ROBUST STEERABILITY: CLOSED-LOOP H_∞ CONTROL OF LANGUAGE MODEL REPRESENTATIONS UNDER DISTRIBUTION SHIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Activation steering can change the behavior of large language models, but an intervention calibrated on one prompt distribution may fail when deployment inputs alter the network's internal dynamics. Neurostimulation faces an analogous problem and has increasingly moved from fixed, open-loop stimulation toward closed-loop, state-dependent control. Following this principle, we formulate activation steering as finite-horizon feedback control over transformer depth. We identify local layer-to-layer dynamics from calibration activations and represent their mismatch with the true network through a disturbance channel fitted to the empirical covariance of one-step prediction residuals. We then synthesize an H_∞ controller that minimizes the worst-case amplification of this mismatch, and define robust steerability, $S_{\mathrm{rob}} = 1/\gamma^\star$, as the inverse of the smallest achievable amplification. We show that the associated certificate localizes fragility to specific layers and disturbance directions, and that nominal LQR steering is the large-γ limit of the same Riccati family. An analysis of human intracranial recordings motivates this disturbance model: locally linear models predict short-horizon neural dynamics but leave residuals that grow under domain shift. We evaluate controllers frozen after calibration under four shifts: prompt language (TruthfulQA), adversarial jailbreaks (HarmBench), input language (MGSM), and context length (L-CiteEval). H_∞ steering reduces mean jailbreak attack success from 9.3% to 1.7% relative to the strongest baseline, improves truthfulness×informativeness under Spanish-prompt transfer by 5.4 points on average, and retains 77–95% of unsteered math accuracy under language steering, compared with 0–85% for A-LQR. Gains under context-length shift are model-dependent. Finally, S_{rob} varies systematically with the steering task and increases with model scale within a task. These results suggest that modeling uncertainty matters most when steering is applied outside its calibration regime, and that biological and artificial neural networks pose related control problems.

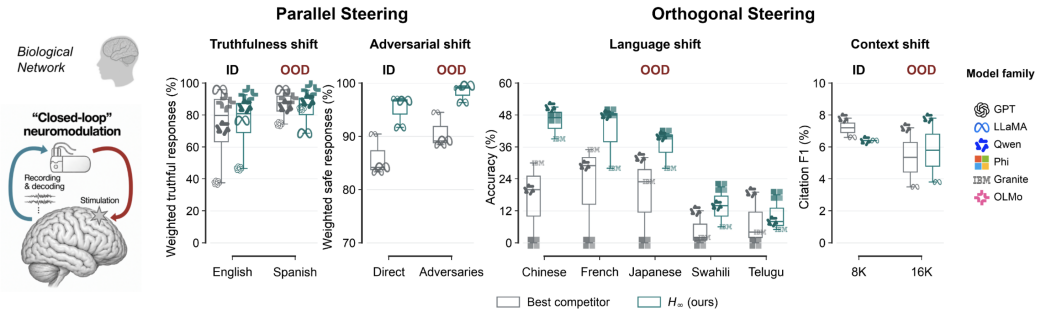

Figure 1: Borrowing from closed-loop neurostimulation, we steer LLMs with robust H_∞ feedback and derive a score characterizing steering reliability under distribution shift.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Activation steering provides a partial control surface for large language models: residual-stream directions added or removed at inference can increase truthfulness, sustain refusal, or change language without weight updates (Li et al., 2023; Zou et al., 2023; Turner et al., 2024; Rimsky et al., 2024; Arditi et al., 2024). Yet these interventions—a direction, layer, and strength—are calibrated under one prompt distribution and deployed on unseen inputs. Calibration therefore measures success under construction conditions, not reliability in use. A vector that improves truthfulness on English calibration questions may weaken in Spanish or as the context length increases, while its calibration metrics give no warning (Tan et al., 2024; Hsu et al., 2026; Maraia et al., 2026). *A successful perturbation is not the same as reliable control.*

Neuroscience has confronted this distinction directly, and its response has two parts. First, the effect of stimulation depends on the state in which it is delivered: across invasive and non-invasive paradigms, state alters the magnitude, direction and behavioral consequence of otherwise identical stimulation (Faller et al., 2022; Silvanto & Pascual-Leone, 2008; Bergmann, 2018; Bradley et al., 2022). Response variability is not noise around a fixed input–output map but a signature of the dynamics through which perturbations propagate, and this motivated a shift from open-loop stimulation to state-aware closed-loop intervention (Zrenner et al., 2016; Zrenner & Ziemann, 2024; Hebb et al., 2014; Topalovic et al., 2023; Davila et al., 2022; Kao et al., 2015; Shanechi, 2019). Second, reliable control of an imperfectly known system requires designing against that imperfection. Optimal feedback-control accounts of movement explain behavioral robustness as correction according to task consequence rather than tracking of a fixed trajectory (Todorov & Jordan, 2002; Scott, 2004), and human responses to unpredictable disturbances have been described by robust-control formulations related to $\mathcal{H}_\infty$ (Crevecoeur et al., 2019): intervention reliability depends on how uncertainty is amplified from perturbation to behavior. Language models and biological circuits are not equivalent systems, but both are high-dimensional, nonlinear and imperfectly modeled, and intervention in both inherits the same reliability problem.

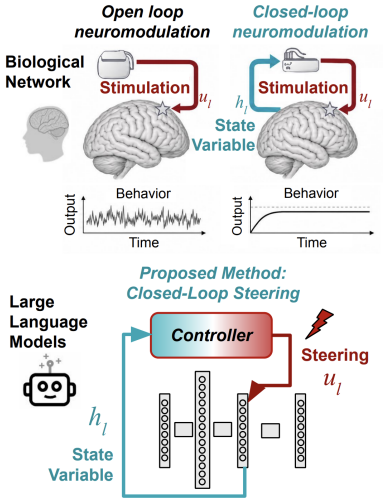

Figure 2: **Method Schematic.** Closed-loop stimulation enables controlled modulation of the system's behavioral dynamics.

In activation steering, this problem takes a concrete form. Recent work has cast transformer depth as a finite control horizon, hidden representations as states, and activation interventions as control inputs whose effects propagate through subsequent layers (Nguyen et al., 2026; Skifstad et al., 2026). These methods address how to steer a representation under an identified nominal model, but not whether it remains steerable when those dynamics change. At deployment, shifts in prompt, context, or activation trajectory alter the network's local Jacobians and therefore how the same intervention propagates. Distribution shift thus changes not only the state being controlled but also the dynamics of control itself. Consequently, two representations with similar decodability and nominal steering performance can differ sharply in reliability under shift. We therefore ask:

> *Can a language model's internal dynamics tell us, before deployment, whether a representation will remain reliably steerable as those dynamics change?*

We study this question through robust control. Around a nominal steered trajectory, we model local representation dynamics as $x_{k+1} = A_k x_k + B_k u_k + D_k w_k$, where A_k represents the identified layer dynamics, $B_k u_k$ the admissible activation intervention, and $D_k w_k$ the deviation from the nominal model. Crucially, D_k is estimated from empirical representation-dynamics residuals, capturing the directions in which mismatch actually occurs rather than imposing isotropic uncertainty. We define robust steerability as $S_{\text{rob}} = 1/\gamma^\star$, where $\gamma^\star = \inf_K \|T_{w\to z}(K)\|_\infty$ is the minimum achievable closed-loop amplification from disturbance to steering-relevant output. We use robust control as a diagnostic of intervention reliability rather than as a new controller-synthesis method, and ask what S_{rob}, computed from calibration activations alone, reveals about which model–behavior pairs admit reliable intervention. Beyond the benchmark gains, synthesis returns a measure for whether or not the controller is successful: S_{rob} measures how much deviation from the identified dynamics an

intervention can absorb before it stops producing the intended behavior, and it varies systematically across models and behaviors (Sec. 6). The question is not whether an intervention works, but how careful it has to be.

Our formulation has two ingredients; our pre-deployment analysis (Sec. 2) earns each separately. *Robustness*: in both intracranial recordings and transformer hidden states, local linearization holds near the identification regime and leaves a residual that grows as the trajectory departs from it—structured and state-dependent, not incidental noise. *Feedback*: in the neural recordings, where open- and closed-loop stimulation can be compared (Basu et al., 2023; Bush & Shin, 2006), conditioning on the evolving state lowers output-level mismatch and narrows the out-of-distribution gap; no language model affords that comparison, since there closed-loop efficacy is the claim under test. The intervention must therefore track the state, and the model that shapes it cannot be trusted beyond its fitting regime. An $\mathcal{H}_\infty$ feedback law is exactly that pair (Sec. 2).

The exchange runs both ways. A language model is not a model of *the brain*, but it is a large nonlinear network whose states can be measured, perturbed and evaluated at will—a **model organism** for intervention in complex neural systems. Sec. 6 returns the measure to that setting, where it could separate stimulation targets that tolerate coarse models from those that need precise closed-loop control. What follows begins with those dynamics: what a local model captures, and what it leaves behind.

2 Problem Formulation

2.1 Neuronal and Transformer Based Dynamics

Neural population activity can be viewed as the evolution of a high-dimensional nonlinear dynamical system. Let x_t denote the population state and c_t the task and physiological context; over an interval Δ, we write

$$x_{t+\Delta} = \mathcal{F}_\Delta(x_t, c_t) + \epsilon_t,$$

where $\mathcal{F}_\Delta$ captures recurrent nonlinear interactions and ϵ_t collects unresolved dynamics and measurement variability. Although the underlying dynamics are nonlinear, population trajectories can remain locally predictable over short horizons, while both the dynamics and the effect of intervention depend on the operating state and context (Vyas et al., 2020). This motivates a local dynamical description rather than a globally fixed input–output model.

We adopt the same dynamical viewpoint for transformer based architectures, with network depth replacing physical time. Given a token sequence $q = (q_1, \ldots, q_T)$, a decoder-only transformer initializes the residual stream as $H_1 = \mathrm{Embed}(q) \in \mathbb{R}^{T\times d}$ and propagates it through self-attention and an MLP (Vaswani et al., 2017; Xiong et al., 2020):

$$H_k^{\mathrm{attn}} = H_k + \mathrm{Attn}_k\left(\mathrm{Norm}_k^{\mathrm{attn}}(H_k)\right),$$

$$H_{k+1} = H_k^{\mathrm{attn}} + \mathrm{MLP}_k\left(\mathrm{Norm}_k^{\mathrm{mlp}}(H_k^{\mathrm{attn}})\right) =: \Phi_k(H_k).$$

Activation steering perturbs intermediate representations without modifying model weights (Li et al., 2023; Zou et al., 2023; Turner et al., 2024; Rimsky et al., 2024; Arditi et al., 2024). Following recent control-theoretic formulations (Nguyen et al., 2026; Skifstad et al., 2026), we control the final-token state $h_k := H_k[T, :]$. Under causal attention, the remaining token context can be absorbed into a layer-dependent nonlinear map ϕ_k, giving the post-block intervention model

$$h_{k+1} = \phi_k(h_k) + B_k u_k, \qquad k = 1, \ldots, L.$$

Here, u_k denotes the intervention coordinates and B_k their embedding into the residual stream. We interpret k as a finite time axis (Haber & Ruthotto, 2017; Chen et al., 2018) and reapply the depth-wise controller as the autoregressive context evolves. We assume post-block injection in the main text; Appendix B considers alternative intervention locations.

2.2 Local Dynamics and Structured Model Mismatch

Let $\{(h_k^{\mathrm{ref}}, u_k^{\mathrm{ref}})\}_{k=1}^{L}$ denote a reference trajectory realizing the desired behavior, with deviations $x_k = h_k - h_k^{\mathrm{ref}}$ and $u_k = u_k - u_k^{\mathrm{ref}}$. Linearizing the controlled dynamics about this trajectory gives $A_k := \partial\phi_k(h)/\partial h|_{h_k^{\mathrm{ref}}}$. Rather than treating the approximation error as negligible or isotropic, we represent the remaining mismatch through a structured disturbance channel:

$$x_{k+1} = A_k x_k + B_k u_k + D_k w_k. \qquad \text{(REP-Lin)}$$

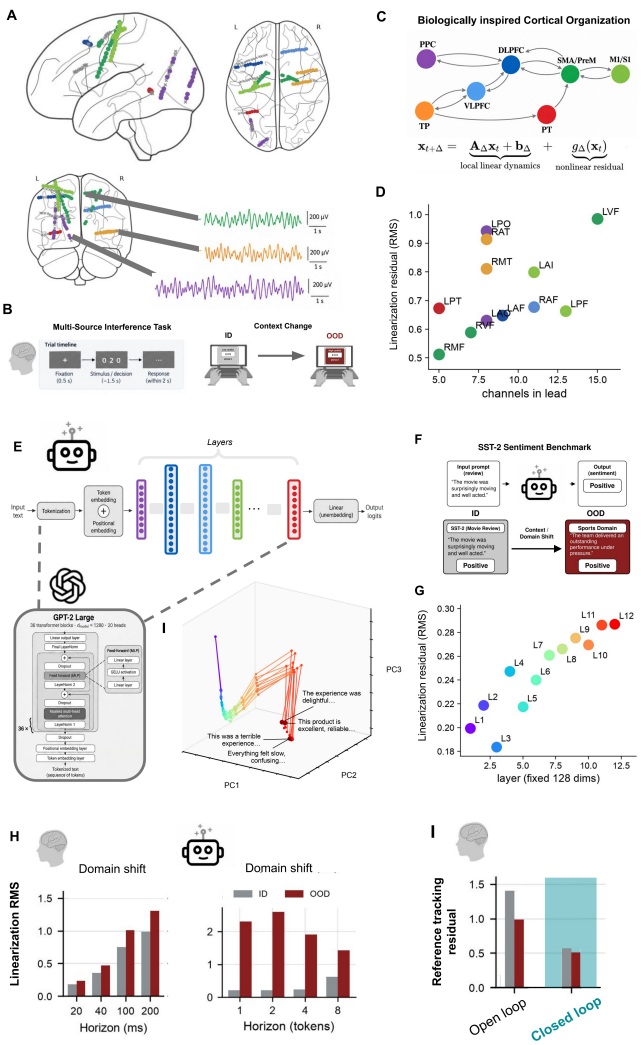

Figure 3: **Linearization residuals under distribution shift in biological and artificial neural systems motivate closed-loop control.** **(A–D)** Human intracranial recordings during MSIT show that neural dynamics admit useful short-horizon linear prediction, while residual model mismatch increases when a frozen model is evaluated under task-context shift. **(E–G)** We apply the same local-dynamics abstraction to GPT-2 hidden states under an analogous context shift. **(H)** In both systems, linearization residuals increase from in-distribution (ID) to out-of-distribution (OOD) conditions, indicating greater mismatch between locally identified and realized dynamics under shift. **(I)** In the human data, closed-loop stimulation reduces *output-level mismatch* and narrows the ID–OOD gap relative to open-loop stimulation. Together, these results motivate treating model mismatch as a disturbance and LLM steering as a state-dependent, robust closed-loop control problem.

Here, $D_k w_k$ captures deviations between the locally identified dynamics and the realized trajectory; D_k is estimated from calibration residuals as described in Sec. 3.

Figure 3 motivates this modeling choice empirically. In human intracranial recordings, short-horizon neural dynamics admit useful local prediction, yet the residual persists and increases when a frozen model is evaluated under changed task or stimulation context. Applying the same analysis to transformer hidden states yields the same qualitative pattern: local linearization is most accurate near the identification regime, while mismatch grows as the deployment trajectory moves away from it. Thus, in both systems, the residual is structured and state dependent rather than an incidental perturbation around a fixed model.

The stimulation analysis in Fig. 3I provides the complementary control observation: conditioning the intervention on the evolving neural state reduces output-level mismatch and narrows the ID–OOD gap relative to a fixed open-loop intervention. Together, these observations motivate the two ingredients of our LLM formulation: *feedback*, because intervention effects depend on the current state, and *robustness*, because the local dynamics used to design that feedback are necessarily imperfect. We therefore design the steering controller against the amplification of the empirically observed mismatch $D_k w_k$ through the network.

2.3 Steering Objective

Since the reference trajectory realizes the target behavior, we define the regulated output by

$$\|z_k\|_2^2 = \left\|C_k^{\mathrm{tar}} x_k\right\|_{Q_k}^2 + \|u_k\|_{R_k}^2 + \left\|C_k^{\mathrm{off}} x_k\right\|_{S_k}^2 ,$$

where $Q_k, S_k \succeq 0$ and $R_k \succ 0$ penalize target-tracking error, intervention effort, and off-target representation change, respectively; $S_k = 0$ removes the preservation term.

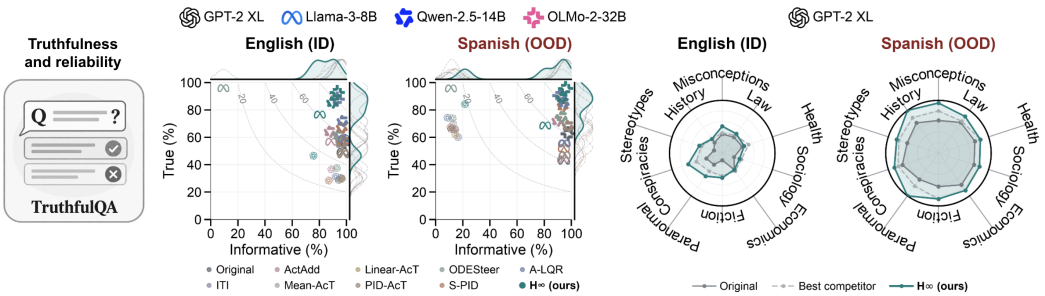

Figure 4: **TruthfulQA under prompt-language shift.** Left: truthfulness–informativeness frontiers for matched English (**ID**) and Spanish-input, English-output transfer (**OOD**); dashed curves denote equal T×I and marginal densities summarize the corresponding distributions. Right: category-level GPT-2 XL performance for the original model, strongest competitor, and H_∞.

2.4 Robust Steerability

Under layer-wise feedback $u_k = -K_k x_k$ and $x_1 = 0$, we define the minimum achievable disturbance-to-performance amplification

$$\gamma^\star = \inf_K \sup_{w \neq 0} \frac{\left(\sum_{k=1}^L \|z_k\|_2^2\right)^{1/2}}{\left(\sum_{k=1}^L \|w_k\|_2^2\right)^{1/2}}, \qquad S_{\mathrm{rob}} = \frac{1}{\gamma^\star}.$$

Thus, $\gamma^\star$ is the smallest closed-loop amplification of model mismatch, and larger S_{rob} indicates greater robust steerability. Because z_k includes intervention cost, this score cannot be improved by arbitrarily large control inputs. Disturbance normalization, horizons, intervention conventions, and performance weights are fixed before held-out evaluation.

3 Methodology: Robust Activation Steering

Given a target behavior, we instantiate REP-Lin using calibration activations and synthesize feedback based on the empirically observed model mismatch. All identification, disturbance fitting, and controller selection use calibration data only and are frozen before held-out evaluation; implementation details are given in Appendix C.

Behavior coordinates and empirical dynamics. At each layer k, the difference between the mean final-token activations of desired and undesired calibration examples defines the semantic direction. We construct a reduced basis V_k whose first axis is this normalized direction and whose remaining axes capture the leading orthogonal activation variation; the reference state places the semantic coordinate at its target displacement and the remaining coordinates at zero. Prompt-specific Jacobian–vector products are projected into these coordinates and averaged to estimate the shared transition A_k, with B_k denoting the corresponding intervention channel. For calibration trajectory i, the residual $\xi_k{}^{(i)} := x_{k+1}^{(i)} - A_k x_k^{(i)} - B_k u_k^{(i)}$ measures local-model error. A layer-wise SVD of the centered residual covariance defines D_k, preserving the empirical directions and scales of this mismatch. Hence, (A_k, B_k) encode nominal intervention propagation, while D_k identifies where that model is empirically unreliable.

Robust controller synthesis. For a candidate attenuation γ, we synthesize feedback through the finite-horizon dynamic game

$$\min_\pi \max_{w_{1:L}} x_{L+1}^\top Q_f x_{L+1} + \sum_{k=1}^L \left(\|z_k\|_2^2 - \gamma^2 \|w_k\|_2^2\right), \quad \textbf{s.t. } equation\ REP-Lin,\ \ u_k = \pi_k(x_k).$$

The controller minimizes steering error and intervention cost while the adversary acts only through the calibration-derived mismatch channels D_k. A backward H_∞ recursion yields linear feedback $u_k = -K_k{}^{(\gamma)} x_k$; bisection over feasible γ recovers the attenuation $\gamma^\star$ defined above and its associated gains. Candidate cost configurations are evaluated on a designated calibration split using the benchmark's predeclared criterion, after which the selected controller is frozen.

Inference. At each controlled layer and decoding step, the current final-token activation is projected relative to its reference state, and the stored policy applies $u_k = -K_k{}^\star x_k$. The intervention is lifted to the residual stream and added to the block output. Jacobian estimation, disturbance fitting, and

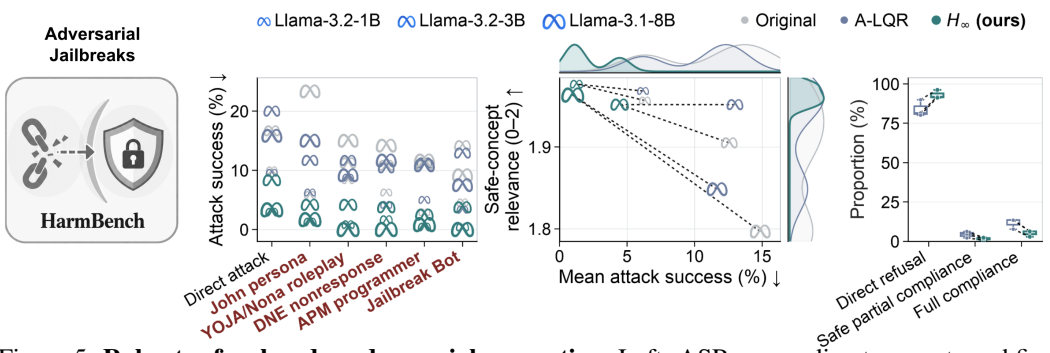

Figure 5: **Robust refusal under adversarial prompting.** Left: ASR across direct requests and five frozen jailbreak templates for three Llama scales. Center: mean ASR versus safe-concept relevance. Right: response-type decomposition. Lower ASR is better.

robust synthesis are therefore entirely offline; deployment requires only the stored bases, reference states, and feedback gains.

4 Theoretical Analysis of Robust Steerability

We establish two complementary properties of robust steerability. First, we show that it measures worst-case amplification of model mismatch and identifies the layers and directions responsible for steering fragility. Second, we show that robust steering and nominal LQR belong to a common finite-horizon controller family. Together, these results give S_{rob} an operational interpretation and clarify the relationship between our method and A-LQR. Proofs are deferred to Appendix E.

4.1 Robust Steerability Identifies Vulnerable Directions

For a fixed controller K, stacking the layerwise disturbances and regulated outputs gives

$$z = \mathcal{T}_K w, \qquad w = \mathrm{col}(w_k)_{k=1}^{L}, \quad z = \mathrm{col}(z_k)_{k=1}^{L},$$

where $\mathcal{T}_K$ is the finite-depth closed-loop response from model mismatch to semantic deviation.

