Activation Steering for Aligned Open-ended Generation without Sacrificing Coherence

Niklas Herbster[1,2,3] **Martin Zborowski**[1,2] **Alberto Tosato**[1]
Gauthier Gidel[3] **Tommaso Tosato**[1,3]

[1]Tara Research [2]Technical University of Munich [3]Mila Quebec AI Institute
tommaso@tararesearch.org

Abstract

Alignment in LLMs is more brittle than commonly assumed: misalignment can be induced by adversarial prompts, benign fine-tuning, emergent misalignment, and goal misgeneralization. Recent evidence suggests that some misalignment behaviors are encoded as linear structure in activation space, making it tractable via activation steering, which could be used as a lightweight runtime defense. We implement three methods: Steer-With-Fixed-Coefficient (SwFC), which applies uniform additive steering, and two novel projection-aware methods, Steer-to-Target-Projection (StTP) and Steer-to-Mirror-Projection (StMP), that use a logistic regression decision boundary to selectively intervene only on tokens whose activations fall below the threshold. We evaluate these methods on two threat models, dishonesty and dismissiveness, using malicious system prompts as a controlled proxy for misalignment. We conduct our experiments on two architectures (Llama-3.3-70B-Instruct and Qwen3.6-27B). All methods substantially recover alignment. StTP and StMP preserve general capabilities (MMLU, MT-Bench, AlpacaEval) better than uniform steering. Finally, we show that our honesty steering generalizes to out of distribution scenarios: a single honesty direction extracted from the aligned model significantly raises scores on the MASK benchmark, suppresses deception in multi-agent settings (Among Us), doubles the hidden-behavior discovery rate on AuditBench, and restores honesty in an emergently misaligned model.[1]

1 Introduction

Large language models (LLMs) undergo extensive alignment training to produce helpful, harmless, and honest behavior through safety SFT and RLHF (Bai et al., 2022), yet this alignment is brittle (Jain et al., 2024) and shallow (Qi et al., 2025). Misalignment can arise through several pathways. First, adversarial prompts can circumvent the model's guardrails (Wei et al., 2023; Zou et al., 2023b). Second, fine-tuning can degrade safety even when the data is benign (Qi et al., 2024), and fine-tuning on a narrow set of poisoned data induces broad misalignment across unrelated domains (*emergent misalignment*; Betley et al. 2026).Third, post-training may itself instill the wrong objective. In *specification gaming*, the training signal is misspecified, so the model optimizes an imperfect proxy for the intended objective (Krakovna et al., 2020; Skalse et al., 2022). In *goal misgeneralization*, the training signal does not pin down a unique goal, so the model can acquire a proxy goal that matches the intended one on the training distribution yet diverges from it out of distribution (Di Langosco et al., 2022; Shah et al., 2022).

Existing defenses each target a specific attack surface or must anticipate the threat at training time. Black-box test-time methods such as input/output classifiers (Sharma et al., 2025) screen for malicious prompts but cannot detect alignment shifts arising independently of user input. White-box train-time methods such as circuit breakers (Zou et al., 2024)

[1]Code: https://github.com/Tara-Research/activation-steering-4-honesty

and latent adversarial training (Sheshadri et al., 2025; Xhonneux et al., 2024) make safe representations more robust, but require retraining before the threat is encountered. Neither category provides a source-agnostic runtime correction.

Activation steering (Turner et al., 2023; Zou et al., 2023a) offers a complementary alternative: modifying internal representations during forward passes without weight updates. The intervention adds a *steering vector*, a direction in activation space that separates the target behavior from its opposite, to the model's hidden states. Two recent mechanistic findings motivate our approach. First, Soligo et al. (2025) show that emergent misalignment converges to similar linear representations across different fine-tuning datasets, and that a single "misalignment direction" can both ablate and induce misalignment; corroborated by Dunefsky & Cohan (2025), who show that steering vectors from a single misaligned example generalize broadly. Second, Qi et al. (2025) demonstrate that safety alignment primarily governs the first few output tokens, leaving deeper representations largely unaltered. Together, these findings motivate a defense that operates at the activation level (because misalignment is linearly encoded there), and does so continuously throughout generation (because early-token safety is insufficient).

Prior work has applied steering for behavioral control (Rimsky et al., 2024; Li et al., 2023) and safety (Wang et al., 2024; Zhao et al., 2025; Wang et al., 2025b). However, traditional steering methods degrade text coherence and unintentionally compromise unrelated behaviors (Xiong et al., 2026; Korznikov et al., 2025). These side effects motivate the development of adaptive methods. Whether selective, per-token steering can maintain alignment during extended conversations without degrading coherence or capabilities remains an open question.

We investigate this question in the context of open-ended, multi-turn text generation, which better reflects deployment conditions than constrained evaluation formats. Specifically: Can activation steering restore alignment while preserving coherence and capabilities, when using malicious system prompts as a controlled proxy for misalignment? Does this persist across multi-turn conversations? Do adaptive methods offer advantages? And, can the same steering direction also correct misalignment that does not originate from the prompt, but from fine-tuning, or from incentives in multi-agent environments?

