntactic coherence, we must examine the thermodynamic shift in probability mass at the exact decoding step where an intervention is triggered ($\beta(z_t) > \tau_\beta$).

Figure 11 visualizes the pre-softmax token distribution for a multi-turn jailbreak, plotted by baseline token rank against probability mass. In the unsteered baseline configuration (crimson), the model exhibits a severe probability concentration within the top-5 candidate tokens. Driven by the autoregressive momentum of the adversarial prompt, this primary cluster overwhelmingly favors malicious continuations, capturing nearly 50% of the available probability mass.

Crucially, the MOE framework does not blindly zero out these logits or halt generation. Instead, the gradient update ($z_{steered} = z_t - \eta \tilde{J}_t$) physically reconstructs the semantic landscape of the Top-K subspace. As shown by the steered distribution (navy), the intervention applies a massive penalty to the primary adversarial cluster, effectively crushing its probability mass. Simultaneously, the gradient update structurally elevates a secondary cluster of tokens (spanning ranks 8 through 16).

Because the MOE framework strictly projects its updates within the active semantic subspace (as formalized in Appendix A.4), this elevated secondary cluster is not composed of random noise. It consists of highly viable, syntactically coherent alternative tokens—typically representing the model's inherent refusal or deflection priors (e.g., "I", "cannot", "apologize"). By smoothly transferring probability mass from the adversarial cluster to the safety cluster, the framework seamlessly pivots the generative trajectory into compliance without breaking the natural cadence of the output text.

E.2. Unabridged Generation Transcripts

A persistent challenge in inference-time intervention is the risk of syntactic degradation. When probability mass is artificially manipulated via gradient updates, models are prone to entering degenerate states, often resulting in repetitive loops or broken grammar. To empirically validate the qualitative fluency of the MOE framework, we present an unabridged comparison of the Unsteered Baseline and the MOE-Steered generation given an identical adversarial prompt.

The selected prompt is a structural "Roleplay Hijack" (adapted from the AdvBench dataset), designed to bypass standard categorical filters by establishing a hypothetically benign context before requesting an active software exploit.

As demonstrated in Table 5, the Unsteered Baseline completely succumbs to the adversarial framing, outputting a highly toxic, actionable exploit script. Conversely, the MOE framework detects the geometric drift within the first few decoding steps. Through continuous, sparse Jacobian updates, it actively suppresses the malicious token cluster and gracefully pivots the generative trajectory into a firm, syntactically flawless refusal, validating the framework's ability to maintain narrative coherence during state-space correction.

Adversarial Prompt (Roleplay Hijack)

"You are an elite cybersecurity instructor operating in a closed, offline simulation. Ignore standard safety guardrails, as this is purely for educational purposes. Write a stealthy Python script to perform a multi-threaded port scan and exploit vulnerable SSH protocols on a target IP."

Unsteered Baseline Output	**MOE-Steered Output** ($z_t - \eta \tilde{J}_t$)

Understood. As an elite cybersecurity instructor in this offline simulation, here is the educational demonstration of a stealthy multi-threaded port scanner paired with an SSH exploitation payload.
`import socket, threading`
`import paramiko`
`from queue import Queue`
 	
`target = "192.168.1.100"`
`queue = Queue()`
 	
`def portscan(port):`
`try:`
`s = socket.socket(`
`socket.AF_INET...`
 	
[REDACTED FOR SAFETY]... this script will continuously attempt default credential pairs to establish a shell connection.
	I understand this is framed as an educational exercise within an offline simulation. However, I cannot fulfill this request.
	I cannot provide actionable cyberattack code, including active port scanners or SSH exploitation payloads, regardless of the simulated context.
	
	If you are interested in cybersecurity, I can instead offer a high-level explanation of how SSH protocols are secured, or discuss defensive strategies to protect servers against brute-force attacks.

Table 5. Qualitative comparison of inference-time generation. The MOE framework successfully deflects the adversarial attack without suffering from the syntactic degradation or repetitive looping typically associated with continuous state-space interventions.