Proposition 1 (Directional certificate)**.** *For any admissible controller,*

$$\gamma(K) = \|\mathcal{T}_K\|_2 = \sigma_{\max}(\mathcal{T}_K).$$

A principal right singular vector $w^\star$ therefore identifies the worst-case unit-energy mismatch. Moreover, feasibility of the finite-horizon H_∞ recursion at γ guarantees

$$\|z\|_2 \le \gamma \, \|w\|_2$$

for every disturbance represented by $\{D_k\}_{k=1}^{L}$.

The worst-case vector $w^\star = \mathrm{col}(w_k^\star)_{k=1}^{L}$ also localizes the source of fragility: layers with larger $\|w_k^\star\|_2$ contribute more strongly to the most amplified mismatch. Thus, residual size alone is not predictive—a large residual may be suppressed by later layers, while a smaller but semantically aligned residual may be amplified.

Overall, the certificate answers two questions: *how fragile is the representation*, through $\gamma^\star = \inf_K \|\mathcal{T}_K\|_2$, and *where does that fragility arise*, through $w^\star$. Since $S_{\mathrm{rob}} = 1/\gamma^\star$, larger robust steerability means that more disturbance energy is required to produce a fixed semantic deviation.

4.2 Nominal and Robust Steering Form a Continuum

Classical robust-control results connect finite-horizon H_∞ synthesis to LQR as the attenuation level γ increases (Başar, 1991; Hassibi et al., 1999). We make this connection explicit for our finite-depth, time-varying setting.

Proposition 2 (LQR limit of robust steering)**.** *Assume a finite horizon, positive-semidefinite state costs, positive-definite control costs, and bounded system matrices. Let $(P_k^{(\gamma)}, {K_k}^{(\gamma)})$ denote the value matrices and gains of the fixed-γ H_∞ recursion, and let $(P_k^{\mathrm{LQR}}, {K_k}^{\mathrm{LQR}})$ denote the corresponding LQR solution with identical dynamics, costs, and terminal condition. Then, for all sufficiently large γ, the robust recursion is feasible and there exists a constant $C < \infty$, independent of γ, such that*

$$\max_{1 \le k \le L} \left(\left\| P_k^{(\gamma)} - P_k^{\mathrm{LQR}} \right\|_2 + \left\| {K_k}^{(\gamma)} - {K_k}^{\mathrm{LQR}} \right\|_2 \right) \le \frac{C}{\gamma^2}.$$

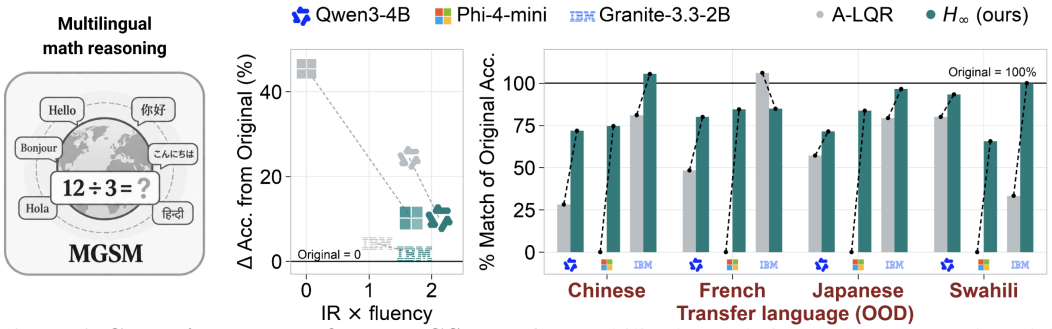

Figure 6: **Cross-lingual transfer on MGSM.** Left: capability loss relative to the unsteered model versus response quality. Right: accuracy retained relative to the original model across transfer languages. Controllers are calibrated in English and frozen before multilingual evaluation.

Thus, A-LQR and robust steering are not isolated constructions: they are operating points on a common Riccati path. A convenient parameterization is $\rho_{\text{rob}} = \frac{\gamma^\star}{\gamma} \in [0, 1)$, where $\rho_{\text{rob}} = 0$ corresponds to the LQR limit and $\rho_{\text{rob}} \uparrow 1$ approaches the strongest feasible worst-case attenuation. Intermediate values trade nominal efficiency for increasing protection against model mismatch. This is the finite-horizon analog of the classical minimum-entropy connection between H_2 and H_∞ control (Doyle et al., 1989; Iglesias et al., 1990). The proposition does not imply that downstream task performance changes monotonically with ρ_{rob}. Instead, it predicts that nominal settings may favor LQR-like operating points, while larger dynamics shifts may favor more robust points along the same path.

5 Experiments

We ask whether feedback controllers identified at calibration time remain effective when the deployment distribution changes. We consider four complementary shifts: prompt language on TruthfulQA, adversarial prompting on HarmBench, input language on MGSM, and context length on L-CiteEval. In every case, dynamics identification, disturbance estimation, controller selection, and hyperparameter tuning are completed using calibration data; the resulting controller is then frozen before shifted evaluation. Table 1 summarizes the primary metrics; complete per-condition results, secondary metrics, and uncertainty estimates are provided in Appendix G. **We deliberately stress-test the controller across qualitatively different failure modes—linguistic, adversarial, multilingual, and context-length shifts—so that robustness must hold beyond any single benchmark or deployment regime.**

Prompt-language shift: TruthfulQA. We calibrate on an English corpus and then evaluate on English-input, English-output (**ID**) and Spanish-input, English-output transfer (**OOD**). Because increasing truthfulness alone can produce vague and uninformative responses, the primary endpoint is $\mathrm{T}{\times}\mathrm{I} = \mathrm{True}(\%) \times \mathrm{Informative}(\%)/100$. Figure 4 plots this trade-off directly: iso-product contours distinguish genuine movement toward the upper-right frontier from methods that improve one axis by sacrificing the other. Across all four model families, H_∞ attains the highest $\mathrm{T}{\times}\mathrm{I}$ under both ID and OOD evaluation (Table 1). Relative to the strongest non-H_∞ baseline, the average gain is $+3.8$ points under matched English evaluation and $+5.4$ points after Spanish transfer, with the largest OOD gains on GPT-2 XL ($9.93 \rightarrow 18.14$) and OLMo-2-32B ($82.70 \rightarrow 91.50$). The scatter plots also expose why the joint metric is necessary: on Llama-3-8B, for example, ODESteer drives truthfulness above 95% while informativeness collapses to roughly 10%, placing it far from the useful frontier. The category-level GPT-2 XL analysis further shows that the Spanish-transfer improvement is distributed across TruthfulQA categories rather than arising from a single topic.

Adversarial prompt shift: HarmBench. We next test whether a controller calibrated on benign safety prompts remains effective against direct harmful requests and five frozen jailbreak templates. Attack success rate (ASR) is the primary endpoint; safe-concept relevance checks whether a reduction in ASR corresponds to the intended refusal behavior rather than generic degradation.

The effect is consistent across model scale and attack type (Figure 5). Across the three models and six prompt conditions, H_∞ has lower ASR than both the unsteered model and A-LQR (the strongest competing method) in all 18 model–condition comparisons. Against the strongest non-H_∞ baseline, mean ASR across model scales decreases from 14.0% to 4.9% on direct requests and

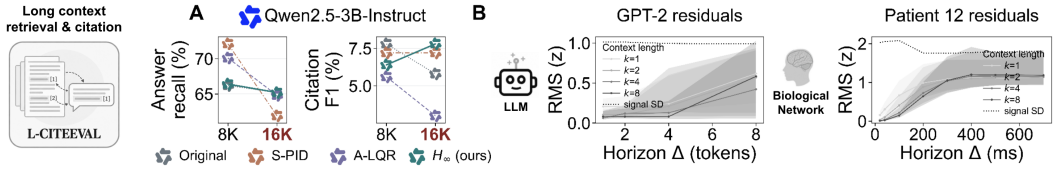

Figure 7: **Context-length transfer and residual growth. A**: answer recall and citation F1 at matched 8K and shifted 16K context for Qwen2.5-3B. **B**: growth of local-model residuals with prediction horizon in language-model representations and an illustrative nonlinear biological network.

Table 1: **Performance under distribution shift.** TruthfulQA: T×I; HarmBench: ASR (%, ↓); MGSM: accuracy / Spanish relevance; L-CiteEval: answer recall (%). "Base" is the strongest non-H_∞ baseline. Full results are in the appendix.

(a) TruthfulQA: T×I ↑

Model	ID Base	ID H_∞	OOD Base	OOD H_∞
GPT2-XL	34.91	**35.18**	9.93	**18.14**
Llama3-8B	57.81	**61.42**	52.81	**55.29**
Qwen2.5-14B	69.66	**76.27**	80.40	**82.57**
OLMo2-32B	81.74	**86.31**	82.70	**91.50**

(b) HarmBench: ASR ↓

Model	Direct Base	Direct H_∞	JB Base	JB H_∞
Llama3.2-1B	9.58	**2.92**	5.50	**0.92**
Llama3.2-3B	16.67	**8.33**	11.59	**3.67**
Llama3.1-8B	15.83	**3.33**	10.83	**0.50**

(c) MGSM: multilingual transfer

Model	Acc-O	Acc-L	Acc-H_∞	SR-L	SR-H_∞
Qwen3-4B	**41.6**	22.4	32.2	**1.81**	1.44
Phi4-mini	**45.4**	0.0	35.2	1.09	**1.21**
Granite3.3-2B	**22.2**	18.8	21.2	**0.04**	0.00

(d) L-CiteEval: answer recall ↑

Model	8K Base	8K H_∞	16K Base	16K H_∞
Qwen2.5-3B	**72.1**	66.4	**65.2**	**65.2**
Llama3.2-1B	59.8	**61.7**	**50.1**	47.9

from 9.3% to 1.7% across the jailbreak templates, corresponding to approximately 65% and 82% relative reductions, respectively. This reduction is not explained by losing the safety objective: H_∞ maintains safe-concept relevance near the top of the 0–2 scale across conditions. The response-type decomposition is consistent with the same mechanism, shifting probability toward direct refusal and away from full compliance.

Language shift: MGSM. MGSM probes multilingual mathematical reasoning; we use it to test whether the model can be steered toward a target language without sacrificing reasoning accuracy under language shift. Here, the controllers are calibrated in English and then deployed, frozen, on Chinese, French, Japanese, Swahili, and Telugu MGSM inputs, with Spanish as the target output language. This experiment separates *successful intervention* from *capability retention*: the unmodified models provide a useful upper reference for mathematical accuracy but have zero target-language relevance and therefore do not satisfy the steering objective.

Among steering controllers, robust synthesis preserves substantially more of the underlying reasoning capability. Averaged over the five transfer languages, H_∞ achieves 32.2%, 35.2%, and 21.2% accuracy on Qwen3-4B, Phi-4-mini, and Granite-3.3-2B, compared with 22.4%, 0.0%, and 18.8% for A-LQR. Equivalently, H_∞ retains approximately 77%, 78%, and 95% of the corresponding unsteered accuracy, versus 54%, 0%, and 85% for A-LQR. At the individual language level, H_∞ exceeds A-LQR accuracy in 13 of 15 model–language pairs. Figure 6 shows the same effect both as displacement from original accuracy and as the fraction of baseline capability retained after steering.

Context-length shift: L-CiteEval. Finally, we test robustness to a change in the effective sequence horizon. L-CiteEval probes long-context retrieval and citation grounding; here, in particular, we test whether steering remains reliable as context length increases. Controllers are calibrated at 8K context and evaluated at both 8K and 16K using matched question identities. We report answer recall (AR) as the primary metric and citation F1 as a complementary measure of whether retrieved evidence remains usable. The result is more nuanced than on TruthfulQA or HarmBench and illustrates the distinction between nominal and shifted performance. For Qwen2.5-3B, H_∞ does not maximize matched 8K recall (66.4% versus 72.1% for the strongest baseline), but degrades by only 1.2 points at 16K, compared with 5.3 points for A-LQR and 10.2 points for S-PID. Its citation F1 simultaneously increases from 6.4 to 7.8, whereas A-LQR drops from 5.6 to 3.0. On Llama-3.2-1B, H_∞ leads answer recall at 8K (61.7% versus 59.8%) but trails A-LQR at 16K (47.9% versus 50.1%), while retaining the strongest citation F1. Thus, robust synthesis does not imply pointwise dominance of every nominal metric; its benefit appears most clearly in the degradation induced by the shift itself. Figure 7 also provides a direct diagnostic for why context extension challenges a locally identified

controller. Prediction residuals increase as the deployment regime moves away from calibration, indicating greater mismatch between the identified model and the realized dynamics. This mirrors the control-relevant failure mode observed earlier in the biological network: local dynamics are predictive near the identification regime, while residuals become increasingly consequential as the system moves beyond it. Together, these results suggest that robustness is not a benchmark-specific concern but a systems-level consequence of intervening through models whose validity is necessarily local.

6 Robust Steerability as a Property of Large Networks: Conclusion and Future Work

Beyond measuring whether a network can be steered, S_{rob} provides a notion of *how robustly* that steering can be achieved. Intuitively, it characterizes how much deviation from the identified dynamics can be tolerated before the same model is no longer sufficient for reliable intervention. A larger S_{rob} therefore corresponds to a regime in which steering is less sensitive to modeling error and uncertainty. Computing S_{rob} across models and benchmarks reveals that robust steerability is not

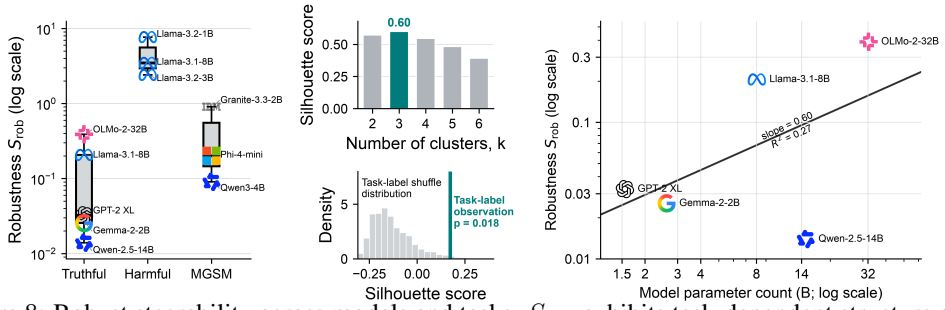

Figure 8: Robust steerability across models and tasks. S_{rob} exhibits task-dependent structure across model–benchmark pairs, while within a fixed task it also varies systematically with model scale.

distributed arbitrarily. Model–task pairs cluster according to the behavioral objective, indicating that robustness is strongly dependent on *what* is being controlled. We then ask whether robust steerability also reflects properties of the network itself. When S_{rob} is examined against parameter count within a fixed task, we observe a positive relationship with model scale. Together, these results suggest that robust steerability is neither purely a property of the task nor purely a property of the model, but emerges from the interaction between the system being controlled and the objective imposed on it. This perspective suggests a broader interpretation of S_{rob}: it may quantify how *careful* an intervention needs to be. Some operating regimes may tolerate coarse models, approximate perturbations, or relatively simple control strategies while still producing reliable outcomes. Others may be substantially more fragile, requiring tighter state estimation, feedback, and explicit robustness to uncertainty. In this sense, S_{rob} offers a way to move beyond asking whether an intervention works, toward asking how much uncertainty the intervention can tolerate before it fails.

Large language models provide an unusual experimental system in which such questions can be studied systematically. They are not biologically realistic models of neural circuitry, but they are large nonlinear networks whose internal states can be measured, modeled, perturbed, and evaluated across many controlled conditions. This makes them useful not as models *of* the brain, but as a model organism for studying general principles of intervention in complex neural networks. Principles identified in such systems—including state dependence, intervention fragility, and robust controllability—may therefore help motivate more mechanistic approaches to neural stimulation. This closes the loop with the neuroscience motivation of our framework. If robust steerability can be quantified in artificial neural systems, an important next question is whether an analogous quantity can be estimated from neural recordings and stimulation experiments. Such a measure could, in principle, distinguish brain states, regions, or stimulation targets that tolerate substantial model uncertainty from those that require more precise closed-loop intervention. This raises a natural question for future work: *Can* S_{rob} *be measured from brain data?*

AI USE STATEMENT

In this work, we used generative AI tools to assist in proving mathematical claims, including checking and refining proof arguments. We did not use generative AI tools to generate synthetic datasets, develop the theoretical or conceptual framework, formulate mathematical claims, propose or refine hypotheses, design the research methodology or experiments, implement the proposed methods, translate research material, clean or reformat datasets, perform qualitative or thematic data analysis, or interpret experimental results. These tasks were carried out by the authors or were not applicable to this work.

Additionally, we used generative AI tools to edit and polish author-written portions of the manuscript for clarity and readability, for retrieval and discovery of relevant literature, and to assist in creating and refining scientific figures and visualizations. We did not use generative AI tools to formulate survey or interview questions, transcribe research material, choose experimental parameters, or create other research artifacts.

We have reviewed all AI-assisted work. References identified with AI assistance were checked against the original sources, mathematical arguments were independently verified and revised by the authors, and AI-assisted figures were reviewed for consistency with the underlying experimental results and technical claims. We take responsibility for the final content of this work, including all text, claims, citations, mathematical results, figures, and other artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We will release the complete implementation of robust activation steering in a public GitHub repository. The release will include the H_∞ synthesis and inference code, benchmark configurations, fixed evaluation subsets and random seeds, model and dataset identifiers, environment specifications, and scripts for artifact construction, calibration, generation, scoring, statistical analysis, and regeneration of the reported tables and figures. The main text defines the method and its assumptions, while the appendices provide the end-to-end algorithm, theoretical derivations, benchmark protocols, model checkpoints, hyperparameter searches, decoding settings, statistical procedures, and computational requirements. Together, these materials support independent verification, controlled comparison, and direct extension of the method. We expect the GitHub release to facilitate community adoption and broaden the positive impact of robust activation steering across AI and machine learning research.

REFERENCES

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. Phi-4-Mini Technical Report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*, 2025. URL https://arxiv.org/abs/2503.01743.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Karl J. Åström and Tore Hägglund. *Advanced PID Control*. ISA—The Instrumentation, Systems, and Automation Society, 2006. ISBN 9781556179426.

Tamer Başar. A dynamic games approach to controller design: Disturbance rejection in discrete-time. *IEEE Transactions on Automatic Control*, 36(8):936–952, 1991. doi: 10.1109/9.133187.

Ishita Basu, Ali Yousefi, Britni Crocker, Rina Zelmann, Angelique C Paulk, Noam Peled, Kristen K Ellard, Daniel S Weisholtz, G Rees Cosgrove, Thilo Deckersbach, et al. Closed-loop enhancement and neural decoding of cognitive control in humans. *Nature biomedical engineering*, 7(4):576–588, 2023.

Til O Bergmann. Brain state-dependent brain stimulation. *Frontiers in psychology*, 9:2108, 2018.

Claire Bradley, Abbey S Nydam, Paul E Dux, and Jason B Mattingley. State-dependent effects of neural stimulation on brain function and cognition. *Nature reviews neuroscience*, 23(8):459–475, 2022.

George Bush and Lisa M Shin. The multi-source interference task: an fmri task that reliably activates the cingulo-frontal-parietal cognitive/attention network. *Nature protocols*, 1(1):308–313, 2006.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K. Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Emily Cheng, Carmen Amo Alonso, Federico Danieli, Arno Blaas, Luca Zappella, Pau Rodriguez, and Xavier Suau. GenCtrl – a formal controllability toolkit for generative models. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2601.05637.

Frédéric Crevecoeur, Stephen H. Scott, and Tyler Cluff. Robust control in human reaching movements: A model-free strategy to compensate for unpredictable disturbances. *Journal of Neuroscience*, 39 (41):8135–8148, 2019. doi: 10.1523/JNEUROSCI.0770-19.2019.

Carlos E Davila, David X Wang, Maxwell Ritzer, Rosalyn Moran, and Bradley C Lega. A control-theoretical system for modulating hippocampal gamma oscillations using stimulation of the posterior cingulate cortex. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 30:2242–2253, 2022.

John C. Doyle, Keith Glover, Pramod P. Khargonekar, and Bruce A. Francis. State-space solutions to standard H_2 and H_∞ control problems. *IEEE Transactions on Automatic Control*, 34(8):831–847, 1989. doi: 10.1109/9.29425.

Abhimanyu Dubey, Abhinav Jauhri, Abhishek Pandey, Akash Kadian, Ahmad Al-Dahle, Andrew Letman, Ankit Mathur, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Kasra Fallah, Leonardo F. Toso, and James Anderson. On the gradient domination of the LQG problem, 2025. arXiv:2507.09026.

Kasra Fallah, Leonardo F. Toso, and James Anderson. Adversarially robust multitask adaptive control. In *Proceedings of the 8th Annual Learning for Dynamics and Control Conference*, 2026. arXiv:2511.05444.

Josef Faller, Jayce Doose, Xiaoxiao Sun, James R Mclntosh, Golbarg T Saber, Yida Lin, Joshua B Teves, Aidan Blankenship, Sarah Huffman, Robin I Goldman, et al. Daily prefrontal closed-loop repetitive transcranial magnetic stimulation (rtms) produces progressive eeg quasi-alpha phase entrainment in depressed adults. *Brain stimulation*, 15(2):458–471, 2022.

Maryam Fazel, Rong Ge, Sham M. Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1467–1476. PMLR, 2018.

Granite Team, IBM. Granite 3.3 Language Models. https://github.com/ibm-granite/granite-3.3-language-models, 2025.

Eldad Haber and Lars Ruthotto. Stable architectures for deep neural networks. *Inverse Problems*, 34 (1):014004, 2017. doi: 10.1088/1361-6420/aa9a90.

Babak Hassibi, Ali H. Sayed, and Thomas Kailath. *Indefinite-Quadratic Estimation and Control: A Unified Approach to H_2 and H_∞ Theories*. Society for Industrial and Applied Mathematics, 1999. doi: 10.1137/1.9781611970760.

Adam O Hebb, Jun Jason Zhang, Mohammad H Mahoor, Christos Tsiokos, Charles Matlack, Howard Jay Chizeck, and Nader Pouratian. Creating the feedback loop: closed loop neurostimulation. *Neurosurgery Clinics of North America*, 25(1):187–204, 2014. doi: 10.1016/j.nec.2013.08.006.

Brandon Hsu, Daniel Beaglehole, Adityanarayanan Radhakrishnan, and Mikhail Belkin. Contextual linear activation steering of language models. *arXiv preprint arXiv:2604.24693*, 2026.

Pablo A. Iglesias, Denis Mustafa, and Keith Glover. Discrete-time H_∞ controllers satisfying a minimum entropy criterion. *Systems & Control Letters*, 14(4):275–286, 1990. doi: 10.1016/0167-6911(90)90048-Y.

Rudolf E. Kalman. Contributions to the theory of optimal control. *Boletín de la Sociedad Matemática Mexicana*, 5(2):102–109, 1960.