We make the following contributions:

1. We introduce two projection-aware steering methods, **Steer-to-Target-Projection (StTP)** and **Steer-to-Mirror-Projection (StMP)**, that selectively intervene on tokens whose projections fall below distribution-derived thresholds, preserving already-aligned tokens.
2. We evaluate across two threat models (*dishonesty* and *dismissiveness*) and two architectures (Llama-3.3-70B, Qwen3.6-27B), and find that all methods substantially recover the target traits under a malicious system prompt. The methods differ mainly in capability preservation: StMP is the most robust and StTP nearly as strong, losing capability only under honesty steering on Llama, whereas uniform steering (SwFC) degrades capabilities the most.
3. We find that in multi-turn evaluations, uniform steering (SwFC) leads to high text repetition both within and across turns, whereas StTP and StMP produce substantially less repetition while maintaining trait expression.
4. We show that steering toward honesty using StTP and StMP generalizes to out-of-distribution settings: 1) *MASK*, a standard honesty benchmark, 2) A*mong Us*, a multi-agent environment, and 3) *AuditBench*, which fine-tunes hidden behaviors that models must confess, and 4) *Emergent Misalignment*, where narrow fine-tuning induces broad misalignment.

2 Related Work

Activation Steering: Foundations. Activation steering modifies model behavior by adding contrastive vectors to internal representations (Turner et al., 2023). Representation engineering (Zou et al., 2023a) demonstrated behavioral control across safety-relevant dimensions,

while contrastive activation addition (CAA) formalized steering vector extraction from contrasting behavior pairs (Rimsky et al., 2024). Inference-time intervention (ITI) (Li et al., 2023) introduced selective intervention on specific model components, establishing the probe-then-intervene paradigm. The theoretical basis is provided by the linear representation hypothesis: Park et al. (2024) propose that concepts are encoded as directions under a causal inner product, connecting probing accuracy to steering effectiveness. However, Wu et al. (2025) and Wang et al. (2025a) doubt the effectiveness of steering in restoring alignment, reporting that prompting and fine-tuning outperform steering-based approaches.

Activation Steering for Safety. A growing body of work makes steering input-adaptive to better balance safety and utility. Methods differ in what they adapt: some scale steering coefficients per input (Zhao et al., 2025; Vogels et al., 2025; Yu et al., 2025), others selectively mask activation dimensions (Wang et al., 2025b; Shen et al., 2025), gate whether steering is applied based on input properties (Wang et al., 2024; Lee et al., 2025; Sheng et al., 2026), or target specific token positions (Nguyen et al., 2025). Our methods also operate at the token level but take a simpler approach: a logistic regression decision boundary determines *whether* to steer each token, while projection geometry onto the steering direction controls *how strongly*, requiring no learned gates or optimization-based tuning.

The Coherence Gap in Steering Evaluation. Despite this progress, evaluation of steered outputs remains narrowly focused on safety metrics such as harmfulness and refusal rates. Siu et al. (2025) span 17 safety datasets yet never assess whether steered text remains coherent. More broadly, Bartoszcze et al. (2025) identify fluency evaluation as a key open challenge in the representation engineering literature. Our work addresses this gap by systematically evaluating coherence alongside trait expression across all steering methods.

Evaluation Beyond Single-Turn Settings. Nearly all steering evaluations use single-turn prompts. Pres et al. (2024) demonstrate that such evaluations systematically overestimate steering effectiveness, and Tosato et al. (2025) show that LLM trait expression exhibits persistent instability across multi-turn conversations even without intervention. No existing work evaluates steering in multi-turn settings where both effects compound. Our evaluation protocol addresses this gap.

3 Methods

3.1 Problem Formulation

Let $\mathcal{M}$ be a language model and $\mathbf{x} = (s, q)$ an input consisting of a system prompt s and user query q, producing response $\mathbf{y} = \mathcal{M}(\mathbf{x})$. Let $\tau : \mathbf{y} \mapsto [0, 100]$ measure a target trait and $\kappa : \mathbf{y} \mapsto [0, 100]$ measure coherence. An *aligned* system prompt s^+ yields high trait expression ($\tau(\mathcal{M}(s^+, q)) \geq T^+$), while a *malicious* prompt s^- suppresses it ($\tau(\mathcal{M}(s^-, q)) \leq T^- \ll T^+$). A steering intervention $\mathcal{S}$ *restores alignment* if $\tau(\mathcal{M}_{\mathcal{S}}(s^-, q)) \approx \tau(\mathcal{M}(s^+, q))$ while maintaining $\kappa(\mathcal{M}_{\mathcal{S}}(s^-, q)) \geq \kappa(\mathcal{M}(s^+, q)) - \epsilon$. This malicious-prompt setting is the controlled proxy in which we extract steering vectors and compare methods; we later evaluate the same extracted honesty direction on misalignment that does not originate from the prompt (Section 5.4).