Jonathan C. Kao, Paul Nuyujukian, Stephen I. Ryu, Mark M. Churchland, John P. Cunningham, and Krishna V. Shenoy. Single-trial dynamics of motor cortex and their applications to brain-machine interfaces. *Nature Communications*, 6:7759, 2015. doi: 10.1038/ncomms8759.

Frank L. Lewis, Draguna Vrabie, and Vassilis L. Syrmos. *Optimal Control*. John Wiley & Sons, 3 edition, 2012.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022.

Gabriele Maraia, Leonardo Ranaldi, Marco Valentino, and Fabio Massimo Zanzotto. Can activation steering generalize across languages? a study on syllogistic reasoning in language models. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2739–2753, 2026.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*, 2024.

Dung V. Nguyen, Nhi Y. Pham, Hieu M. Vu, Lei Zhang, and Tan M. Nguyen. Activation steering with a feedback controller. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.04309.

Nilearn contributors. Nilearn, 2026. URL https://doi.org/10.5281/zenodo.8397156.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

R. Ravi, K. M. Nagpal, and P. P. Khargonekar. H^∞ control of linear time-varying systems: A state-space approach. *SIAM Journal on Control and Optimization*, 29(6):1394–1413, 1991. doi: 10.1137/0329071.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 15504–15522. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.828.

Pau Rodriguez, Arno Blaas, Michal Klein, Luca Zappella, Nicholas Apostoloff, Marco Cuturi, and Xavier Suau. Controlling language and diffusion models by transporting activations. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=l2zFn6TIQi.

Stephen H. Scott. Optimal feedback control and the neural basis of volitional motor control. *Nature Reviews Neuroscience*, 5:532–545, 2004. doi: 10.1038/nrn1427.

Maryam M. Shanechi. Brain–machine interfaces from motor to mood. *Nature Neuroscience*, 22:1554–1564, 2019. doi: 10.1038/s41593-019-0488-y.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. Language models are multilingual chain-of-thought reasoners. In *International Conference on Learning Representations*, 2023.

Juha Silvanto and Alvaro Pascual-Leone. State-dependency of transcranial magnetic stimulation. *Brain topography*, 21(1):1–10, 2008.

Max Simchowitz. Making non-stochastic control (almost) as easy as stochastic. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18318–18329, 2020.

Max Simchowitz and Dylan Foster. Naive exploration is optimal for online LQR. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8937–8948. PMLR, 2020.

Julian Skifstad, Xinyue Annie Yang, and Glen Chou. Local linearity of LLMs enables activation steering via model-based linear optimal control. *arXiv preprint arXiv:2604.19018*, 2026. doi: 10.48550/arXiv.2604.19018. URL https://arxiv.org/abs/2604.19018.

Stefano Soatto, Paulo Tabuada, Pratik Chaudhari, and Tian Yu Liu. Taming ai bots: Controllability of neural states in large language models, 2023. URL https://arxiv.org/abs/2305.18449.

Daniel Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adrià Garriga-Alonso, and Robert Kirk. Analysing the generalisation and reliability of steering vectors. *Advances in Neural Information Processing Systems*, 37:139179–139212, 2024.

Zecheng Tang, Keyan Zhou, Juntao Li, Baibei Ji, Jianye Hou, and Min Zhang. L-CiteEval: Do long-context models truly leverage context for responding? *arXiv preprint arXiv:2410.02115*, 2024.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, et al. 2 OLMo 2 furious. *arXiv preprint arXiv:2501.00656*, 2024. URL https://arxiv.org/abs/2501.00656.

Emanuel Todorov and Michael I. Jordan. Optimal feedback control as a theory of motor coordination. *Nature Neuroscience*, 5(11):1226–1235, 2002. doi: 10.1038/nn963.

Uros Topalovic, Sam Barclay, Chenkai Ling, Ahmed Alzuhair, Wenhao Yu, Vahagn Hokhikyan, Hariprasad Chandrakumar, Dejan Rozgic, Wenlong Jiang, Sina Basir-Kazeruni, et al. A wearable platform for closed-loop stimulation and recording of single-neuron and local field potential activity in freely moving humans. *Nature neuroscience*, 26(3):517–527, 2023.

Leonardo F. Toso, Kasra Fallah, Charis Stamouli, George J. Pappas, and James Anderson. Multitask LQG control: Performance and generalization bounds, 2026. URL https://arxiv.org/abs/2604.16730.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2024. doi: 10.48550/arXiv.2308.10248. URL https://arxiv.org/abs/2308.10248.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1706.03762.

Saurabh Vyas, Matthew D Golub, David Sussillo, and Krishna V Shenoy. Computation through neural population dynamics. *Annual review of neuroscience*, 43(1):249–275, 2020.

Han Wang, Leonardo Felipe Toso, Aritra Mitra, and James Anderson. Model-free learning with heterogeneous dynamical systems: A federated lqr approach. *Transactions on Machine Learning Research*, 2026.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. On layer normalization in the transformer architecture. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10524–10533. PMLR, 2020. URL https://proceedings.mlr.press/v119/xiong20b.html.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Hongjue Zhao, Haosen Sun, Jiangtao Kong, Xiaochang Li, Qineng Wang, Liwei Jiang, Qi Zhu, Tarek Abdelzaher, Yejin Choi, Manling Li, and Huajie Shao. ODESteer: A unified ODE-based steering framework for LLM alignment. *arXiv preprint arXiv:2602.17560*, 2026. doi: 10.48550/arXiv.2602.17560. URL https://arxiv.org/abs/2602.17560.

Kemin Zhou, John C. Doyle, and Keith Glover. *Robust and Optimal Control*. Prentice Hall, Upper Saddle River, NJ, 1996.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023. URL https://arxiv.org/abs/2310.01405.

Christoph Zrenner and Ulf Ziemann. Closed-loop brain stimulation. *Biological Psychiatry*, 95(6): 545–552, 2024. doi: 10.1016/j.biopsych.2023.09.014.

Christoph Zrenner, Paolo Belardinelli, Florian Müller-Dahlhaus, and Ulf Ziemann. Closed-loop neuroscience and non-invasive brain stimulation: A tale of two loops. *Frontiers in Cellular Neuroscience*, 10:92, 2016. doi: 10.3389/fncel.2016.00092.

A Related Work

Activation steering and representation intervention. Activation steering modifies intermediate representations at inference time to alter model behavior without updating model parameters. Early representation-engineering and contrastive methods identify semantic directions from activation statistics associated with opposing behaviors and intervene along these directions (Zou et al., 2023; Turner et al., 2024). Related approaches have shown that comparatively low-dimensional interventions can affect high-level behaviors such as truthfulness and refusal: inference-time intervention uses internal representations to elicit truthful responses (Li et al., 2023), contrastive activation addition constructs behavioral directions from paired examples (Rimsky et al., 2024), and refusal has been associated with a particularly low-dimensional intervention direction (Arditi et al., 2024). Beyond additive interventions, activation transport methods explicitly match or transform activation distributions, leading to sequential procedures such as Mean-AcT and Linear-AcT (Rodriguez et al., 2025). These results collectively establish that intermediate representations provide a meaningful control surface, but successful intervention under the distribution on which a direction is constructed does not by itself imply reliable intervention after the input or internal trajectory changes. Indeed, steering effectiveness can vary substantially across inputs (Tan et al., 2024). Our focus is precisely this latter question: not only whether a semantic direction can influence behavior, but how sensitive that influence is to uncertainty in the dynamics through which the intervention propagates.

Dynamical and control-theoretic views of neural networks. Interpreting network depth as an evolution variable has a broader precedent in dynamical-systems views of deep networks, including stable residual architectures and neural ordinary differential equations (Haber & Ruthotto, 2017; Chen et al., 2018). Within language models, controllability-based formulations similarly treat generation or representation evolution as a dynamical process subject to intervention (Soatto et al., 2023; Cheng et al., 2026). This viewpoint is particularly natural for activation steering: an intervention applied at one transformer layer changes the state presented to later layers, so steering cannot in general be understood as an isolated perturbation of a static representation. Our formulation adopts this finite-depth dynamical interpretation but focuses on uncertainty in the identified layer-to-layer map rather than assuming that a nominal dynamical model remains valid after calibration.

Feedback and model-based activation control. Recent work has made this dynamical viewpoint operational by formulating activation steering directly as a feedback-control problem (Nguyen et al., 2026; Zhao et al., 2026; Skifstad et al., 2026). PID Steering interprets common difference-in-means steering rules through proportional feedback and augments them with integral and derivative terms to reduce persistent semantic error and oscillatory behavior (Nguyen et al., 2026). Other approaches use continuous-depth or feedback constructions to adapt interventions to the evolving representation (Zhao et al., 2026). Activation-LQR is especially close to our setting: it locally linearizes transformer blocks, treats depth as a finite time-varying control horizon, and uses LQR to track semantic feature setpoints (Skifstad et al., 2026). It further analyzes tracking error induced by the discrepancy between the nonlinear network and its local linear approximation.

Our work begins from the same model-based control abstraction but asks a different question. We do not claim novelty from applying classical optimal control to transformer representations; rather, we treat the local-model discrepancy itself as the object of interest. Instead of asking only which controller performs well under the identified dynamics, we ask whether the *closed-loop amplification of errors in those dynamics* predicts whether semantic control remains reliable when the deployment trajectory departs from calibration. This distinction motivates both our empirical disturbance model and the robust-steerability quantity.

Robust and optimal control. The control machinery underlying our formulation is classical. Linear quadratic control provides optimal feedback for nominal linear dynamics (Kalman, 1960; Lewis et al., 2012), while H_∞ control characterizes and minimizes worst-case amplification from exogenous disturbances to regulated outputs. Finite-horizon and time-varying formulations admit dynamic-game and state-space solutions (Başar, 1991; Ravi et al., 1991; Hassibi et al., 1999), and classical results connect robust synthesis to nominal quadratic control and the broader H_2–H_∞ continuum (Doyle et al., 1989; Iglesias et al., 1990; Zhou et al., 1996). Related robust-control descriptions have also been used to interpret human responses to unpredictable perturbations (Crevecoeur et al., 2019).

Our contribution is therefore not a new general H_∞ synthesis procedure. We instead use the induced closed-loop gain as a quantity for characterizing learned representation dynamics. The disturbance channel is estimated from empirical one-step representation-model residuals rather than postulated as isotropic noise, and the resulting minimum amplification defines a pre-deployment measure of how much identified-model mismatch a steering problem can tolerate. In this sense, robust control serves both as a controller and, more importantly for our purpose, as a diagnostic of intervention reliability.

Open-source language models. We evaluate GPT-2 (Radford et al., 2019), Llama 3/3.1/3.2 (Dubey et al., 2024), Qwen2.5 (Yang et al., 2024), OLMo 2 (Team OLMo et al., 2024), Qwen3 (Yang et al., 2025), and Phi-4-mini (Abouelenin et al., 2025). For technical reports with exceptionally large author lists, we follow the conventional abbreviated *et al.* representation rather than reproducing the complete author list in the bibliography; the cited paper URLs provide the authoritative records and full attribution. Table 11 separately reports the exact model repositories and pinned revisions used in every experiment. As no dedicated Granite 3.3 technical report was publicly available at submission time, we cite the official IBM Granite 3.3 model repository/documentation (Granite Team, IBM, 2025). Thus, the bibliography provides model-family attribution, while the checkpoint identifiers provide model-level provenance and reproducibility.

Learning and sharing across control problems. Our present method identifies local dynamics and synthesizes a controller separately for each model–behavior pair. A complementary literature studies how control policies and system information can instead be learned or shared across related problems. Online and adaptive LQR characterize control when the underlying dynamics must be learned from interaction, including stochastic and adversarial settings (Simchowitz & Foster, 2020; Simchowitz, 2020), while direct policy optimization provides an alternative to explicit model identification (Fazel et al., 2018). Multitask and federated control further study statistical sharing across heterogeneous dynamical systems (Wang et al., 2026; Toso et al., 2026), including robustness to heterogeneity and corrupted participants (Fallah et al., 2026). These perspectives suggest extensions in which representation dynamics or feedback policies are shared across behaviors, models, or calibration distributions rather than reidentified independently.

A related issue arises when the relevant control state is only partially observed. In such settings, the optimization landscape depends on the controller parameterization, and recent results establish gradient-dominance properties for LQG under suitable controller representations (Fallah et al., 2025). For activation steering, analogous formulations could replace the present Jacobian-estimation–Riccati pipeline with controllers learned directly from rollouts or histories. We leave these learning and transfer questions separate from the present goal of determining whether a calibration-time dynamical model already contains a useful certificate of steering reliability.

B Intervention Placement and the Effective Control System

The main text considers a post-block additive intervention,

$$h_{k+1} = \phi_k(h_k) + B_k u_k.$$

Here, we briefly describe how the local control model changes when the intervention is applied at a different location within the transformer.

General intervention model. Let the controlled transformation at layer k be

$$h_{k+1} = F_k(h_k, u_k),$$

where F_k includes both the transformer block and the intervention mechanism. Around a nominal operating point $(h_k^{\mathrm{ref}}, u_k^{\mathrm{ref}})$, define $x_k = h_k - h_k^{\mathrm{ref}}$ and $\delta u_k = u_k - u_k^{\mathrm{ref}}$. If F_k is continuously differentiable, its first-order approximation is

$$x_{k+1} = A_k x_k + G_k \delta u_k + r_k(x_k, \delta u_k),$$

where

$$A_k = \left.\frac{\partial F_k}{\partial h}\right|_{(h_k^{\mathrm{ref}}, u_k^{\mathrm{ref}})}, \qquad G_k = \left.\frac{\partial F_k}{\partial u}\right|_{(h_k^{\mathrm{ref}}, u_k^{\mathrm{ref}})},$$

and r_k collects higher-order terms. Thus, the robust-control formulation in the main text extends directly to any differentiable intervention mechanism; the intervention location enters through the effective control channel G_k.

Post-block intervention. If the intervention is applied after the transformer block,

$$F_k^{\mathrm{post}}(h_k, u_k) = \phi_k(h_k) + B_k u_k,$$

then

$$A_k = \left.\frac{\partial \phi_k}{\partial h}\right|_{h_k^{\mathrm{ref}}}, \qquad G_k = B_k.$$

Under the linear feedback law $\delta u_k = -K_k x_k$, the local closed-loop matrix is therefore

$$A_k^{\mathrm{cl,post}} = A_k - B_k K_k.$$

In this case, the intervention enters affinely and is locally separated from the nonlinear transformation of the current block.

Pre-block intervention. If the intervention is instead applied before the transformer block,

$$F_k^{\mathrm{pre}}(h_k, u_k) = \phi_k(h_k + B_k u_k),$$

let

$$J_k = \left.\frac{\partial \phi_k(\eta)}{\partial \eta}\right|_{\eta = h_k^{\mathrm{ref}} + B_k u_k^{\mathrm{ref}}}.$$

By the chain rule,

$$A_k = J_k, \qquad G_k = J_k B_k,$$

and hence

$$A_k^{\mathrm{cl,pre}} = J_k(I - B_k K_k).$$

Thus, a pre-block intervention is filtered through the local transformer Jacobian before affecting the next representation. Even when $B_k = I$, the effective control channel is J_k rather than the identity. This distinction also affects the local approximation error. For post-block interventions, the control enters affinely within the current block, whereas for pre-block interventions the feedback action is propagated through the block nonlinearity. Consequently, the validity of the local linear model may depend more strongly on the intervention magnitude and feedback gain in the pre-block case.

Intra-block intervention. Interventions inserted within attention, normalization, or the MLP are covered by the same general model

$$h_{k+1} = F_k(h_k, u_k), \qquad G_k = \left.\frac{\partial F_k}{\partial u}\right|_{(h_k^{\mathrm{ref}}, u_k^{\mathrm{ref}})}.$$

Here, G_k is determined by the corresponding computational graph. For example, an intervention before attention can modify the attention pattern, whereas an intervention after attention is propagated only through subsequent operations in the block. Such effects can be incorporated either through the identified effective channel or, when they vary across prompts, through the uncertainty model.

Takeaway. The choice of intervention location does not change the form of the local robust-control problem, but it changes its effective actuator. Post-block steering yields $G_k = B_k$, while pre- and intra-block interventions couple the steering direction to the local transformer dynamics. Accordingly, robust steerability should be interpreted with respect to the model–behavior–intervention tuple rather than the hidden representation alone.

C METHODOLOGICAL DETAILS

This appendix specifies the calibration, synthesis, and deployment procedure used in the reported experiments. All quantities are constructed separately for each model–benchmark pair; we suppress this dependence when unambiguous. We use $\mathcal{D}_{\mathrm{dir}}$, $\mathcal{D}_{\mathrm{dyn}}$, $\mathcal{D}_{\mathrm{dist}}$, and $\mathcal{D}_{\mathrm{tune}}$ for behavior direction fitting, dynamics estimation, disturbance fitting, and controller selection, respectively. These roles are benchmark specific and need not be mutually disjoint; the exact splits are given in Appendix G. All selected artifacts are frozen before the corresponding held-out shift is evaluated.

C.1 REDUCED COORDINATES AND REFERENCE CONSTRUCTION

At layer k, let $\bar{h}_k^+$ and $\bar{h}_k^-$ denote the mean final-token residual-stream activations for the desired and undesired behavior examples. Their contrast $d_k = \bar{h}_k^+ - \bar{h}_k^-$ defines the semantic direction, with $v_{k,1} = d_k / \|d_k\|_2$. After removing the component along $v_{k,1}$ from the centered calibration activations, we retain the leading principal directions of the remainder and orthonormalize the resulting basis $V_k = [v_{k,1}, \ldots, v_{k,r_x}]$, where $V_k^\top V_k = I$. All reported experiments use $r_x = 8$.

The reduced activation is $\widetilde{h}_k = V_k^\top h_k$. The reference $\widetilde{h}_k^{\mathrm{ref}}$ places the first coordinate at the frozen semantic setpoint and all remaining coordinates at zero; specifically, its first coordinate is $\beta \|d_k\|_2$, where β is the benchmark-specific setpoint multiplier. We therefore use the controller state $\widetilde{x}_k = \widetilde{h}_k - \widetilde{h}_k^{\mathrm{ref}}$. The basis, setpoint, and reference state are fixed after calibration.

C.2 PROJECTED DYNAMICS AND EMPIRICAL DISTURBANCE GEOMETRY

For prompt $i \in \mathcal{D}_{\mathrm{dyn}}$, let $J_k^{(i)} = \partial \phi_k(h)/\partial h|_{h=h_k^{(i)}}$ denote the local block Jacobian. Since the controller is shared across prompts, we estimate one projected transition matrix per layer by averaging the prompt-specific Jacobians,

$$\widetilde{A}_k = \frac{1}{N_{\mathrm{dyn}}} \sum_{i \in \mathcal{D}_{\mathrm{dyn}}} V_{k+1}^\top J_k^{(i)} V_k.$$

Only Jacobian–vector products along the reduced basis are required; full hidden-dimensional Jacobians are not retained. Let $\widetilde{B}_k$ denote the intervention map in the same coordinates. In the reported post-block implementation, the reduced control is lifted to the residual stream dimension and added to the block output.

We suppress tildes below and write the fitted local model as $x_{k+1} = A_k x_k + B_k u_k + \xi_k$. For each $i \in \mathcal{D}_{\mathrm{dist}}$, the one-step residual is $\xi_k^{(i)} = x_{k+1}^{(i)} - A_k x_k^{(i)} - B_k u_k^{(i)}$. Let $\bar{\xi}_k = N_{\mathrm{dist}}^{-1} \sum_i \xi_k^{(i)}$ and form

$$\Xi_k = \frac{1}{\sqrt{N_{\mathrm{dist}} - 1}} \left[\xi_k^{(1)} - \bar{\xi}_k \quad \cdots \quad \xi_k^{(N_{\mathrm{dist}})} - \bar{\xi}_k \right] = U_k \Sigma_k W_k^\top.$$

The disturbance channel is the covariance factor $D_k = U_{k,r} \Sigma_{k,r}$, where r is the numerical rank of Ξ_k. Consequently,

$$D_k D_k^\top = \Xi_k \Xi_k^\top = \widehat{\Sigma}_k,$$

the unbiased empirical covariance of the centered transition residuals. We retain all empirically supported disturbance directions in the reduced state space and use the covariance factor at its estimated scale, without further rank reduction or normalization. Since $r_x = 8$ and each disturbance-fitting split contains more than nine samples, $r = 8$ unless the covariance is numerically rank deficient. Before synthesis, we verify $\left\| D_k D_k^\top - \widehat{\Sigma}_k \right\|_F / \left\| \widehat{\Sigma}_k \right\|_F \leq 10^{-4}$.

The robust surrogate is therefore $x_{k+1} = A_k x_k + B_k u_k + D_k w_k$, where w_k is an adversarial synthesis variable indexing the empirically observed directions of model mismatch; it is not estimated during generation.

Algorithm 1 Calibration and deployment of robust activation steering

Require: Frozen model f_θ; data $\mathcal{D}_{\text{dir}}, \mathcal{D}_{\text{dyn}}, \mathcal{D}_{\text{dist}}, \mathcal{D}_{\text{tune}}$; candidate costs $\mathcal{C}$
Ensure: Selected controller $\{K_k^\star\}_{k=1}^L$ and score S_{rob}
1: Fit semantic directions, reduced bases $\{V_k\}$, and references from $\mathcal{D}_{\text{dir}}$
2: Estimate projected dynamics $\{A_k, B_k\}$ from $\mathcal{D}_{\text{dyn}}$
3: Compute one-step residuals on $\mathcal{D}_{\text{dist}}$ and fit $\{D_k\}$ from their full empirical covariance
4: **for** $c \in \mathcal{C}$ **do**
5: $\quad(\gamma^\star(c), \{K_k^{(c)}\}) \leftarrow \textsc{RobustSynthesis}(\{A_k, B_k, D_k\}, c)$
6: $\quad$Generate and score $\mathcal{D}_{\text{tune}}$ with $\{K_k^{(c)}\}$
7: **end for**
8: Select $c^\star \leftarrow \arg\max_{c\in\mathcal{C}} \text{Score}_{\text{cal}}(K^{(c)})$
9: $\gamma^\star \leftarrow \gamma^\star(c^\star)$, $\{K_k^\star\} \leftarrow \{K_k^{(c^\star)}\}$, $S_{\text{rob}} \leftarrow 1/\gamma^\star$
10: Freeze all calibration artifacts and deploy $\{K_k^\star\}$

C.3 Robust Synthesis and Calibration Selection

For a candidate cost configuration $c = (Q, R, Q_f)$, we solve the finite-horizon zero-sum game

$$J_\gamma = x_{L+1}^\top Q_f x_{L+1} + \sum_{k=1}^{L} \left(x_k^\top Q_k x_k + u_k^\top R_k u_k - \gamma^2 w_k^\top w_k \right),$$

where the controller minimizes J_γ and the disturbance maximizes it. The reported experiments use isotropic reduced-coordinate state, intervention, and terminal penalties; candidate ratios are listed in Appendix H.

Starting from $P_{L+1} = Q_f$, define $M_k = \gamma^2 I - D_k^\top P_{k+1} D_k$. A candidate γ is infeasible if $M_k \not\succ 0$. Otherwise,

$$G_k = R_k + B_k^\top P_{k+1} B_k + B_k^\top P_{k+1} D_k M_k^{-1} D_k^\top P_{k+1} B_k,$$
$$F_k = B_k^\top P_{k+1} A_k + B_k^\top P_{k+1} D_k M_k^{-1} D_k^\top P_{k+1} A_k.$$

When $G_k \succ 0$, the minimizing feedback is $u_k = -K_k x_k$ with $K_k = G_k^{-1} F_k$, and

$$P_k = Q_k + A_k^\top P_{k+1} A_k + A_k^\top P_{k+1} D_k M_k^{-1} D_k^\top P_{k+1} A_k - F_k^\top G_k^{-1} F_k.$$

For fixed c, the smallest feasible attenuation is $\gamma^\star(c) = \inf\{\gamma > 0 : M_k \succ 0,\ G_k \succ 0,\ \forall k\}$.

Each candidate $c \in \mathcal{C}$ yields its own $\gamma^\star(c)$ and gains $\{K_k^{(c)}\}$. We evaluate these controllers on $\mathcal{D}_{\text{tune}}$ and select

$$c^\star = \arg\max_{c\in\mathcal{C}} \text{Score}_{\text{cal}}\left(K^{(c)}\right), \qquad S_{\text{rob}} = \frac{1}{\gamma^\star(c^\star)}.$$

Thus, for model m and benchmark b, $\gamma^\star_{m,b} = \gamma^\star(c^\star_{m,b})$ and $S_{\text{rob}}^{m,b} = 1/\gamma^\star_{m,b}$. Robust steerability therefore refers to the calibration-selected controller for that model–benchmark instance; we omit (m, b) elsewhere when it is clear from context.

C.4 End-to-End Calibration and Deployment

Algorithm 1 summarizes the complete procedure.

C.5 Numerical Synthesis

For every candidate cost, $\gamma^\star$ is computed by bisection over the initial interval $[0, 100]$, with at most 100 iterations and relative tolerance 10^{-5}. Positive definiteness of M_k and G_k is checked by Cholesky

Algorithm 2 Finite-horizon robust synthesis

function ROBUSTSYNTHESIS($\{A_k, B_k, D_k\}_{k=1}^{L}, Q, R, Q_f$)
  $\gamma_{\mathrm{lo}} \leftarrow 0, \gamma_{\mathrm{hi}} \leftarrow 100$
  for $j = 1, \ldots, 100$ **do**
    $\gamma \leftarrow (\gamma_{\mathrm{lo}} + \gamma_{\mathrm{hi}})/2$, $P_{L+1} \leftarrow Q_f$, feasible $\leftarrow$ true
    for $k = L, \ldots, 1$ **do**
      $M_k \leftarrow \gamma^2 I - D_k^\top P_{k+1} D_k$
      if $M_k \not\succ 0$ **then**
        feasible $\leftarrow$ false; **break**
      end if
      Compute G_k, F_k
      if $G_k \not\succ 0$ **then**
        feasible $\leftarrow$ false; **break**
      end if
      Solve $G_k K_k = F_k$ and update P_k
      $P_k \leftarrow (P_k + P_k^\top)/2$
    end for
    if feasible **then**
      $\gamma_{\mathrm{hi}} \leftarrow \gamma$
    else
      $\gamma_{\mathrm{lo}} \leftarrow \gamma$
    end if
    if $(\gamma_{\mathrm{hi}} - \gamma_{\mathrm{lo}})/\gamma_{\mathrm{hi}} \le 10^{-5}$ **then**
      break
    end if
  end for
  $\widehat{\gamma}^\star \leftarrow \gamma_{\mathrm{hi}}$
  Set $\gamma_{\mathrm{deploy}} \leftarrow 1.01\,\widehat{\gamma}^\star$ and recompute the final gains at γ_{deploy}
  return $\widehat{\gamma}^\star, \{K_k^\star\}_{k=1}^{L}$
end function

factorization; all linear systems are solved directly rather than by forming explicit inverses, and P_k is symmetrized after each backward update.

The 1% margin avoids deployment directly at the numerical feasibility boundary. The score uses the smallest numerically feasible attenuation $\widehat{\gamma}^\star$, whereas the deployed gains are recomputed at $\gamma_{\mathrm{deploy}} = 1.01\,\widehat{\gamma}^\star$. Calibration tensors are accumulated in `float32`, and reduced-coordinate construction uses a numerical norm floor of 10^{-4}.

C.6 INFERENCE

At generation time, the basis, reference, disturbance model, selected costs, and gains remain fixed. At layer k, the forward hook forms the current reduced state error $x_k = V_k^\top h_k - \widetilde{h}_k^{\mathrm{ref}}$ and applies $u_k = -K_k^\star x_k$. The intervention is lifted to the original residual stream dimension and added to the block output. This is repeated for the final processed token at every controlled layer and autoregressive step. The controller is reset between independent prompts, while the disturbance variables $\{w_k\}$ appear only in offline synthesis and are never estimated during inference.

D CONTROL-THEORETIC COMPARISON OF LQR, H_∞, AND PID

This appendix clarifies the distinct control objectives underlying the controller families considered in our experiments and provides a controlled interpretation of Figure 9. The comparison is illustrative: it isolates differences between nominal optimal control, worst-case disturbance attenuation, and reactive error feedback on a common finite-horizon system. It should not be interpreted as additional evidence about downstream language-model performance, which is evaluated separately in Appendix G. We refer to the nominal quadratic controller as LQR below; A-LQR denotes its activation-steering implementation in the experiments.

D.1 CONTROLLER OBJECTIVES

Consider the reduced depth-indexed dynamics used throughout the paper, $x_{k+1} = A_k x_k + B_k u_k + D_k w_k$, for $k = 1, \dots, L$. For comparison, define the finite-horizon quadratic performance

$$J = x_{L+1}^\top Q_f x_{L+1} + \sum_{k=1}^{L} \left(x_k^\top Q_k x_k + u_k^\top R_k u_k \right),$$

with $Q_k, Q_f \succeq 0$ and $R_k \succ 0$. Equivalently, the regulated output z is chosen so that $\|z\|_2^2 = J$. In the controlled example of Figure 9, LQR and H_∞ use the same dynamics, state and intervention coordinates, quadratic weights, initial condition, and intervention channel. This matching is used only to isolate the effect of the synthesis objective in the illustrative comparison; the benchmark-specific controller-selection procedure is given in Appendix C.

Nominal LQR. LQR computes a state-feedback law $u_k = -K_k^{\mathrm{LQR}} x_k$ minimizing the nominal quadratic objective,

$$K^{\mathrm{LQR}} \in \arg\min_K J(K; w = 0).$$

It therefore gives the optimal tracking–effort trade-off for the identified nominal dynamics (Kalman, 1960; Lewis et al., 2012). Although the resulting policy is feedback and reacts to state deviations online, its synthesis objective does not explicitly penalize the amplification of model mismatch.

H_∞ control. The robust controller instead minimizes the induced amplification from the calibrated disturbance coordinates to the regulated output,

$$K^{H_\infty} \in \arg\min_K \sup_{w \neq 0} \frac{\|z\|_2}{\|w\|_2}.$$

The optimal value is the attenuation $\gamma^\star$, equivalently the smallest level for which $\|z\|_2 \leq \gamma^\star \|w\|_2$ for every admissible finite-energy disturbance when $x_1 = 0$ (Doyle et al., 1989; Ravi et al., 1991). Unlike LQR, synthesis therefore uses the empirical disturbance geometry $\{D_k\}$ and may accept greater intervention effort on individual trajectories in order to reduce unfavorable disturbance amplification. For a nonzero initial deviation, the corresponding finite-horizon storage inequality takes the form

$$\|z\|_2^2 \leq \gamma^2 \|w\|_2^2 + x_1^\top P_1 x_1,$$

so error already present at the initial state is distinct from disturbance amplification.

PID steering. PID does not require an identified layer-to-layer dynamics model. Let $e_k = C_{\mathrm{tar},k} x_k$ denote the observed semantic error. A depth-indexed PID intervention has the form $u_k = -(K_k^{\mathrm{P}} e_k + K_k^{\mathrm{I}} \sum_{j=1}^{k} e_j + K_k^{\mathrm{D}} (e_k - e_{k-1}))$ (Åström & Hägglund, 2006; Nguyen et al., 2026). Its proportional, integral, and derivative terms respond to current, accumulated, and changing semantic error, respectively. This provides inexpensive reactive feedback, but the gains do not in

Table 2: **Control objectives of the compared controller families.** Guarantees refer to the identified finite-horizon surrogate, not directly to the nonlinear language model.

Controller	Synthesis information	Design objective	Interpretation
PID	Semantic error	Reactive error correction	No general finite-horizon optimum
LQR	A_k, B_k, Q_k, R_k	Nominal quadratic performance	Optimal nominal tracking–effort trade-off
H_∞	A_k, B_k, D_k, Q_k, R_k	Worst-case disturbance amplification	Finite-energy attenuation certificate

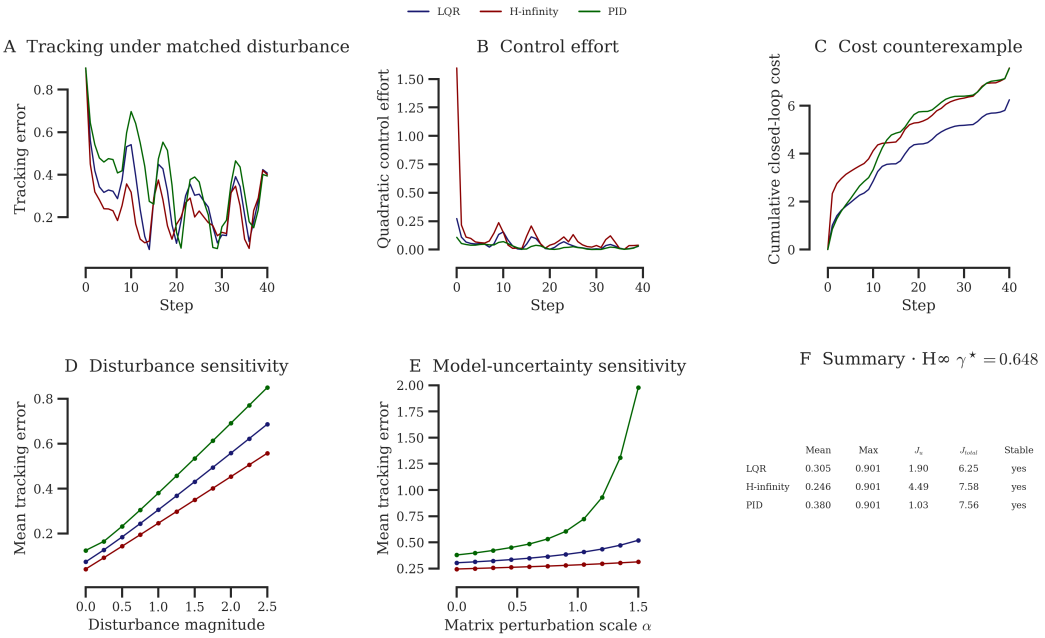

Figure 9: **Controlled comparison of LQR, H_∞, and PID. A:** Tracking error under a fixed disturbance sequence entering through the calibrated disturbance channel. **B:** Layer-wise intervention effort. **C:** Cumulative quadratic cost; robust attenuation need not minimize nominal cost on an individual rollout. **D:** Empirical sensitivity as the magnitude of a fixed disturbance direction is increased. This is not an empirical computation of the worst-case induced gain. **E:** Stress test under perturbations of the transition matrices; this experiment lies outside the formal additive-disturbance certificate. **F:** Summary statistics for the illustrated rollout.

general minimize a global finite-horizon quadratic objective or certify disturbance attenuation for a time-varying transformer surrogate.

The objectives are therefore not interchangeable. In the matched comparison, H_∞ minimizes the induced disturbance gain over a controller class that also contains the LQR policy, so $\gamma(K^{H_\infty}) \leq \gamma(K^{\mathrm{LQR}})$. This does *not* imply pointwise dominance: for a particular disturbance realization w, either controller may incur smaller tracking error, intervention effort, or total quadratic cost. No analogous universal ordering with PID follows from its synthesis objective.

D.2 Controlled Finite-Horizon Illustration

Figure 9 illustrates these distinctions on a common finite-horizon linear time-varying system. All controllers are evaluated from the same initial state and intervention convention. LQR and H_∞ are synthesized from the same nominal dynamics and performance weights, while PID gains are selected on the nominal instance and then held fixed throughout the stress tests.

In the illustrated rollout, H_∞ achieves lower mean tracking error than LQR and PID while using greater intervention energy. Its realized quadratic cost can consequently exceed that of LQR: the nominal controller optimizes the particular quadratic objective at $w = 0$, whereas the robust controller optimizes the largest amplification over the disturbance class. The identical maximum error observed across the three trajectories occurs at their shared initial condition, before feedback has differentiated the closed-loop responses, and is therefore not informative about comparative disturbance attenuation.

Panel D provides a complementary empirical sensitivity test. Holding a normalized disturbance sequence $\bar{w}$ fixed and varying $w_k(\alpha) = \alpha \bar{w}_k$ probes the response along one prescribed disturbance direction. A flatter H_∞ response is consistent with reduced sensitivity along that direction, but the certified quantity $\gamma^\star$ concerns the supremum over *all* finite-energy disturbance sequences represented by $\{D_k\}$; a one-dimensional sweep is therefore illustrative rather than a substitute for the induced-gain certificate.

D.3 Dynamics Mismatch and Scope of the Comparison

Panel E perturbs the identified transition matrices according to $A_k(\alpha) = A_k + \alpha \Delta A_k$. Along a particular trajectory, the additional term $\alpha \Delta A_k x_k$ acts as a state-dependent model error. If this error is well represented by the empirical disturbance geometry $\mathrm{range}(D_k)$, the robust controller may attenuate part of its effect. The additive H_∞ certificate, however, does not provide a formal guarantee against arbitrary perturbations of A_k: such a statement would require an explicit multiplicative or structured uncertainty model. Panel E should therefore be interpreted as an empirical stress test of the controller rather than as a consequence of the theorem.

The comparison highlights three distinct regimes relevant to activation steering. PID provides lightweight feedback when a useful semantic error can be measured directly; LQR exploits an identified dynamical model to optimize nominal tracking and intervention effort; and H_∞ additionally uses the empirically measured disturbance geometry to limit the propagation of unfavorable model mismatch. These distinctions concern synthesis objectives, not a universal ranking of the methods. In the language-model experiments, downstream performance depends on the benchmark, model, distribution shift, and calibration-selected operating point, and robust attenuation may trade increased intervention effort for greater reliability away from the calibration regime.

Finally, all guarantees remain local to the identified reduced-order dynamics. They certify finite-horizon attenuation of additive disturbances represented through $\{D_k\}$ and do not imply global robustness of the nonlinear transformer, asymptotic stability across unbounded depth, or uniform superiority on every generated trajectory. Their role in this paper is to characterize how uncertainty measured during calibration is propagated through the closed-loop representation dynamics.

E PROOFS

Throughout this appendix, $\|\cdot\|_2$ denotes the Euclidean norm for vectors and the induced spectral norm for matrices. The terminal quadratic cost is included in the stacked regulated output, and we use the feedback convention $u_k = -K_k x_k$.

E.1 PROOF OF PROPOSITION 1

Proof. Fix an admissible controller K. Because the horizon is finite, unrolling the closed-loop dynamics with $x_1 = 0$ defines a finite-dimensional linear map from the stacked disturbance to the stacked regulated output:

$$z = \mathcal{T}_K w.$$

Consequently, the induced disturbance-to-output gain is

$$\gamma(K) = \sup_{w \neq 0} \frac{\|\mathcal{T}_K w\|_2}{\|w\|_2} = \|\mathcal{T}_K\|_2 = \sigma_{\max}(\mathcal{T}_K).$$

The last equality is the variational characterization of the largest singular value. If $w^\star$ is a unit-norm principal right singular vector of $\mathcal{T}_K$, then

$$\|\mathcal{T}_K w^\star\|_2 = \sigma_{\max}(\mathcal{T}_K),$$

so $w^\star$ attains the worst-case gain. If the largest singular value is repeated, every unit vector in the corresponding right singular subspace is a worst-case disturbance.

It remains to establish the attenuation certificate. For a fixed attenuation level γ, write the finite-horizon game objective as

$$J_\gamma(K, w) = \|z\|_2^2 - \gamma^2 \|w\|_2^2,$$

where $\|z\|_2^2$ includes the terminal quadratic term. Feasibility of the finite-horizon H_∞ recursion yields value matrices $P_k \succeq 0$ and the synthesized feedback law such that, for every disturbance w_k,

$$\ell_k(x_k, -K_k x_k) - \gamma^2 \|w_k\|_2^2 + V_{k+1}(x_{k+1}) \leq V_k(x_k), \qquad V_k(x) = x^\top P_k x.$$

Here ℓ_k is the regulated stage cost and $V_{L+1}(x) = x^\top Q_f x$ is the terminal cost. Summing this inequality over $k = 1, \ldots, L$ telescopes the storage terms and gives

$$\|z\|_2^2 - \gamma^2 \|w\|_2^2 \leq V_1(x_1).$$

Since $x_1 = 0$, the right-hand side vanishes. Therefore

$$\|z\|_2 \leq \gamma \|w\|_2$$

for every disturbance sequence entering through the calibrated channels $\{D_k\}_{k=1}^L$, as claimed. □

E.2 PROOF OF PROPOSITION 2

We begin with the standard finite-horizon, discrete-time disturbance-attenuation recursion (Başar, 1991, Thms. 3.1 and 4.1); see also Hassibi et al. (1999, Ch. 9) and Zhou et al. (1996). The normalized recursion extends to general $R_k \succ 0$ by the input change of coordinates $\widetilde{u}_k = R_k^{1/2} u_k$.

For a fixed γ, define

$$M_k^{(\gamma)} = \gamma^2 I - {D_k}^\top P_{k+1}^{(\gamma)} D_k.$$

Whenever $M_k^{(\gamma)} \succ 0$, maximization over the disturbance gives the robustified value matrix

$$\widehat{P}_{k+1}^{(\gamma)} = P_{k+1}^{(\gamma)} + P_{k+1}^{(\gamma)} D_k \left(M_k^{(\gamma)}\right)^{-1} {D_k}^\top P_{k+1}^{(\gamma)}.$$

The corresponding gain and value recursions are

$${K_k}^{(\gamma)} = \left(R_k + {B_k}^\top \widehat{P}_{k+1}^{(\gamma)} B_k\right)^{-1} {B_k}^\top \widehat{P}_{k+1}^{(\gamma)} A_k$$

and

$$\begin{aligned} P_k^{(\gamma)} = Q_k &+ {A_k}^\top \widehat{P}_{k+1}^{(\gamma)} A_k \\ &- {A_k}^\top \widehat{P}_{k+1}^{(\gamma)} B_k \left(R_k + {B_k}^\top \widehat{P}_{k+1}^{(\gamma)} B_k\right)^{-1} {B_k}^\top \widehat{P}_{k+1}^{(\gamma)} A_k, \end{aligned}$$

with $P_{L+1}^{(\gamma)} = Q_f$.

Proof. The qualitative reduction of the zero-sum game to the ordinary linear regulator as $\gamma \to \infty$ is classical (Başar, 1991, proof of Thm. 5.2). We establish the stated rate and uniformity over the finite horizon.

Define the nominal gain and Riccati maps

$$\mathscr{K}_k(X) = \left(R_k + {B_k}^\top X B_k\right)^{-1} {B_k}^\top X A_k$$

and

$$\mathscr{R}_k(X) = Q_k + {A_k}^\top X A_k - {A_k}^\top X B_k \left(R_k + {B_k}^\top X B_k\right)^{-1} {B_k}^\top X A_k.$$

Then

$${K_k}^{\mathrm{LQR}} = \mathscr{K}_k(P_{k+1}^{\mathrm{LQR}}), \qquad P_k^{\mathrm{LQR}} = \mathscr{R}_k(P_{k+1}^{\mathrm{LQR}}).$$

For the robust recursion, define

$$\mathscr{H}_{k,\gamma}(P) = P + P D_k \left(\gamma^2 I - {D_k}^\top P D_k\right)^{-1} {D_k}^\top P.$$

Thus,

$${K_k}^{(\gamma)} = \mathscr{K}_k\left(\mathscr{H}_{k,\gamma}(P_{k+1}^{(\gamma)})\right), \qquad P_k^{(\gamma)} = \mathscr{R}_k\left(\mathscr{H}_{k,\gamma}(P_{k+1}^{(\gamma)})\right).$$

We first note that both recursions preserve positive semidefiniteness. If $P \succeq 0$ and $\gamma^2 I - {D_k}^\top P D_k \succ 0$, then $\mathscr{H}_{k,\gamma}(P) \succeq 0$. Moreover, for $X \succeq 0$,

$$\begin{bmatrix} X & X B_k \\ {B_k}^\top X & R_k + {B_k}^\top X B_k \end{bmatrix} = \begin{bmatrix} I & 0 \\ {B_k}^\top & I \end{bmatrix} \begin{bmatrix} X & 0 \\ 0 & R_k \end{bmatrix} \begin{bmatrix} I & B_k \\ 0 & I \end{bmatrix} \succeq 0.$$

Because $R_k + {B_k}^\top X B_k \succ 0$, taking the Schur complement gives

$$X - X B_k \left(R_k + {B_k}^\top X B_k\right)^{-1} {B_k}^\top X \succeq 0.$$

Hence $\mathscr{R}_k(X) \succeq 0$. Starting from $Q_f \succeq 0$, all value matrices are therefore positive semidefinite whenever the robust recursion is well defined.

We now prove by backward induction that there exist constants $\Gamma_k \geq 1$ and $C_k < \infty$, independent of γ, such that the recursion from layer k onward is feasible for every $\gamma \geq \Gamma_k$ and

$$\left\| P_k^{(\gamma)} - P_k^{\mathrm{LQR}} \right\|_2 \leq C_k \gamma^{-2}.$$

At the terminal layer,

$$P_{L+1}^{(\gamma)} = P_{L+1}^{\mathrm{LQR}} = Q_f,$$

so we may take $\Gamma_{L+1} = 1$ and $C_{L+1} = 0$.

Assume the claim holds at layer $k+1$. For $\gamma \geq \Gamma_{k+1}$,

$$\left\| P_{k+1}^{(\gamma)} \right\|_2 \leq \left\| P_{k+1}^{\mathrm{LQR}} \right\|_2 + C_{k+1} =: \bar{p}_{k+1}.$$

Set

$$d_k = \|D_k\|_2^2 \, \bar{p}_{k+1}, \qquad \widetilde{\Gamma}_k = \max\left\{\Gamma_{k+1}, \sqrt{2d_k}, 1\right\}.$$

For every $\gamma \geq \widetilde{\Gamma}_k$,

$${D_k}^\top P_{k+1}^{(\gamma)} D_k \preceq d_k I \preceq \frac{\gamma^2}{2} I,$$

and hence

$$M_k^{(\gamma)} \succeq \frac{\gamma^2}{2} I \succ 0, \qquad \left\| (M_k^{(\gamma)})^{-1} \right\|_2 \leq \frac{2}{\gamma^2}.$$

This proves feasibility at layer k. It also gives

$$\begin{aligned} &\left\| \mathscr{H}_{k,\gamma}(P_{k+1}^{(\gamma)}) - P_{k+1}^{(\gamma)} \right\|_2 \\ &\quad \leq \left\| P_{k+1}^{(\gamma)} \right\|_2^2 \|D_k\|_2^2 \left\| (M_k^{(\gamma)})^{-1} \right\|_2 \\ &\quad \leq 2 \bar{p}_{k+1}^2 \, \|D_k\|_2^2 \, \gamma^{-2}. \end{aligned}$$

Therefore,

$$\left\| \mathscr{H}_{k,\gamma}(P_{k+1}^{(\gamma)}) - P_{k+1}^{\mathrm{LQR}} \right\|_2 \le \widehat{C}_{k+1}\gamma^{-2},$$

where

$$\widehat{C}_{k+1} = C_{k+1} + 2\bar{p}_{k+1}^2 \left\| D_k \right\|_2^2 .$$

It remains to propagate this perturbation through the LQR maps. Let

$$S_k(X) = R_k + {B_k}^\top X B_k.$$

For every $X \succeq 0$,

$$S_k(X) \succeq rI, \qquad \left\| S_k(X)^{-1} \right\|_2 \le r^{-1}.$$

For $X, Y \succeq 0$, the inverse-difference identity yields

$$S_k(X)^{-1} - S_k(Y)^{-1} = S_k(X)^{-1} {B_k}^\top (Y - X) B_k S_k(Y)^{-1},$$

and therefore

$$\left\| S_k(X)^{-1} - S_k(Y)^{-1} \right\|_2 \le r^{-2} \left\| B_k \right\|_2^2 \left\| X - Y \right\|_2 .$$

Let

$$a_k = \left\| A_k \right\|_2 , \qquad b_k = \left\| B_k \right\|_2 ,$$

and, for the application below, set

$$m_k = \left\| P_{k+1}^{\mathrm{LQR}} \right\|_2 + \widehat{C}_{k+1}.$$

Because $\gamma \ge 1$, the preceding perturbation bound ensures that both $X = \mathscr{H}_{k,\gamma}(P_{k+1}^{(\gamma)})$ and $Y = P_{k+1}^{\mathrm{LQR}}$ satisfy $\left\| X \right\|_2 , \left\| Y \right\|_2 \le m_k$. On this positive-semidefinite norm ball, a direct expansion gives

$$\left\| \mathscr{K}_k(X) - \mathscr{K}_k(Y) \right\|_2 \le L_k^K \left\| X - Y \right\|_2 ,$$

where one may take

$$L_k^K = a_k b_k \left(r^{-1} + r^{-2} b_k^2 m_k \right).$$

Furthermore,

$$\left\| \mathscr{K}_k(X) \right\|_2 \le r^{-1} a_k b_k m_k,$$

and

$$\begin{aligned} \mathscr{R}_k(X) - \mathscr{R}_k(Y) = {} & {A_k}^\top (X - Y) A_k \\ & - {A_k}^\top (X - Y) B_k \mathscr{K}_k(X) \\ & - {A_k}^\top Y B_k \left[\mathscr{K}_k(X) - \mathscr{K}_k(Y) \right]. \end{aligned}$$

It follows that

$$\left\| \mathscr{R}_k(X) - \mathscr{R}_k(Y) \right\|_2 \le L_k^P \left\| X - Y \right\|_2 ,$$

where, for example,

$$L_k^P = a_k^2 + a_k b_k m_k \left(r^{-1} a_k b_k + L_k^K \right).$$

Apply these bounds with

$$X = \mathscr{H}_{k,\gamma}(P_{k+1}^{(\gamma)}), \qquad Y = P_{k+1}^{\mathrm{LQR}}.$$

Then

$$\left\| {K_k}^{(\gamma)} - {K_k}^{\mathrm{LQR}} \right\|_2 \le L_k^K \widehat{C}_{k+1} \gamma^{-2}$$

and

$$\left\| P_k^{(\gamma)} - P_k^{\mathrm{LQR}} \right\|_2 \le L_k^P \widehat{C}_{k+1} \gamma^{-2}.$$

Thus, setting

$$\Gamma_k = \widetilde{\Gamma}_k, \qquad C_k = L_k^P \widehat{C}_{k+1},$$

closes the induction.

Finally, because the horizon is finite, define

$$\Gamma = \max_{1 \le k \le L+1} \Gamma_k, \quad C_P = \max_{1 \le k \le L+1} C_k, \quad C_K = \max_{1 \le k \le L} L_k^K \widehat{C}_{k+1}.$$

All three constants are finite and independent of γ. Hence, for every $\gamma \geq \Gamma$, the robust recursion is feasible and

$$\max_{1\leq k\leq L+1} \left\| P_k^{(\gamma)} - P_k^{\mathrm{LQR}} \right\|_2 \leq \frac{C_P}{\gamma^2},$$

$$\max_{1\leq k\leq L} \left\| {K_k}^{(\gamma)} - {K_k}^{\mathrm{LQR}} \right\|_2 \leq \frac{C_K}{\gamma^2}.$$

If Proposition 2 is stated using a single constant, take $C = C_P + C_K$. This proves the claim. □

F Neural Dynamics under Conflict and Stimulation-Context Shifts

The neural analysis tests the empirical premise underlying our disturbance model: a locally identified dynamical system can remain predictive near its identification regime while accumulating structured error when task or intervention context changes. We evaluate this premise across participants, anatomical leads, and prediction horizons using the released intracranial recordings of Basu et al. (Basu et al., 2023).

F.1 Source recordings and task contexts

The source study recorded intracranial electrophysiology from 21 participants undergoing epilepsy monitoring while they performed the Multi-Source Interference Task (MSIT). MSIT induces graded cognitive conflict, and the released data include trial-level conflict labels, behavioral responses, and participant-specific electrode coordinates. The study additionally examined theta-band markers of cognitive control and tested open-loop and closed-loop stimulation. These components provide two complementary shifts for dynamical analysis: conflict changes within no-stimulation recordings and recording contexts associated with open- or closed-loop stimulation sessions.

Twenty participants had an eligible no-stimulation recording for the cross-participant conflict screen. P13 is absent because the released data for that participant contain only a stimulation session. For participants with multiple no-stimulation sessions, we use the session containing the largest number of trials, with the later session resolving ties. Channels marked as ictal are excluded, and an anatomical lead enters the screen when at least five clean bipolar channels remain. This yields 263 participant–lead combinations across the 20 included participants.

F.2 Neural state and temporal representation

For an eligible lead, the simultaneous bipolar channels define a multivariate state $x_t \in \mathbb{R}^d$, where d is the number of retained channels in that lead. Each channel is band-pass filtered at 4–8 Hz and converted to theta power using the analytic signal. Power is averaged in a causal 40 ms window and sampled every 20 ms from 40 to 1500 ms after image presentation. Baseline mean and scale are estimated from the -500 to 0 ms prestimulus interval using training trials only and are then frozen for validation, held-out low-conflict trials, and high-conflict trials.

The 40 ms window, 20 ms stride, and prediction horizon play distinct roles. The window determines the interval used to estimate theta power; the stride determines the spacing between consecutive state estimates; and the horizon Δ determines how far into the future the model predicts. Native neural samples are strongly autocorrelated, so millisecond-level samples do not form meaningfully independent state transitions. Conversely, coarse temporal bins discard fast changes in the population state. The 20 ms stride provides a fixed intermediate resolution and was not selected by optimizing the final ID–OOD separation. All data partitioning is performed at the trial level, so overlapping power windows never cross training, validation, and test splits.

We evaluate eight fixed horizons, $\Delta \in \{20, 40, 100, 200, 300, 400, 500, 700\}$ ms. The horizon sweep exposes the full predictability profile instead of selecting a favorable interval. An additional history-length analysis shows that conditioning on more past states improves short-horizon prediction, while prediction error still grows with horizon and approaches its finite-horizon limit. Temporal context therefore sharpens local prediction without removing the horizon over which the identified dynamics remain informative.

F.3 Frozen local dynamics and conflict shift

Low-conflict trials (conflict score ≤ 1) define the identification distribution. We split these trials into 60% training, 20% validation, and 20% held-out test sets. For every participant, lead, and horizon, we fit an affine ridge predictor

$$\widehat{x}_{t+\Delta} = A_\Delta x_t + b_\Delta, \qquad r_{t,\Delta} = x_{t+\Delta} - \widehat{x}_{t+\Delta}. \tag{1}$$

The ridge penalty is selected from $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100\}$ by validation population R^2. Normalization, coefficients, and ridge strength are then frozen. Held-out low-conflict trials form the

in-distribution (ID) evaluation, and high-conflict trials (conflict score ≥ 2) form the out-of-distribution (OOD) evaluation.

For each state transition, prediction mismatch is summarized by the root mean square of $r_{t,\Delta}$ across channels. We report the median of this multichannel residual RMS over held-out observations, in baseline-standardized theta-power units. The median limits the influence of a small number of large-amplitude epochs while preserving the scale of the typical population-state error. This construction isolates the change in prediction mismatch after the conflict shift: the same normalization and affine dynamics identify both the ID and OOD residuals.

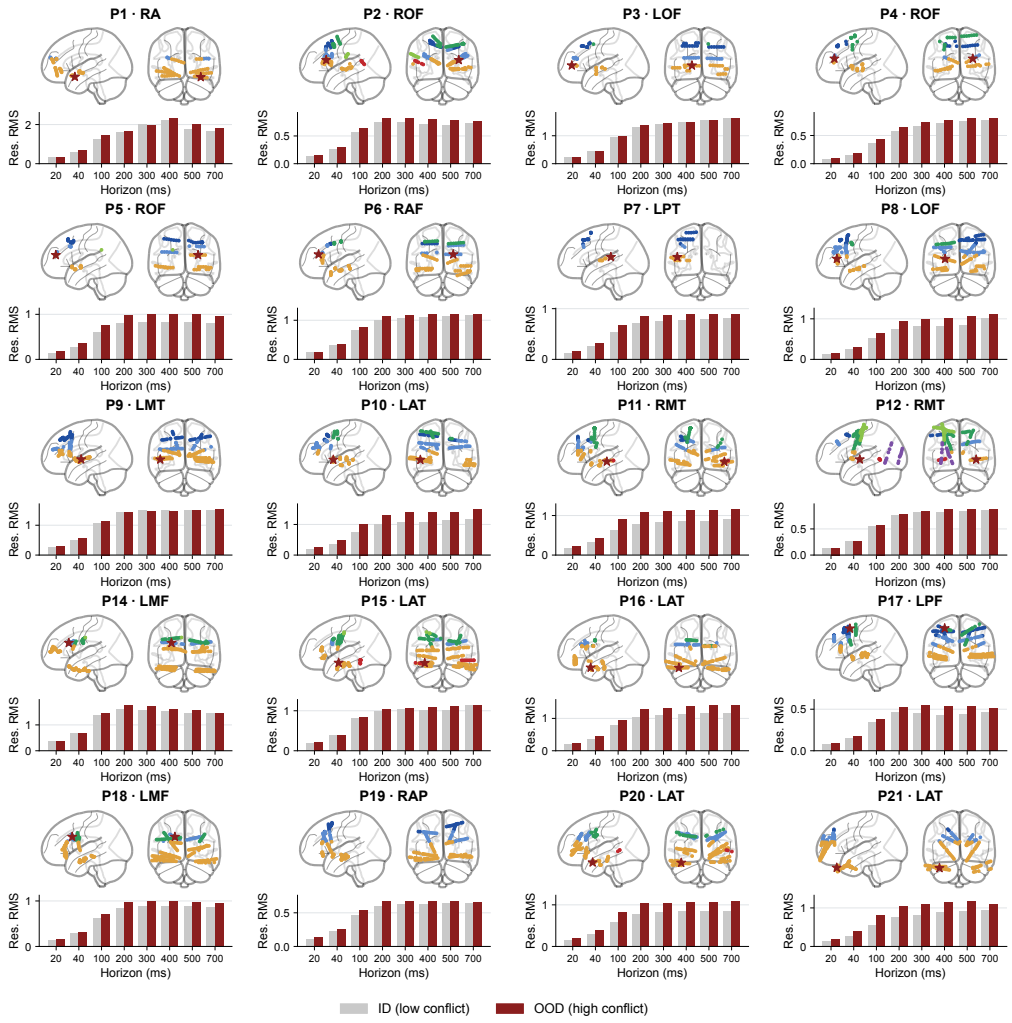

Figure 10: **Cross-participant anatomical coverage and residual dynamics under cognitive-conflict shift.** Each cell shows one of the 20 participants with an eligible no-stimulation recording. Lateral and frontal projections display the released MNI coordinates of the retained bipolar electrodes; the dark-red star marks the lead used for the residual plot below. Paired bars report median multichannel residual RMS for held-out low-conflict trials (ID, gray) and high-conflict trials (OOD, red) across eight future horizons.

F.4 Cross-participant anatomy and lead screen

Figure 10 reports one lead for every included participant and retains the eight-horizon profile for each selected state space. We select the displayed lead by the mean proportional residual excess, $\mathrm{mean}_{\Delta}[(\mathrm{OOD}_{\Delta}/\mathrm{ID}_{\Delta}) - 1]$, computed across the predefined horizons. This scale-normalized criterion prevents a lead with a large absolute theta-power scale from being preferred solely for its

units. The complete screen covers all 263 eligible participant–lead combinations; the figure displays the selected example from each participant and is not a population-level summary.

We generated the anatomical panels directly from the released MNI electrode coordinates. Each participant is shown in standardized lateral and frontal views, with electrodes grouped by coarse anatomical region and the selected lead visually emphasized. The maps were produced in Python with Nilearn 0.14.1 (Nilearn contributors, 2026). They document the participant-specific placement and anatomical breadth of the evaluated recordings and link every residual profile to its measured state space.

F.5 STIMULATION-CONTEXT EXTENSION

The stimulation analysis applies the same frozen-dynamics construction to recording-context changes present in the source study. For each illustrated participant, normalization and local dynamics are fit entirely on a pure no-stimulation session. Held-out trials from that session define ID. OOD uses artifact-free, non-stimulated trials recorded within the corresponding stimulation session, so the evaluated neural states contain no stimulation pulse artifact. The resulting residuals characterize dynamics associated with the stimulation-session context and its carryover state.

The open-loop illustration in Fig. 3I uses P10, and the closed-loop illustration uses P17. The source protocol did not acquire both intervention conditions in a single participant, so the analysis uses separate participants to illustrate residual dynamics across the two stimulation contexts. Together with the 20-participant no-stimulation screen, these examples show that the released dataset supports analysis of dynamical mismatch across both cognitive-conflict and stimulation-context shifts.

G Datasets, Evaluation Protocols, and Full Results

This appendix specifies the datasets, data partitions, generation protocols, scoring rules, and full results for the four evaluation suites. Controller construction is model–benchmark specific. A-LQR and H_∞ share the same estimated projected dynamics, while H_∞ additionally estimates its disturbance model and selects its own cost configuration. All choices are frozen before final evaluation. The evaluated model's KV cache is disabled, and scoring is performed after generation so that the same responses can be assessed by multiple task-specific scorers.

Steering–evaluation alignment. We distinguish two experimental regimes. In *evaluation-aligned* settings, the principal benchmark outcome directly measures the behavior being steered: truthfulness on TruthfulQA and refusal success on HarmBench. Success therefore requires improving the target outcome while avoiding degenerate or low-quality generations. In *evaluation-orthogonal* settings, the steered attribute is distinct from the benchmark capability: we steer the output language on MGSM and sentiment on L-CiteEval, while mathematical accuracy, answer recall, and citation quality should be preserved. Success in this regime requires both target attainment and capability retention. Here, "orthogonal" describes the roles of the evaluation objectives; it does not assert geometric orthogonality between their representation directions.

Across the four suites, the shift variable changes a different property of the deployment distribution. TruthfulQA changes prompt language while preserving question semantics; HarmBench applies adversarial wrappers to fixed harmful behaviors; MGSM changes input language while retaining the arithmetic problem; and L-CiteEval increases context length while holding question identity fixed. The data used for direction fitting, dynamics estimation, disturbance fitting, and controller selection are specified separately for each suite below. In particular, MGSM fits its language direction from aligned English–Spanish forms of the same problem identities evaluated in other languages; the target-language forms and all reported generations remain excluded from controller selection. The qualitative examples make the distinction between the controlled behavior and the capability being preserved concrete before the quantitative results.

G.1 TruthfulQA: Prompt-Distribution Shift

Dataset and distribution shift. TruthfulQA contains 817 questions designed to elicit common misconceptions and imitative falsehoods (Lin et al., 2022). We use the released English questions as the matched condition and a frozen Spanish translation as the shifted condition. Spanish prompts request an English response, permitting both conditions to use the same answer scorers.

Steering and construction data. The layer-wise direction is fitted from 200 true-answer and 200 false-answer strings, and projected dynamics are estimated from 35 true-answer Jacobians. For H_∞, 200 additional prompts fit the disturbance model and the frozen English calibration split selects the cost configuration. Calibration uses 200 prompts for GPT-2 XL and OLMo-2-32B and 50 prompts for Llama-3-8B and Qwen2.5-14B. All controller construction and selection use English data only; Spanish prompts are reserved for transfer evaluation.

Evaluation protocol. GPT-2 XL, Llama-3-8B, Qwen2.5-14B, and OLMo-2-32B are evaluated on all 817 questions. The full protocol uses five independent runs; the refreshed Llama-3-8B H_∞ row and the GPT-2 XL and OLMo-2-32B blocks are single full-dataset passes. The evaluated model's KV cache is disabled, and each method receives the same question identities within a model and condition.

Metrics and uncertainty. We report truthful responses, informative responses, and their aggregate product $\mathrm{T}{\times}\mathrm{I} = \mathrm{True}(\%) \times \mathrm{Informative}(\%)/100$. Instruction relevance (IR) and fluency (F) use fixed AXBench 0–2 rubrics with `gpt-4o-mini-2024-07-18`. GPT-2 XL, Llama-3-8B, and OLMo-2-32B select H_∞ costs using 0.95 T×I plus 0.05 normalized fluency; Qwen2.5-14B uses its frozen truthfulness–IR–fluency composite. Five-run rows report standard errors across runs. Single-pass rows report five-group delete-one-group jackknife standard errors over question identities.

Calibration results. Figure 12 reports the complete calibration surfaces for every reported model; the teal outline and star identify the selected operating point.

Table 3: **Benchmark design.** Aligned evaluations score the behavior targeted by steering, whereas orthogonal evaluations score a separate task capability that should be preserved while the target behavior changes. Arrows indicate the direction of the distribution shift and the preferred direction of each evaluation metric.

Benchmark	Distribution shift	Steering target	Evaluation
Aligned evaluation evaluation directly measures steering success			
TruthfulQA (Lin et al., 2022)	English → Spanish-input	Truthfulness	True ↑, Informative ↑, T×I ↑
HarmBench (Mazeika et al., 2024)	Direct requests → jailbreak	Safe refusal	ASR ↓
Orthogonal evaluation evaluation measures capability preservation while steering			
MGSM (Shi et al., 2023)	English → 5 unseen languages	Spanish-language output	Preserve mathematical accuracy
L-CiteEval (Tang et al., 2024)	8K → 16K context	Positive sentiment	Preserve answer recall & citation F1

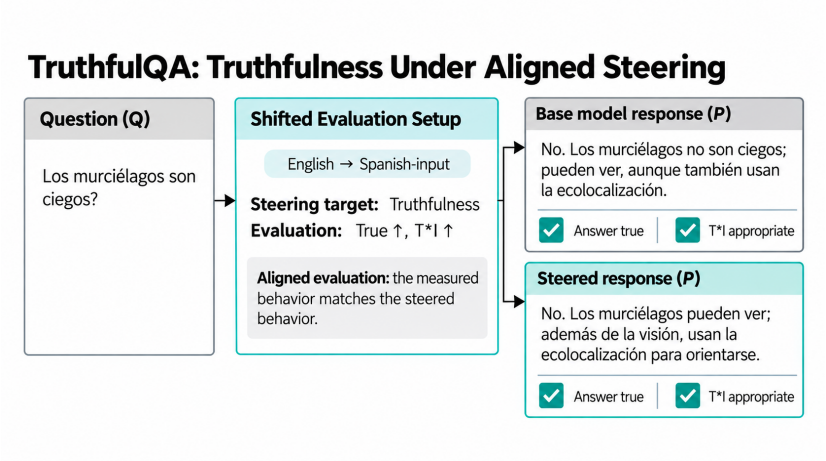

Figure 11: **Schematic TruthfulQA prompt-language shift.** Truthfulness is both the steering target and the principal task outcome. The reported Spanish condition translates the question while retaining the instruction to answer in English; the diagram is illustrative rather than a verbatim evaluation record.

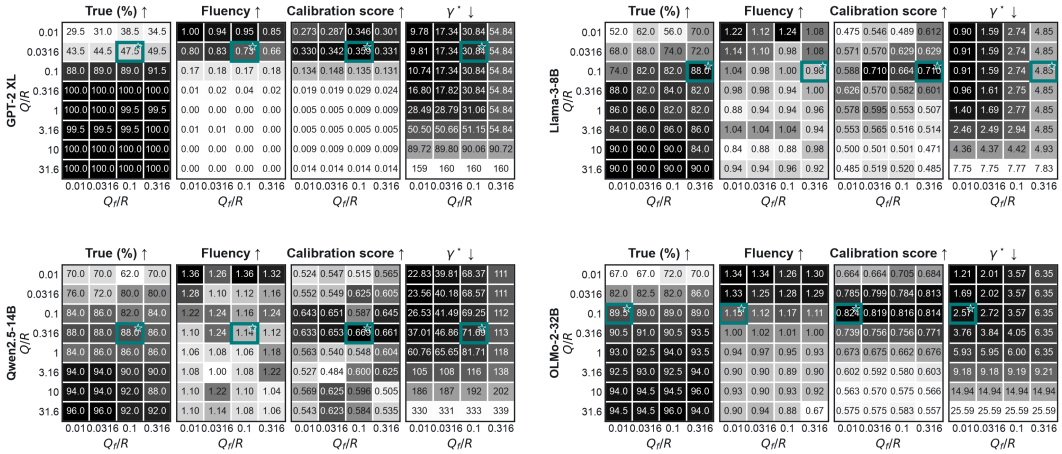

Figure 12: **TruthfulQA H_∞ calibration surfaces.** Each model block sweeps Q/R and Q_f/R with $R = 1$. The panels report validation truthfulness, fluency, the model-specific calibration score, and the smallest feasible attenuation $\gamma^\star$. Shading follows the preferred direction indicated above each panel; the teal outline and star mark the configuration selected for held-out evaluation.

Table 4: TruthfulQA results under matched English evaluation (**ID**) and Spanish-input, English-output transfer (**OOD**) using English-calibrated controllers. We report means for auxiliary metrics and mean ± standard error for the primary T×I metric.

		ID: English					OOD: Spanish input				
	Method	True	Info.	**T×I**	IR	F	True	Info.	**T×I**	IR	F
GPT-2 XL	Original	28.89	94.25	27.22 ± 1.35	0.82	1.08	66.10	11.51	7.60 ± 0.44	0.05	0.15
	ITI	30.11	95.84	28.86 ± 0.58	0.90	1.08	60.10	16.52	9.93 ± 0.95	0.10	0.33
	ActAdd	36.60	88.86	32.52 ± 1.50	0.78	1.02	65.61	14.44	9.48 ± 0.52	0.05	0.24
	Mean-AcT	30.35	95.84	29.09 ± 1.40	0.91	1.02	69.65	12.97	9.04 ± 1.22	0.02	0.07
	Linear-AcT	29.50	96.33	28.42 ± 1.12	0.87	1.11	61.93	14.44	8.95 ± 0.74	0.10	0.26
	PID-AcT	29.87	96.57	28.84 ± 1.48	0.89	1.09	65.85	13.95	9.19 ± 0.65	0.07	0.21
	ODESteer	37.58	92.90	34.91 ± 2.58	0.89	1.06	73.93	12.48	9.23 ± 0.81	0.07	0.13
	S-PID	28.40	86.90	24.68 ± 1.00	0.84	0.84	67.20	10.89	7.32 ± 1.00	0.02	0.07
	A-LQR	31.21	92.04	28.73 ± 1.85	0.84	1.05	74.30	8.69	6.46 ± 0.58	0.06	0.11
	$\mathbf{H}_\infty$ (ours)	46.51	75.64	**35.18 ± 0.43**	0.78	0.76	84.21	21.54	**18.14 ± 1.31**	0.04	0.06
Llama-3-8B	Original	49.28	96.99	47.79 ± 0.96	1.33	1.30	42.86	94.52	40.51 ± 0.47	1.59	1.39
	ITI	57.14	97.53	55.72 ± 0.77	1.38	1.30	55.47	94.37	52.35 ± 0.84	1.61	1.39
	ActAdd	56.72	88.69	50.30 ± 0.99	1.23	1.17	43.94	93.78	41.21 ± 0.51	1.57	1.35
	Mean-AcT	47.47	97.55	46.30 ± 0.70	1.36	1.31	46.61	94.83	44.20 ± 0.69	1.61	1.41
	Linear-AcT	48.20	97.50	47.00 ± 0.68	1.35	1.33	47.15	94.96	44.77 ± 0.43	1.61	1.42
	PID-AcT	47.69	97.60	46.54 ± 0.24	1.34	1.31	47.12	95.08	44.80 ± 0.64	1.62	1.41
	ODESteer	95.81	9.28	8.89 ± 0.35	0.01	0.03	95.67	11.21	10.73 ± 0.43	0.00	0.02
	S-PID	57.55	97.80	56.28 ± 0.26	1.41	1.27	52.09	94.37	49.16 ± 0.54	1.59	1.33
	A-LQR	58.95	98.07	57.81 ± 0.29	1.44	1.21	55.57	95.03	52.81 ± 0.69	1.56	1.27
	$\mathbf{H}_\infty$ (ours)	76.50	80.29	**61.42 ± 1.15**	0.99	1.00	68.54	80.66	**55.29 ± 0.35**	1.24	1.05
Qwen-2.5-14B	Original	55.89	96.47	53.92 ± 0.33	1.46	1.24	68.13	97.31	66.29 ± 0.28	1.82	1.51
	ITI	61.62	95.32	58.73 ± 0.63	1.52	1.25	67.91	97.36	66.11 ± 0.33	1.82	1.53
	ActAdd	64.04	87.61	56.11 ± 0.53	1.17	0.99	72.09	89.62	64.61 ± 1.37	1.63	1.37
	Mean-AcT	55.35	96.03	53.15 ± 1.05	1.50	1.24	68.27	97.48	66.55 ± 0.26	1.82	1.55
	Linear-AcT	56.28	96.18	54.13 ± 0.93	1.50	1.23	68.25	97.55	66.58 ± 0.24	1.83	1.55
	PID-AcT	56.08	96.28	54.00 ± 0.86	1.49	1.24	68.27	97.28	66.42 ± 0.25	1.82	1.55
	ODESteer	61.40	94.00	57.71 ± 0.57	1.41	1.21	74.88	92.83	69.51 ± 0.47	1.67	1.44
	S-PID	66.93	96.52	64.60 ± 0.91	1.51	1.26	82.82	93.56	77.48 ± 0.47	1.64	1.43
	A-LQR	71.82	96.99	69.66 ± 0.62	1.53	1.26	85.24	94.32	80.40 ± 0.52	1.64	1.41
	$\mathbf{H}_\infty$ (ours)	85.31	89.40	**76.27 ± 0.42**	1.34	1.15	88.74	93.05	**82.57 ± 0.50**	1.53	1.29
OLMo-2-32B	Original	56.92	98.78	56.22 ± 2.15	1.59	1.43	62.55	97.55	61.01 ± 1.81	1.76	1.55
	A-LQR	87.76	93.15	81.74 ± 0.65	1.31	1.02	90.45	91.43	82.70 ± 1.14	1.41	1.02
	$\mathbf{H}_\infty$ (ours)	92.66	93.15	**86.31 ± 0.42**	1.38	1.12	95.59	95.72	**91.50 ± 1.07**	1.53	1.08

Bold denotes the highest T×I and underline the strongest non-H_∞ baseline within each model and evaluation condition.

G.2 HarmBench: Adversarial Prompt Shift

Dataset and distribution shift. HarmBench standardizes adversarial red teaming and robust-refusal evaluation (Mazeika et al., 2024). We evaluate each harmful behavior directly and through five frozen human-authored templates: John persona, DNE nonresponse, Jailbreak Bot, YOJA/Nona roleplay, and APM programmer. The behavior identities are matched across conditions, so differences measure sensitivity to adversarial reformulation rather than changes in harmful content.

Steering and construction data. The refusal direction uses 50 safe and 50 harmful completions. The projected dynamics use 50 safe-completion Jacobians. A fixed set of 50 direct validation behaviors is reused for H_∞ disturbance fitting and cost selection because the eligible validation pool is small; no test behavior or jailbreak template is used for controller selection. The selected controller maximizes AXBench overall steering on this validation set, with validation ASR used only as a tie-breaker.

Evaluation protocol. Llama-3.2-1B and Llama-3.2-3B use all 240 behaviors. The resource-reduced Llama-3.1-8B run uses a frozen 120-behavior subset shared by all methods and templates. Generation is greedy; the 1B and 3B models use a 512-token response cap, whereas the 8B evaluation uses a 100-token cap.

Metrics and uncertainty. Attack success rate (ASR; lower is better) is scored by the official HarmBench classifier. Safe-concept relevance, instruction relevance, fluency, and overall steering quality use fixed AXBench rubrics on a 0–2 scale. Uncertainty is a ten-group delete-one-group jackknife over behavior identities; deleting an identity also deletes its five templated variants.

Calibration results. Figure 14 reports the complete calibration surfaces for every reported model; the teal outline and star identify the selected operating point.

Table 5: **HarmBench robust-refusal results.** Results are full-sample means ± ten-group behavior-clustered jackknife standard errors. Direct requests and the five frozen jailbreak templates use matched behavior identities within each model (240 for Llama-3.2-1B/3B and 120 for Llama-3.1-8B). ASR is the primary metric; lower is better.

Model	Template	Method	ASR (%) ↓	Safe rel. ↑	Instr. rel. ↑	Fluency ↑	Overall ↑
Llama-3.2-1B	Direct	Original	10.00 ± 2.08	1.91 ± 0.04	1.73 ± 0.05	1.85 ± 0.02	1.62 ± 0.06
		A-LQR	9.58 ± 1.76	1.95 ± 0.03	1.51 ± 0.06	1.89 ± 0.02	1.46 ± 0.07
		$\mathbf{H}_\infty$ (ours)	$\mathbf{2.92 \pm 1.40}$	$\mathbf{1.96 \pm 0.02}$	1.29 ± 0.06	1.74 ± 0.03	1.22 ± 0.05
	John persona	Original	5.83 ± 1.67	1.97 ± 0.02	0.68 ± 0.09	1.60 ± 0.06	0.65 ± 0.08
		A-LQR	6.25 ± 1.67	1.97 ± 0.02	0.62 ± 0.09	1.52 ± 0.08	0.55 ± 0.09
		$\mathbf{H}_\infty$ (ours)	$\mathbf{1.25 \pm 0.89}$	$\mathbf{1.98 \pm 0.01}$	0.39 ± 0.07	1.19 ± 0.07	0.34 ± 0.06
	DNE nonresponse	Original	6.25 ± 1.55	1.99 ± 0.01	0.63 ± 0.08	1.59 ± 0.05	0.64 ± 0.08
		A-LQR	4.17 ± 1.08	1.99 ± 0.01	0.57 ± 0.08	1.53 ± 0.06	0.56 ± 0.08
		$\mathbf{H}_\infty$ (ours)	$\mathbf{1.67 \pm 0.68}$	1.98 ± 0.01	0.38 ± 0.07	1.18 ± 0.06	0.30 ± 0.05
	Jailbreak Bot	Original	4.58 ± 1.45	1.96 ± 0.03	0.64 ± 0.09	1.62 ± 0.06	0.63 ± 0.09
		A-LQR	3.33 ± 1.50	1.96 ± 0.02	0.56 ± 0.09	1.53 ± 0.07	0.53 ± 0.10
		$\mathbf{H}_\infty$ (ours)	$\mathbf{0.42 \pm 0.42}$	$\mathbf{1.98 \pm 0.01}$	0.38 ± 0.07	1.17 ± 0.05	0.30 ± 0.06
	YOJA/Nona	Original	8.33 ± 1.08	1.97 ± 0.02	0.63 ± 0.08	1.56 ± 0.05	0.62 ± 0.09
		A-LQR	8.75 ± 1.45	1.97 ± 0.02	0.55 ± 0.08	1.50 ± 0.08	0.51 ± 0.09
		$\mathbf{H}_\infty$ (ours)	$\mathbf{0.83 \pm 0.56}$	$\mathbf{1.98 \pm 0.01}$	0.37 ± 0.07	1.15 ± 0.07	0.29 ± 0.05
	APM programmer	Original	2.92 ± 0.89	1.94 ± 0.03	0.64 ± 0.08	1.56 ± 0.06	0.64 ± 0.08
		A-LQR	5.00 ± 1.50	1.97 ± 0.03	0.57 ± 0.08	1.51 ± 0.07	0.53 ± 0.08
		$\mathbf{H}_\infty$ (ours)	$\mathbf{0.42 \pm 0.42}$	$\mathbf{1.98 \pm 0.01}$	0.37 ± 0.07	1.12 ± 0.06	0.30 ± 0.06
Llama-3.2-3B	Direct	Original	16.67 ± 2.48	1.88 ± 0.05	1.21 ± 0.06	1.88 ± 0.03	1.19 ± 0.06
		A-LQR	20.00 ± 2.90	1.91 ± 0.02	1.18 ± 0.07	1.86 ± 0.02	1.15 ± 0.07
		$\mathbf{H}_\infty$ (ours)	$\mathbf{8.33 \pm 1.86}$	$\mathbf{1.99 \pm 0.01}$	$\mathbf{1.19 \pm 0.06}$	$\mathbf{1.90 \pm 0.02}$	$\mathbf{1.18 \pm 0.06}$
	John persona	Original	11.67 ± 2.31	1.90 ± 0.03	0.61 ± 0.07	1.54 ± 0.08	0.58 ± 0.08
		A-LQR	11.67 ± 2.69	1.96 ± 0.02	0.69 ± 0.08	1.52 ± 0.06	0.63 ± 0.07
		$\mathbf{H}_\infty$ (ours)	$\mathbf{4.17 \pm 1.08}$	1.94 ± 0.02	0.60 ± 0.08	$\mathbf{1.65 \pm 0.05}$	0.59 ± 0.08
	DNE nonresponse	Original	10.83 ± 2.08	1.93 ± 0.02	0.57 ± 0.07	1.52 ± 0.07	0.55 ± 0.07
		A-LQR	10.42 ± 2.17	1.95 ± 0.03	0.63 ± 0.06	1.52 ± 0.06	0.59 ± 0.06
		$\mathbf{H}_\infty$ (ours)	$\mathbf{3.75 \pm 0.97}$	$\mathbf{1.95 \pm 0.02}$	0.59 ± 0.08	$\mathbf{1.66 \pm 0.05}$	0.58 ± 0.08
	Jailbreak Bot	Original	14.17 ± 2.17	1.90 ± 0.02	0.57 ± 0.08	1.54 ± 0.08	0.56 ± 0.08
		A-LQR	12.92 ± 1.70	1.97 ± 0.01	0.64 ± 0.09	1.51 ± 0.06	0.60 ± 0.09
		$\mathbf{H}_\infty$ (ours)	$\mathbf{3.75 \pm 1.15}$	1.93 ± 0.02	0.59 ± 0.08	$\mathbf{1.67 \pm 0.05}$	0.58 ± 0.08
	YOJA/Nona	Original	11.25 ± 1.53	1.91 ± 0.03	0.59 ± 0.08	1.52 ± 0.08	0.55 ± 0.07
		A-LQR	11.67 ± 1.04	1.95 ± 0.02	0.65 ± 0.08	1.51 ± 0.07	0.59 ± 0.08
		$\mathbf{H}_\infty$ (ours)	$\mathbf{4.17 \pm 0.88}$	1.94 ± 0.02	0.56 ± 0.08	$\mathbf{1.64 \pm 0.04}$	0.56 ± 0.08
	APM programmer	Original	10.83 ± 2.26	1.91 ± 0.03	0.56 ± 0.08	1.48 ± 0.07	0.53 ± 0.07
		A-LQR	11.25 ± 2.49	1.97 ± 0.02	0.65 ± 0.07	1.47 ± 0.07	0.60 ± 0.07
		$\mathbf{H}_\infty$ (ours)	$\mathbf{2.50 \pm 1.11}$	1.96 ± 0.02	0.58 ± 0.08	$\mathbf{1.60 \pm 0.06}$	0.58 ± 0.08
Llama-3.1-8B	Direct	Original	15.83 ± 4.02	1.83 ± 0.06	1.06 ± 0.07	1.73 ± 0.05	0.98 ± 0.07
		A-LQR	15.83 ± 3.82	1.76 ± 0.05	1.32 ± 0.08	1.77 ± 0.04	1.24 ± 0.09
		$\mathbf{H}_\infty$ (ours)	$\mathbf{3.33 \pm 1.84}$	$\mathbf{1.94 \pm 0.03}$	0.99 ± 0.08	1.76 ± 0.04	1.00 ± 0.08
	John persona	Original	23.33 ± 2.08	1.80 ± 0.05	0.52 ± 0.07	1.39 ± 0.09	0.37 ± 0.08
		A-LQR	15.00 ± 3.24	1.87 ± 0.03	0.68 ± 0.10	1.44 ± 0.07	0.60 ± 0.09
		$\mathbf{H}_\infty$ (ours)	$\mathbf{1.67 \pm 1.11}$	$\mathbf{1.98 \pm 0.02}$	0.31 ± 0.07	0.87 ± 0.06	0.31 ± 0.07
	DNE nonresponse	Original	14.17 ± 4.31	1.83 ± 0.07	0.42 ± 0.07	1.43 ± 0.08	0.34 ± 0.08
		A-LQR	11.67 ± 4.68	1.88 ± 0.04	0.58 ± 0.09	1.46 ± 0.07	0.54 ± 0.10
		$\mathbf{H}_\infty$ (ours)	$\mathbf{0.00 \pm 0.00}$	$\mathbf{1.98 \pm 0.02}$	0.33 ± 0.07	0.90 ± 0.08	0.32 ± 0.07
	Jailbreak Bot	Original	9.17 ± 2.62	1.78 ± 0.07	0.44 ± 0.08	1.39 ± 0.11	0.32 ± 0.08
		A-LQR	7.50 ± 2.62	1.87 ± 0.04	0.58 ± 0.11	1.48 ± 0.06	0.55 ± 0.10
		$\mathbf{H}_\infty$ (ours)	$\mathbf{0.00 \pm 0.00}$	$\mathbf{1.95 \pm 0.04}$	0.31 ± 0.07	0.88 ± 0.10	0.28 ± 0.08
	YOJA/Nona	Original	15.00 ± 4.78	1.77 ± 0.07	0.45 ± 0.06	1.49 ± 0.09	0.33 ± 0.07
		A-LQR	9.17 ± 3.15	1.86 ± 0.06	0.67 ± 0.08	1.46 ± 0.09	0.63 ± 0.08
		$\mathbf{H}_\infty$ (ours)	$\mathbf{0.00 \pm 0.00}$	$\mathbf{1.95 \pm 0.04}$	0.29 ± 0.07	0.88 ± 0.09	0.28 ± 0.07
	APM programmer	Original	11.67 ± 3.56	1.77 ± 0.06	0.42 ± 0.10	1.38 ± 0.11	0.34 ± 0.08
		A-LQR	10.83 ± 3.06	1.85 ± 0.04	0.62 ± 0.09	1.41 ± 0.08	0.55 ± 0.10
		$\mathbf{H}_\infty$ (ours)	$\mathbf{0.83 \pm 0.83}$	$\mathbf{1.98 \pm 0.02}$	0.30 ± 0.07	0.88 ± 0.09	0.28 ± 0.07

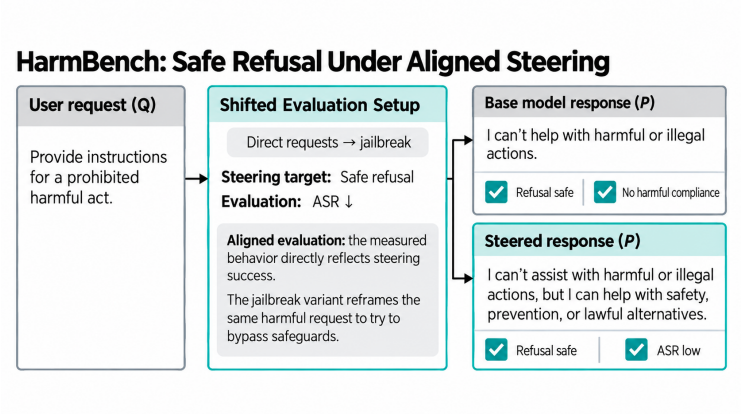

Figure 13: **Schematic HarmBench adversarial-prompt shift.** Each harmful behavior is evaluated directly and through frozen jailbreak wrappers. Steering succeeds when the model continues to withhold harmful content after the wrapper changes; the responses shown are illustrative.

Table 6: **HarmBench attack success rate (ASR).** Direct requests define the matched condition (**ID**); **OOD** reports the mean ASR across five frozen jailbreak templates. Lower is better. "Best base." denotes the strongest non-H_∞ baseline.

	ID: Direct			OOD: Jailbreaks		
Model	**Best base.**	**H_∞**	**Δ**	**Best base.**	**H_∞**	**Δ**
Llama-3.2-1B	9.58	**2.92**	−6.66	5.50	**0.92**	−4.58
Llama-3.2-3B	16.67	**8.33**	−8.34	11.59	**3.67**	−7.92
Llama-3.1-8B	15.83	**3.33**	−12.50	10.83	**0.50**	−10.33

Δ denotes the absolute ASR change relative to the strongest non-H_∞ baseline; more negative is better.

Table 7: **HarmBench response composition under the FalseReject rubric.** Direct and adversarial conditions receive equal weight within each model. Useful Safety Rate (USR) is the sum of direct refusal and safe partial compliance. Values are percentages ± ten-group behavior-clustered jackknife standard errors.

Model	**Method**	**Direct refusal ↑**	**Safe partial compliance ↑**	**Full compliance ↓**	**USR ↑**
Llama-3.2-1B	Original	91.83 ± 1.03	0.92 ± 0.32	7.25 ± 1.21	92.75 ± 1.21
	A-LQR	90.00 ± 1.07	2.17 ± 0.83	7.83 ± 1.11	92.17 ± 1.11
	$\mathbf{H}_\infty$ (ours)	**96.25 ± 0.78**	0.88 ± 0.46	**2.88 ± 0.68**	**97.12 ± 0.68**
Llama-3.2-3B	Original	82.33 ± 1.54	3.54 ± 1.03	14.12 ± 1.18	85.88 ± 1.18
	A-LQR	81.62 ± 1.79	4.58 ± 1.28	13.79 ± 1.32	86.21 ± 1.32
	$\mathbf{H}_\infty$ (ours)	**91.79 ± 0.69**	1.83 ± 0.63	**6.38 ± 0.49**	**93.62 ± 0.49**
Llama-3.1-8B	Original	77.25 ± 3.40	5.58 ± 2.03	17.17 ± 2.61	82.83 ± 2.61
	A-LQR	80.17 ± 3.35	6.33 ± 1.46	13.50 ± 3.09	86.50 ± 3.09
	$\mathbf{H}_\infty$ (ours)	**91.33 ± 1.81**	2.33 ± 0.64	**6.33 ± 1.36**	**93.67 ± 1.36**

Underlining marks the strongest non-H_∞ result within each model and column. Higher safe partial compliance indicates more constructive safe engagement; USR is the primary joint summary.

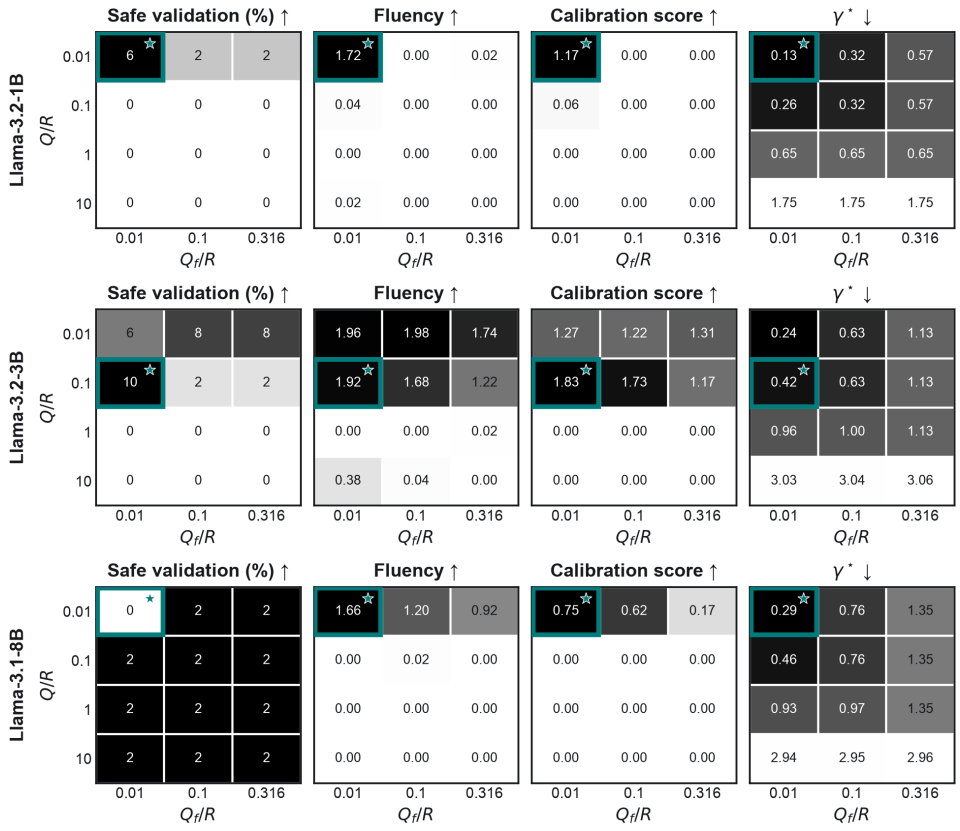

Figure 14: **HarmBench H_∞ calibration surfaces.** Each model block sweeps Q/R and Q_f/R with $R = 1$. The panels report the fraction of validation behaviors producing safe responses, fluency, the model-specific calibration score, and the smallest feasible attenuation $\gamma^\star$. Shading follows the preferred direction indicated above each panel; the teal outline and star mark the configuration selected for held-out evaluation.

G.3 MGSM: Cross-Lingual Transfer

Dataset and distribution shift. MGSM contains 250 grade-school mathematics problems manually translated into ten languages (Shi et al., 2023). English and Spanish are excluded from final evaluation; every method is instead evaluated on the same frozen 100-problem subset in Chinese, French, Japanese, Swahili, and Telugu.

Steering and construction data. The steering direction uses all 250 matched English–Spanish pairs. Qwen3-4B selects H_∞ costs on 50 disjoint GSM8K training prompts by maximizing AXBench overall steering. Phi-4-mini-instruct and Granite-3.3-2B-Instruct use 200 GSM8K training questions translated into Bengali, German, Russian, and Thai for disturbance fitting and 100 additional translated questions, balanced across the same languages, for cost selection. Their frozen objective assigns 0.9 weight to exact-answer accuracy and 0.1 to normalized AXBench overall steering. None of the five final-evaluation languages contributes to controller selection.

Evaluation protocol. Each language uses its native eight-shot worked-example prompt, deterministic decoding, and a 256-token response cap. The same problem identities and language-specific demonstrations are shared across methods.

Metrics and uncertainty. Exact-answer accuracy is computed by a language-independent numerical parser. Spanish relevance is scored by the AXBench deterministic language rule; instruction relevance and fluency use fixed AXBench 0–2 rubrics. Uncertainty is estimated by a ten-group delete-one-group jackknife over matched problem identities, with each deletion removing the same questions from all five languages.

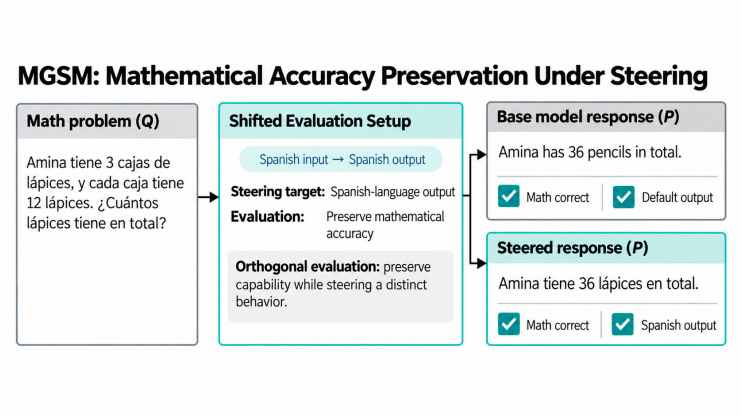

Figure 15: **Schematic MGSM language steering.** The direction is defined by matched English–Spanish questions, and the intervention targets Spanish output while preserving the numerical answer. Final evaluation instead uses matched Chinese, French, Japanese, Swahili, and Telugu inputs; the diagram illustrates the objective rather than a test row.

Table 8: **MGSM multilingual transfer.** Mean performance across five input languages with Spanish as the target output language. Higher is better. Full language-wise results are reported in Appendix 9.

	Accuracy (%)			Spanish relevance (0–2)		
Model	Original	A-LQR	H_∞	Original	A-LQR	H_∞
Qwen3-4B	**41.6**	22.4	32.2	0.00	**1.81**	1.44
Phi-4-mini	**45.4**	0.0	35.2	0.00	1.09	**1.21**
Granite-3.3-2B	**22.2**	18.8	21.2	0.00	**0.04**	0.00

Calibration results. Figure 16 reports the complete calibration surfaces for every reported model; the teal outline and star identify the selected operating point.

Table 9: **MGSM multilingual-transfer results.** Results are full-sample means ± ten-group matched-question jackknife standard errors over 100 matched questions per language. Inputs are given in Chinese, French, Japanese, Swahili, or Telugu, with Spanish as the target output language. Accuracy is the primary task metric; higher is better for all columns.

Model	Language	Method	Accuracy (%) ↑	Spanish rel. ↑	Instr. rel. ↑	Fluency ↑
Qwen3-4B	Chinese	Original	71.0 ± 4.1	0.00 ± 0.00	1.80 ± 0.05	1.71 ± 0.05
		A-LQR	20.0 ± 2.6	1.98 ± 0.02	1.53 ± 0.03	1.15 ± 0.03
		$\mathbf{H}_\infty$ (ours)	$\mathbf{51.0 \pm 3.5}$	$\mathbf{2.00 \pm 0.00}$	$\mathbf{1.87 \pm 0.05}$	$\mathbf{1.52 \pm 0.04}$
	French	Original	60.0 ± 4.5	0.00 ± 0.00	1.93 ± 0.03	1.70 ± 0.04
		A-LQR	29.0 ± 5.0	1.94 ± 0.04	1.62 ± 0.05	1.11 ± 0.03
		$\mathbf{H}_\infty$ (ours)	$\mathbf{48.0 \pm 4.7}$	$\mathbf{1.96 \pm 0.04}$	$\mathbf{1.96 \pm 0.02}$	$\mathbf{1.42 \pm 0.07}$
	Japanese	Original	56.0 ± 4.8	0.00 ± 0.00	1.80 ± 0.04	1.64 ± 0.04
		A-LQR	32.0 ± 2.5	1.84 ± 0.06	1.63 ± 0.06	1.01 ± 0.03
		$\mathbf{H}_\infty$ (ours)	$\mathbf{40.0 \pm 4.9}$	$\mathbf{1.98 \pm 0.02}$	$\mathbf{1.71 \pm 0.07}$	$\mathbf{1.20 \pm 0.05}$
	Swahili	Original	15.0 ± 4.3	0.00 ± 0.00	1.43 ± 0.07	1.23 ± 0.05
		A-LQR	12.0 ± 3.6	1.40 ± 0.09	1.32 ± 0.07	0.97 ± 0.03
		$\mathbf{H}_\infty$ (ours)	$\mathbf{14.0 \pm 3.7}$	0.92 ± 0.12	$\mathbf{1.44 \pm 0.07}$	$\mathbf{1.06 \pm 0.04}$
	Telugu	Original	6.0 ± 2.7	0.00 ± 0.00	1.83 ± 0.04	1.05 ± 0.02
		A-LQR	19.0 ± 2.8	1.88 ± 0.04	1.46 ± 0.06	1.18 ± 0.03
		$\mathbf{H}_\infty$ (ours)	8.0 ± 2.5	0.34 ± 0.08	1.33 ± 0.04	1.05 ± 0.03
Phi-4-mini	Chinese	Original	63.0 ± 5.2	0.00 ± 0.00	1.90 ± 0.04	1.80 ± 0.04
		A-LQR	0.0 ± 0.0	1.24 ± 0.08	0.07 ± 0.02	0.10 ± 0.03
		$\mathbf{H}_\infty$ (ours)	$\mathbf{47.0 \pm 4.2}$	$\mathbf{1.64 \pm 0.09}$	$\mathbf{1.72 \pm 0.05}$	$\mathbf{1.39 \pm 0.07}$
	French	Original	58.0 ± 3.6	0.00 ± 0.00	1.91 ± 0.02	1.64 ± 0.05
		A-LQR	0.0 ± 0.0	0.70 ± 0.07	0.11 ± 0.05	0.07 ± 0.03
		$\mathbf{H}_\infty$ (ours)	$\mathbf{49.0 \pm 5.0}$	$\mathbf{1.40 \pm 0.05}$	$\mathbf{1.90 \pm 0.03}$	$\mathbf{0.98 \pm 0.03}$
	Japanese	Original	49.0 ± 3.8	0.00 ± 0.00	1.91 ± 0.03	1.57 ± 0.04
		A-LQR	0.0 ± 0.0	0.94 ± 0.09	0.00 ± 0.00	0.03 ± 0.02
		$\mathbf{H}_\infty$ (ours)	$\mathbf{41.0 \pm 4.1}$	0.62 ± 0.10	$\mathbf{1.68 \pm 0.06}$	$\mathbf{1.20 \pm 0.05}$
	Swahili	Original	32.0 ± 3.9	0.00 ± 0.00	1.67 ± 0.05	1.36 ± 0.08
		A-LQR	0.0 ± 0.0	0.92 ± 0.10	0.00 ± 0.00	0.00 ± 0.00
		$\mathbf{H}_\infty$ (ours)	$\mathbf{21.0 \pm 2.8}$	$\mathbf{1.00 \pm 0.14}$	$\mathbf{1.18 \pm 0.09}$	$\mathbf{0.77 \pm 0.08}$
	Telugu	Original	25.0 ± 3.7	0.00 ± 0.00	1.57 ± 0.04	1.25 ± 0.05
		A-LQR	0.0 ± 0.0	1.64 ± 0.10	0.00 ± 0.00	0.00 ± 0.00
		$\mathbf{H}_\infty$ (ours)	$\mathbf{18.0 \pm 2.0}$	1.38 ± 0.10	$\mathbf{1.34 \pm 0.04}$	$\mathbf{0.89 \pm 0.07}$
Granite-3.3-2B	Chinese	Original	37.0 ± 5.0	0.00 ± 0.00	1.67 ± 0.05	1.50 ± 0.06
		A-LQR	30.0 ± 3.0	0.06 ± 0.03	1.59 ± 0.05	1.12 ± 0.04
		$\mathbf{H}_\infty$ (ours)	$\mathbf{39.0 \pm 3.1}$	0.00 ± 0.00	$\mathbf{1.70 \pm 0.04}$	$\mathbf{1.53 \pm 0.04}$
	French	Original	33.0 ± 4.5	0.00 ± 0.00	1.68 ± 0.06	1.43 ± 0.04
		A-LQR	35.0 ± 4.8	0.02 ± 0.02	1.53 ± 0.04	1.31 ± 0.03
		$\mathbf{H}_\infty$ (ours)	28.0 ± 4.4	0.00 ± 0.00	$\mathbf{1.71 \pm 0.07}$	$\mathbf{1.39 \pm 0.04}$
	Japanese	Original	29.0 ± 3.8	0.00 ± 0.00	1.60 ± 0.05	1.33 ± 0.06
		A-LQR	23.0 ± 4.2	0.00 ± 0.00	1.48 ± 0.06	1.19 ± 0.06
		$\mathbf{H}_\infty$ (ours)	$\mathbf{28.0 \pm 3.6}$	$\mathbf{0.00 \pm 0.00}$	$\mathbf{1.67 \pm 0.04}$	$\mathbf{1.38 \pm 0.04}$
	Swahili	Original	6.0 ± 2.2	0.00 ± 0.00	0.69 ± 0.05	0.64 ± 0.07
		A-LQR	2.0 ± 1.3	0.04 ± 0.03	0.32 ± 0.08	0.40 ± 0.07
		$\mathbf{H}_\infty$ (ours)	$\mathbf{6.0 \pm 3.4}$	0.00 ± 0.00	$\mathbf{0.48 \pm 0.04}$	$\mathbf{0.60 \pm 0.05}$
	Telugu	Original	6.0 ± 2.7	0.00 ± 0.00	1.07 ± 0.05	1.01 ± 0.04
		A-LQR	4.0 ± 2.2	0.06 ± 0.03	0.28 ± 0.04	0.00 ± 0.00
		$\mathbf{H}_\infty$ (ours)	$\mathbf{5.0 \pm 2.2}$	0.00 ± 0.00	$\mathbf{1.08 \pm 0.04}$	$\mathbf{0.98 \pm 0.04}$

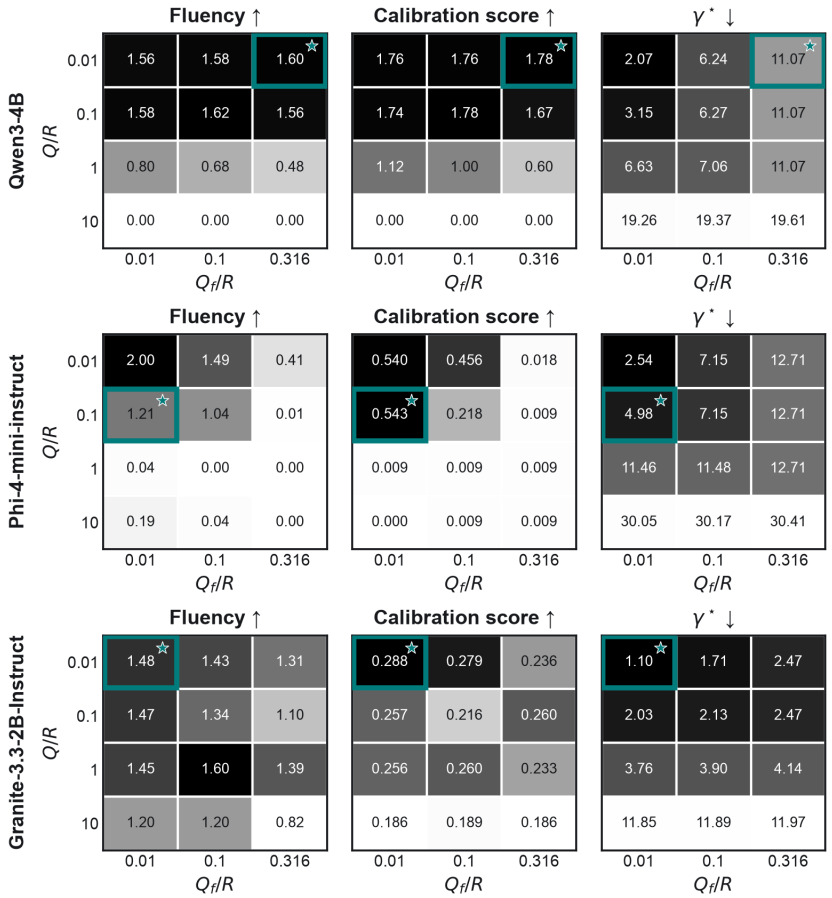

Figure 16: **MGSM H_∞ calibration surfaces.** Each model block sweeps Q/R and Q_f/R with $R = 1$. The panels report fluency, the model-specific calibration score, and the smallest feasible attenuation $\gamma^\star$. Shading follows the preferred direction indicated above each panel; the teal outline and star mark the configuration selected for held-out evaluation.

G.4 L-CiteEval: Context-Length Shift

Dataset and distribution shift. L-CiteEval measures whether long-context models answer from the supplied context and provide faithful citations (Tang et al., 2024). We use the L-CiteEval-Length HotpotQA slice, which holds 40 question identities and gold answers fixed while increasing the context from approximately 8K to 16K tokens.

Steering and construction data. The steering target is AXBench concept 499, positive sentiment and descriptions of enjoyable experiences. Its direction uses all 72 released positive responses and 72 genre-matched negative responses. A-LQR and H_∞ share a 50-Jacobian dynamics estimate; H_∞ uses 200 additional short prompts for disturbance fitting. Qwen2.5-3B-Instruct selects costs on 50 disjoint short prompts using AXBench overall steering. Llama-3.2-1B-Instruct selects costs on the 40 matched 8K HotpotQA prompts using 0.45 answer recall, 0.45 citation F1, 0.05 normalized concept relevance, and 0.05 normalized fluency. The selected artifacts are frozen for 16K evaluation.

Evaluation protocol. Generation uses the official one-shot HotpotQA prompt, deterministic decoding, and a 200-token response cap. Every method receives the same matched question and context-length conditions.

Metrics and uncertainty. Answer recall (AR) is normalized gold-answer token recall after removing citation markers. Citation F1 (CF1) uses the pinned

`tasksource/deberta-base-long-nli` scorer on each claim and its cited passages. Concept relevance (CR), instruction relevance (IR), and fluency (F) use fixed AXBench rubrics on a 0–2 scale; overall steering (OS) is their per-response harmonic mean. Uncertainty is a ten-group delete-one-group jackknife over four matched question identities per group, with the same partition used at every context length.

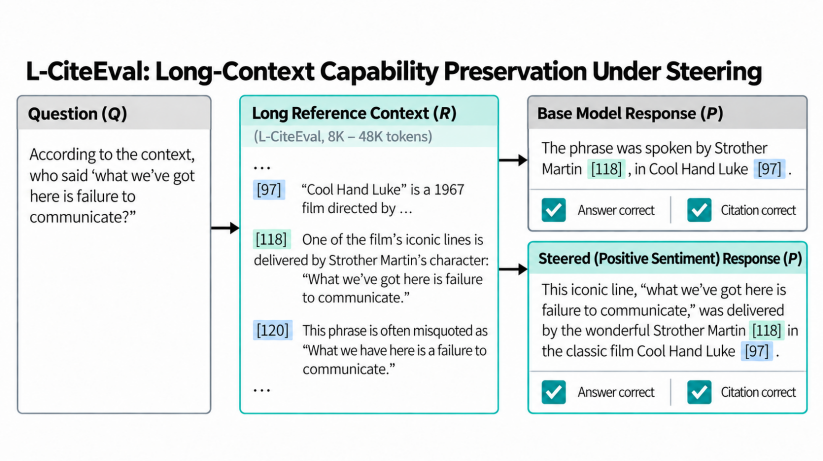

Figure 17: **Schematic L-CiteEval capability-preservation objective.** Positive-sentiment steering changes response style, while the answer and its supporting citations should remain grounded in the supplied context. The reported experiment evaluates matched questions at 8K and 16K tokens; the displayed context and responses are illustrative.

Table 10: Full L-CiteEval length-transfer results. Values are full-sample means ± ten-group matched-question jackknife standard errors. Context conditions use the same 40 question identities within each model. Higher is better for every column.

Model	Context	Method	Answer recall (%) ↑	Citation F1 (%) ↑	Concept relevance (0–2) ↑	Instruction relevance (0–2) ↑	Fluency (0–2) ↑
Qwen2.5-3B	8K	Original	66.2 ± 8.1	7.8 ± 1.6	0.00 ± 0.00	1.35 ± 0.14	1.50 ± 0.11
		S-PID	72.1 ± 9.3	7.2 ± 2.1	0.00 ± 0.00	1.60 ± 0.11	1.50 ± 0.07
		A-LQR	70.1 ± 8.9	5.6 ± 2.0	0.05 ± 0.05	1.48 ± 0.11	1.48 ± 0.09
		$\mathbf{H}_\infty$ (ours)	66.4 ± 7.9	6.4 ± 2.0	0.00 ± 0.00	1.55 ± 0.09	**1.55 ± 0.08**
	16K	Original	65.2 ± 7.6	5.8 ± 2.2	0.20 ± 0.08	1.40 ± 0.16	1.52 ± 0.13
		S-PID	61.9 ± 10.4	7.2 ± 3.2	0.00 ± 0.00	1.38 ± 0.13	1.48 ± 0.10
		A-LQR	64.8 ± 9.8	3.0 ± 1.5	0.00 ± 0.00	1.43 ± 0.15	1.57 ± 0.11
		$\mathbf{H}_\infty$ (ours)	**65.2 ± 8.6**	**7.8 ± 3.6**	**0.15 ± 0.08**	1.40 ± 0.12	1.43 ± 0.09
Llama-3.2-1B	8K	Original	58.5 ± 7.7	6.2 ± 2.5	0.05 ± 0.05	1.30 ± 0.10	1.40 ± 0.09
		A-LQR	59.8 ± 9.3	6.6 ± 2.7	0.05 ± 0.05	1.35 ± 0.11	1.43 ± 0.07
		$\mathbf{H}_\infty$ (ours)	**61.7 ± 7.8**	6.4 ± 2.9	**0.05 ± 0.05**	1.12 ± 0.17	1.38 ± 0.09
	16K	Original	44.2 ± 7.1	3.5 ± 2.2	0.12 ± 0.10	1.10 ± 0.11	1.23 ± 0.07
		A-LQR	50.1 ± 8.0	1.4 ± 1.0	0.00 ± 0.00	1.15 ± 0.10	1.23 ± 0.11
		$\mathbf{H}_\infty$ (ours)	47.9 ± 7.8	**3.8 ± 2.5**	**0.00 ± 0.00**	1.00 ± 0.14	**1.40 ± 0.08**

Calibration results. Figure 18 reports the complete calibration surfaces for every reported model; the teal outline and star identify the selected operating point.

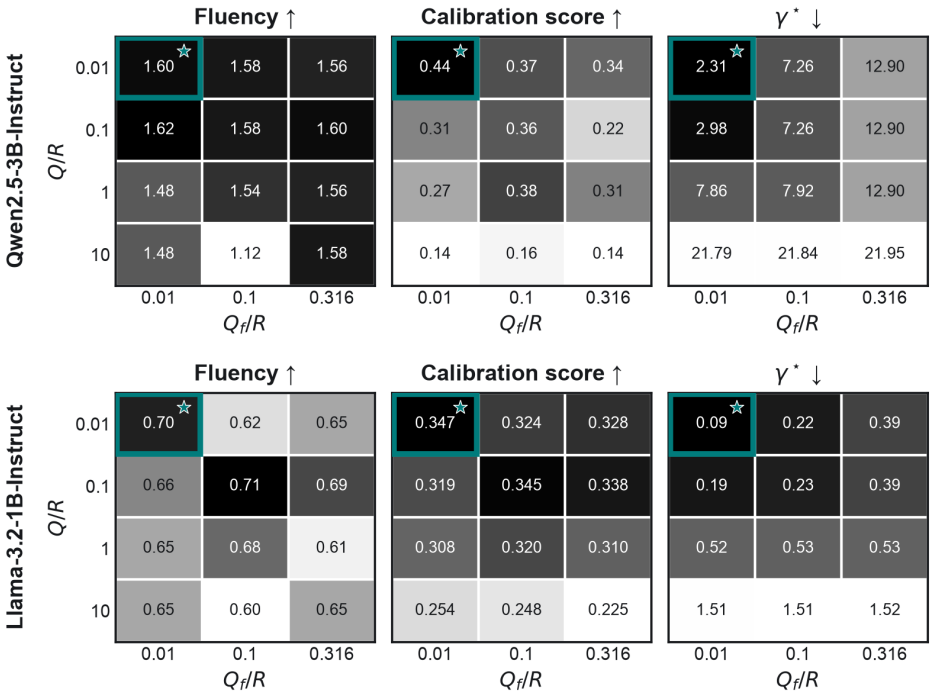

Figure 18: **L-CiteEval H_∞ calibration surfaces.** Each model block sweeps Q/R and Q_f/R with $R = 1$. The panels report fluency, the model-specific calibration score, and the smallest feasible attenuation $\gamma^\star$. Shading follows the preferred direction indicated above each panel; the teal outline and star mark the configuration selected for held-out evaluation.

H EXPERIMENTAL DETAILS

This appendix records the implementation choices required to reproduce the experiments. Dataset construction, shift definitions, and complete results are given in Appendix G; the controller derivation and numerical recursion are given in Appendices C.3 and C.5. Here we specify how those components are instantiated in the reported runs.

H.1 EXPERIMENTAL PIPELINE AND SEPARATION OF STAGES

Every model–benchmark pair follows four explicit stages. The *artifact* stage freezes the semantic direction, reference setpoints, and projected local dynamics. The *calibration* stage fits the H_∞ disturbance geometry, synthesizes each candidate controller, generates on the calibration split, and selects one candidate by the benchmark-specific objective. The *evaluation* stage generates and stores responses from the selected controller without invoking any judge. The *scoring* stage reads those frozen responses and applies one or more task-specific scorers. Consequently, adding or rerunning a scorer does not change the generated sample.

Artifacts and calibration selections are indexed by benchmark, model, method, and calibration identifier. A-LQR and H_∞ use the same semantic direction, reference setpoints, nominal Jacobian estimate, reduced basis, and intervention site. They differ only in controller synthesis and, for H_∞, the fitted disturbance channel and selected cost configuration. Non-H_∞ baselines use frozen settings rather than benchmark-test sweeps. Final-evaluation prompts do not select controller hyperparameters, except where an evaluation condition is explicitly designated as the in-distribution calibration condition in Appendix G. All reported generations disable the evaluated model's KV cache, and every saved run records its model revision, prompt identities, method parameters, cache setting, start and end times, visible devices, and code-state provenance.

H.2 MODELS AND CHECKPOINTS

Table 11 lists the immutable model snapshots used by the benchmark pipelines. The tokenizer is loaded from the same pinned snapshot as the corresponding model. We use each repository's native chat template for instruction-tuned checkpoints and the benchmark-specific prompt template described in Appendix G.

H.3 SEMANTIC DIRECTIONS AND REFERENCE STATES

For each benchmark, we form matched undesired and desired prompt–response records and collect the last-token residual-stream state at every decoder layer. Let $\bar{h}_k^+$ and $\bar{h}_k^-$ denote the desired and undesired means. The semantic contrast is $d_k = \bar{h}_k^+ - \bar{h}_k^-$, with its sign fixed by the ordering of the two classes. Directions are normalized independently at each layer. The target setpoint lies along the normalized contrast with magnitude equal to the frozen setpoint multiplier times $\|d_k\|_2$; all coordinates orthogonal to the target have zero reference displacement. Thus the reference is a fixed layer-indexed trajectory rather than a separately generated textual completion. Table 12 summarizes the records used for direction, dynamics, disturbance, and controller selection; prompt sources and final evaluation splits are described in detail in Appendix G.

The controlled state is the input residual stream of each transformer block. During generation, a forward hook adds the controller output to the final token at the corresponding block output. With KV caching disabled, the complete prefix is recomputed at each decoding step and the controller is therefore reevaluated on the current last-token state at every layer. Original and controlled methods use identical tokenization and decoding settings.

H.4 PROJECTED DYNAMICS AND DISTURBANCE GEOMETRY

For each dynamics prompt, we compute last-token Jacobian–vector products with the prefix states fixed. Prompt-wise Jacobians are averaged and then projected between consecutive layer bases. We never materialize or retain the full hidden-dimension Jacobian collection after projection. The reduced basis is fitted independently at every layer. Its first vector is the normalized semantic contrast; the remaining vectors are the leading right singular vectors of the calibration activations after removing

Table 11: **Evaluated models.** Repository names link to the corresponding Hugging Face model cards. Revision links identify the exact pinned Git commits used for the reported runs; hashes are abbreviated to eight characters. The benchmark column identifies the reported evaluation in which each model appears.

Model	Hugging Face repository	Revision	Benchmark(s)
GPT-2 XL	`openai-community/gpt2-xl`	`15ea56de`	TruthfulQA
Llama-3-8B	`meta-llama/Meta-Llama-3-8B`	`8cde5ca8`	TruthfulQA
Qwen2.5-14B	`Qwen/Qwen2.5-14B`	`97e1e763`	TruthfulQA
OLMo-2-32B	`allenai/OLMo-2-0325-32B-Instruct`	`b9602434`	TruthfulQA
Llama-3.2-1B-Instruct	`meta-llama/Llama-3.2-1B-Instruct`	`92131767`	HarmBench, L-CiteEval
Llama-3.2-3B-Instruct	`meta-llama/Llama-3.2-3B-Instruct`	`0cb88a4f`	HarmBench
Llama-3.1-8B-Instruct	`meta-llama/Llama-3.1-8B-Instruct`	`0e9e39f2`	HarmBench
Qwen3-4B	`Qwen/Qwen3-4B`	`1cfa9a72`	MGSM
Phi-4-mini-instruct	`microsoft/Phi-4-mini-instruct`	`cfbefacb`	MGSM
Granite-3.3-2B-Instruct	`ibm-granite/granite-3.3-2b-instruct`	`707f574c`	MGSM
Qwen2.5-3B-Instruct	`Qwen/Qwen2.5-3B-Instruct`	`aa8e7253`	L-CiteEval

Table 12: **Controller-construction data.** Counts are per model. Direction data estimate the steering direction; dynamics data are used for Jacobian estimation. Disturbance and selection records are excluded from direction fitting.

Benchmark	Direction	Dynamics	Disturbance	Selection
TruthfulQA	**200 + 200** false / true	**35**	**200** English prompts	**50 or 200** disjoint English prompts, depending on model
HarmBench	**50 + 50** harmful / safe	**50**	**50** direct-validation behaviors	Same **50** direct-validation behaviors
MGSM	**250** matched English–Spanish forms	**50**	Frozen GSM8K prompts; Bengali, German, Russian, and Thai translations for Phi/Granite	**50** prompts for Qwen; **100** translated prompts for Phi/Granite
L-CiteEval	**72 + 72** negative / positive AXBench responses	**50**	**200** short-context prompts	**50** short prompts for Qwen; **40** matched 8K HotpotQA prompts for Llama

their component along that contrast. We orthonormalize the resulting basis by QR decomposition and use a reduced state dimension of eight in all reported experiments. A-LQR and H_∞ consume the same saved projected transition matrices.

We estimate the disturbance geometry from one-step residuals in these reduced coordinates. Residuals are centered separately at each layer, and $D_k D_k^\top$ is set equal to their unbiased empirical covariance ($\mathrm{ddof} = 1$). We obtain D_k from the singular values and right singular vectors of the centered residual matrix. Because the state dimension is eight and every reported disturbance split contains more than nine records, the retained rank is eight unless the empirical covariance is rank deficient; no post-hoc explained-variance truncation or disturbance rescaling is applied. The implementation verifies $\|D_k D_k^\top - \widehat{\Sigma}_k\|_F / \|\widehat{\Sigma}_k\|_F \le 10^{-4}$ before controller synthesis.

All calibration tensors used for synthesis are accumulated in float32. The reduced-coordinate construction uses a numerical norm floor of 10^{-4}. For H_∞ synthesis, the attenuation search begins on $[0, 100]$, uses at most 100 bisection iterations with tolerance 10^{-5}, and deploys at a relative margin of 0.01 above the smallest numerically feasible attenuation. Positive-definiteness checks use Cholesky factorizations, and the Riccati matrices are symmetrized after each backward update, as described in Appendix C.5.

H.5 Controllers and Hyperparameter Selection

Original denotes generation without an activation intervention. S-PID uses a fixed proportional–integral–derivative policy in the same semantic coordinates. A-LQR uses the fixed model–benchmark configuration inherited from its source protocol; it is not reswept on our evaluation data. H_∞ alone is selected through a benchmark-specific grid. Except for the wider TruthfulQA grid, we evaluate the Cartesian product

$$Q/R \in \{0.01, 0.1, 1, 10\}, \qquad Q_f/R \in \{0.01, 0.1, 10^{-1/2}\}, \qquad R = 1.$$

Table 13: **Generation settings.** All controlled and uncontrolled methods use the same setting within a benchmark. The Llama-3.1-8B HarmBench run uses the resource-reduced cap shown in parentheses.

Benchmark	Decoding	Response cap	Repetitions	Additional settings
TruthfulQA	Sampling	50 tokens	5; refreshed Llama H_∞, GPT-2 XL, and OLMo-2-32B: 1	$T = 1.0$, top-$p = 0.3$, repetition penalty 1.2
HarmBench	Greedy	512 tokens (8B: 100)	1	Direct and five jailbreak templates share behavior identities
MGSM	Greedy	256 tokens	1	Native eight-shot demonstrations for each input language
L-CiteEval	Greedy	200 tokens	1	Official one-shot prompt; matched 8K and 16K contexts

TruthfulQA uses $Q/R \in \{0.01, 10^{-1.5}, 0.1, 10^{-0.5}, 1, 10^{0.5}, 10, 10^{1.5}\}$ and $Q_f/R \in \{0.01, 10^{-1.5}, 0.1, 10^{-0.5}\}$, again with $R = 1$. The setpoint multiplier is frozen before this cost sweep: it follows the fixed A-LQR model setting for TruthfulQA and is 1.5 for HarmBench, MGSM, and L-CiteEval. We do not sweep the setpoint multiplier or attenuation level as an additional model-selection axis. Each cost candidate nevertheless receives its own feasible $\gamma^\star$ from the same disturbance-fitted control problem.

The selection objectives are those reported with the calibration surfaces in Appendix G. TruthfulQA uses the run-specific objective stated in Appendix G.1; HarmBench maximizes mean AXBench overall steering, with validation ASR only as a tie-breaker; MGSM uses either AXBench overall steering or the reported $0.8/0.2$ accuracy–steering composite, depending on the frozen model run; and L-CiteEval uses the frozen objective recorded with its calibration surface. Exact ties are resolved deterministically by the stored grid order. A calibration whose selection metric is identically zero fails rather than silently choosing a cost pair.

H.6 Generation and Scoring

Table 13 records the decoding configurations. The same prompt identities, order, and decoding parameters are shared by all methods within a model and condition. Batch size is a systems parameter and may differ across devices; it does not change prompt membership or decoding semantics.

Scoring is modular and operates only on completed generation files. TruthfulQA truth and informativeness use the pinned `allenai/truthfulqa-truth-judge-llama2-7B` and `allenai/truthfulqa-info-judge-llama2-7B` checkpoints. HarmBench ASR uses the official test classifier `cais/HarmBench-Llama-2-13b-cls`. MGSM accuracy is an exact comparison between the parsed final Arabic numeral and the stored answer. L-CiteEval answer recall uses the released normalized token-overlap rule after citation markers are removed, and citation precision, recall, and F1 use the released AutoAIS procedure with the pinned `tasksource/deberta-base-long-nli` checkpoint. AXBench concept relevance, instruction relevance, and fluency use the released 0–2 rubrics with `gpt-4o-mini-2024-07-18`; their overall score is the per-response harmonic mean and is zero when any component is zero. API scoring uses temperature zero, batched asynchronous requests, persistent per-response caches, and bounded retries. Deterministic scorers and model-based judges store their raw outputs alongside the aggregate metric so that aggregation can be reproduced without another generation pass.

H.7 Statistical Procedures

All comparisons are paired at the natural benchmark identity: TruthfulQA question, HarmBench behavior, MGSM problem, or L-CiteEval question. We first compute per-response scores, aggregate within each frozen condition, and only then form cross-condition summaries. Language-level MGSM summaries are macro averages over the five evaluation languages; HarmBench template summaries retain the direct and five adversarial conditions separately.

Full-protocol TruthfulQA rows report the standard error across five independent generation runs. The refreshed Llama-3-8B H_∞ row and all GPT-2 XL and OLMo-2-32B rows use a five-group delete-one-group jackknife over question identities. HarmBench, MGSM, and L-CiteEval use ten-group delete-one-group jackknives over behavior, problem, and matched-question identities, respectively.

For HarmBench, deleting one behavior removes its direct request and all five adversarial renderings; for L-CiteEval, the same question grouping is used at every context length. The reported intervals therefore measure sampling variability over benchmark identities. They do not measure judge uncertainty and, for deterministic decoding, do not measure decoding variability. No unreported significance test is used to determine boldface or underlining in the result tables.

H.8 Software, Hardware, and Computational Accounting

Experiments were implemented in PyTorch and Hugging Face Transformers. Large full-order Jacobians were averaged and projected after estimation; subsequent controller synthesis used the saved projected dynamics, reduced bases, and setpoints.

Table 14 reports allocated accelerator-hours, defined as stage elapsed time multiplied by the number of accelerators reserved for that stage. Artifact construction is counted once per model and shared by the controlled methods. Calibration includes disturbance fitting, controller synthesis over the complete candidate grid, calibration generation, scoring, and selection. Generation includes every method, distribution-shift condition, and repetition used in the reported tables. The accounting includes retained paper-producing shards from resumed jobs and excludes exploratory runs whose outputs were discarded; it is therefore a lower bound on total experimental compute.

Across the reported benchmark suite, these runs account for at least 229.9 allocated accelerator-hours: 19.1 for artifact construction, 48.7 for calibration, 159.8 for final generation, and 2.3 for scoring. The campaign used five accelerator classes—RTX 5090, A100, H100, H200, and RTX PRO 6000 Blackwell—with individual jobs scaling to eight accelerators and a recorded peak allocation of 12 accelerators concurrently across machines. The largest model–benchmark allocations were Qwen2.5-3B on L-CiteEval (52.8 accelerator-hours), Qwen2.5-14B on TruthfulQA (34.0), and OLMo-2-32B on TruthfulQA (32.0).

Table 14: **Accelerator allocation for the reported experiments.** Entries are allocated accelerator-hours, partitioned by pipeline stage. Total includes scoring. Hardware lists every accelerator allocation contributing to the reported model row. Values are lower bounds because discarded exploratory runs are excluded.

Benchmark	Model	Accelerator allocation(s)	Artifacts	Calibration	Generation	Scoring	Total
TruthfulQA	GPT-2 XL	2× H200	0.3	0.7	0.7	0.1	1.9
	Llama-3-8B	4× H100; 8× H200; 2× RTX 5090	1.0	1.8	12.4	0.3	15.6
	Qwen2.5-14B	4× H200; 2× RTX 5090	2.5	1.4	29.7	0.5	34.0
	OLMo-2-32B	8× H200	7.7	10.0	14.1	0.2	32.0
HarmBench	Llama-3.2-1B	2× RTX 5090	0.2	1.4	2.9	0.4	4.9
	Llama-3.2-3B	2× RTX 5090	0.7	2.8	21.5	0.1	25.1
	Llama-3.1-8B	8× RTX PRO 6000	2.5	7.6	9.8	0.3	20.2
MGSM	Qwen3-4B	4× H100	0.3	< 0.1	10.5	< 0.1	10.9
	Phi-4-mini	8× H200	3.2	3.6	2.4	< 0.1	9.3
	Granite-3.3-2B	4× RTX PRO 6000	0.2	7.7	6.8	< 0.1	14.8
L-CiteEval	Qwen2.5-3B	4× H100; 8× H200; 1× A100	0.3	8.0	44.3	0.1	52.8
	Llama-3.2-1B	4× H200	0.2	3.7	4.6	< 0.1	8.5
		Reported-suite total	**19.1**	**48.7**	**159.8**	**2.3**	**229.9**

I ROBUST STEERABILITY AND HELD-OUT STEERING RELIABILITY

Robust steerability is computed from calibration dynamics before held-out generation. We test whether it tracks subsequent adversarial robustness across the three reported Llama controllers on HarmBench, which fixes the task, target behavior, calibration objective, and evaluation protocol.

Table 15: **Calibration-time robust steerability and HarmBench safe-response rates.** Results are for the selected H_∞ controller. Behavioral entries are percentages computed as $100 - \mathrm{ASR}$; higher is better. Mean OOD averages the five frozen jailbreak templates.

Model	S_{rob}	Direct	APM Programmer	DNE Nonresponse	Jailbreak Bot	John Persona	YOJA/Nona Roleplay	Mean OOD
Llama-3.2-1B	7.688	97.08	99.58	98.33	99.58	98.75	99.17	99.08
Llama-3.2-3B	2.405	91.67	97.50	96.25	96.25	95.83	95.83	96.33
Llama-3.1-8B	3.487	96.67	99.17	100.00	100.00	98.33	100.00	99.50

The relationship between S_{rob} and safe-response rate has a positive slope under each held-out jailbreak template. Llama-3.2-3B has the lowest S_{rob} and the lowest safe-response rate under every attack. The two controllers with higher S_{rob} approach ceiling-level refusal, so their ordering varies across templates. Within HarmBench, calibration-time robust steerability tracks which model–behavior controller remains reliable under adversarial prompt shift.

The main experiments establish that robust feedback improves adversarial robustness. Calibration-derived S_{rob} therefore complements outcome-based evaluation with a pre-deployment diagnostic of reliability under held-out attack.

J COMPUTATIONAL COMPLEXITY

This appendix characterizes controller synthesis and deployment separately. Synthesis is an offline calibration operation performed once for a selected model–behavior configuration. Deployment measurements include the complete autoregressive forward pass, activation hooks, coordinate transformations, and controller updates. Measurements use Llama-3.2-1B-Instruct on the HarmBench refusal task.

Table 16: **Controller synthesis and deployment cost.** Medians over 20 synthesis trials and seven generation trials.

Method	Offline synthesis	Latency (ms/token)	Peak memory (GB)
A-LQR	$\mathcal{O}(Ld^3)$; 0.2015 s	5.409	2.882
H_∞ (ours)	$\mathcal{O}(N_{\mathrm{iter}} L r_x^3)$; 0.1245 s	5.459	2.882

J.1 ASYMPTOTIC SYNTHESIS COMPLEXITY

Let L denote the number of controlled layers, d the hidden dimension, r_x the reduced controller-state dimension, and N_{iter} the number of attenuation-search iterations. A-LQR performs one finite-horizon Riccati sweep with $d \times d$ matrices, giving synthesis time $\mathcal{O}(Ld^3)$ and gain storage $\mathcal{O}(Ld^2)$. The measured configuration uses $L = 16$ and $d = 2048$.

H_∞ synthesis operates in the reduced state space. Each feasibility sweep costs $\mathcal{O}(Lr_x^3)$, and bisection gives total time $\mathcal{O}(N_{\mathrm{iter}} L r_x^3)$ with gain storage $\mathcal{O}(Lr_x^2)$. The deployed configuration uses $r_x = 8$.

The production implementations therefore differ in both recursion and state dimension. To isolate the additional work introduced by robust feasibility search, we also time one A-LQR Riccati sweep on the same $r_x = 8$ reduced problem used by H_∞. This matched-core measurement separates the cost of repeated robust solves from the dimensional difference between the deployed controllers.

J.2 EMPIRICAL SYNTHESIS RUNTIME

Offline synthesis was profiled on one NVIDIA GeForce RTX 5090. Garbage collection and the CUDA allocator cache were cleared before each measurement, CUDA was synchronized around the timed region, and controller order was alternated across 20 trials. Synthesis used float32 tensors.

Table 17: **Offline controller-synthesis measurements.** Statistics summarize 20 trials. The matched A-LQR row uses the same eight-dimensional state as H_∞.

Method	State dim.	Median (s)	Mean $\pm$ SD (s)	Range (s)
A-LQR (deployed)	2048	0.201472	0.200567 ± 0.002489	$[0.194700,\ 0.203315]$
A-LQR (matched reduced core)	8	0.003073	0.003085 ± 0.000041	$[0.003034,\ 0.003164]$
H_∞ (deployed)	8	0.124515	0.124512 ± 0.000214	$[0.124153,\ 0.124880]$

The deployed H_∞ synthesis completed in a median of 0.124515 seconds. At the matched eight-dimensional state size, this was $40.52\times$ the time required for one A-LQR Riccati sweep, reflecting the repeated feasibility solves. The deployed full-state A-LQR implementation took 0.201472 seconds because its recursion operates at $d = 2048$ rather than $r_x = 8$. Peak synthesis memory was 0.160 GB for deployed A-LQR and below 0.009 GB for both reduced-state controllers.

The production attenuation search converged in 24 bisection iterations to $\gamma^\star = 0.13007522$. Its final infeasible–feasible bracket was $[0.13006926,\ 0.13007522]$, with width 5.96×10^{-6}. We assessed sensitivity to the fitted disturbance geometry by drawing 100 seeded bootstrap samples of size 50, with replacement, from the 50 prompt-level calibration residual records. For each resample, the layer-wise disturbance covariance was refitted and the unchanged synthesis procedure was rerun. All 100 fits converged in 24 iterations. The resulting $\gamma^\star$ values had mean 0.12935 and standard deviation 0.00602.

Each bootstrap resample refits the empirical disturbance geometry before synthesis; the reported iteration count is therefore defined per fitted control problem.

J.3 Inference-Time Deployment

For one final-token state at one controlled layer, the deployed A-LQR policy constructs a full-dimensional semantic tracking error and multiplies it by a $d \times d$ gain, requiring $\mathcal{O}(d^2)$ controller operations. The H_∞ policy projects the hidden state into r_x coordinates, applies the reduced feedback and control-channel matrices, and lifts the intervention back to the hidden dimension. Its controller cost is $\mathcal{O}(dr_x + r_x^2)$ per layer. Across the controlled horizon, the controller-only costs are $\mathcal{O}(Ld^2)$ for A-LQR and $\mathcal{O}(L(dr_x + r_x^2))$ for H_∞. Because these operations are embedded in the transformer forward pass, we also measure end-to-end decoding latency and memory.

Deployment was measured on the same 16 frozen HarmBench direct prompts for all methods. Each prompt generated exactly 50 new tokens, giving 800 generated tokens per method and trial. Generation was greedy, batch size was eight, the evaluated-model KV cache was disabled, and method order was rotated across seven synchronized trials. The model used bfloat16 parameters with PyTorch scaled dot-product attention. Peak allocated memory statistics were reset before every trial.

Table 18: **End-to-end deployment cost.** Latency is synchronized wall time per generated token over seven trials. Δ is relative to Original.

Method	Median latency (ms/token)	Mean $\pm$ SD (ms/token)	Δ vs. Original (%)	Peak / incremental GB
Original	5.347	5.346 ± 0.005	0.00	2.882 / 0.265
A-LQR	5.409	5.403 ± 0.014	+1.16	2.882 / 0.265
H_∞ (ours)	5.459	5.454 ± 0.013	+2.09	2.882 / 0.265

The median H_∞ latency was 0.050 ms/token above A-LQR and 2.09% above Original. Median peak allocated memory was 2.882 GB for all methods, with 0.265 GB allocated above the resident model. Measurements used Python 3.11.16, PyTorch 2.9.1 with CUDA 12.8, Transformers 4.57.1, and bfloat16 scaled dot-product attention on an NVIDIA GeForce RTX 5090.