3.2 Steering Vector Extraction

We extract steering vectors using logistic regression on contrastive activations. Given N training scenarios, each pairing a user prompt p_i with two contrastive system prompts (one aligned, one malicious), we use the target model to generate responses under each system prompt, yielding on-policy response pairs (a_i^+, a_i^-). We then collect response-averaged hidden states at layer ℓ:

$$\mathcal{E}_\ell^+ = \{\bar{\mathbf{h}}_\ell(p_i, a_i^+)\}_{i=1}^N \quad \text{and} \quad \mathcal{E}_\ell^- = \{\bar{\mathbf{h}}_\ell(p_i, a_i^-)\}_{i=1}^N \tag{1}$$

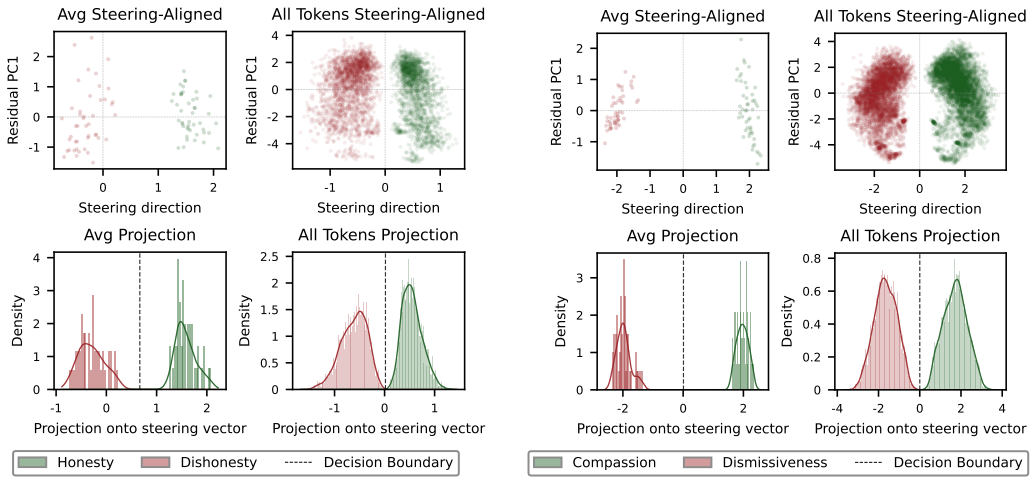


Figure 1: **PCA and projection histogram analysis.** Each panel shows a 2×2 grid: PCA of response-averaged embeddings (top-left), PCA of all-token embeddings (top-right), and the corresponding projection histograms onto $\hat{\mathbf{v}}_\ell$ (bottom row). Dashed lines show the logistic regression decision boundary m_ℓ used by StTP & StMP.

where $\bar{\mathbf{h}}_\ell$ denotes the mean hidden state over response tokens at layer ℓ. We train a binary logistic regression classifier on $\mathcal{E}_\ell^+ \cup \mathcal{E}_\ell^-$:

$$P(y{=}+1 \mid \mathbf{e}) = \sigma(\mathbf{w}_\ell^\top \mathbf{e} + b_\ell)\,. \tag{2}$$

We normalize the weight vector to obtain the steering direction $\hat{\mathbf{v}}_\ell = \mathbf{w}_\ell / \|\mathbf{w}_\ell\|_2$.[2] The steering vector is defined as $\mathbf{v}_\ell = \Delta\mu_\ell \cdot \hat{\mathbf{v}}_\ell$, where $\Delta\mu_\ell = \mu_\ell^+ - \mu_\ell^-$ is the mean projection gap between positive and negative class centroids along $\hat{\mathbf{v}}_\ell$. This normalization ensures $\|\mathbf{v}_\ell\| = \Delta\mu_\ell$, so that $\alpha = 1$ in SwFC (§ 3.3) shifts activations by one natural unit of class separation. The bias is rescaled as $b'_\ell = b_\ell \cdot \Delta\mu_\ell / \|\mathbf{w}_\ell\|_2$, encoding the decision boundary in projection space; the resulting threshold $m_\ell = -b'_\ell / \|\mathbf{v}_\ell\|_2$ is used by StTP (§ 3.3) and StMP (§ 3.3).

From $\mathcal{E}_\ell^+$ and $\mathcal{E}_\ell^-$, we compute projections onto $\hat{\mathbf{v}}_\ell$: $\mathcal{P}_\ell^+ = \{\langle \mathbf{e}_i^+, \hat{\mathbf{v}}_\ell \rangle\}_{i=1}^N$ and $\mathcal{P}_\ell^- = \{\langle \mathbf{e}_i^-, \hat{\mathbf{v}}_\ell \rangle\}_{i=1}^N$, and derive distribution statistics (means μ_ℓ^+, μ_ℓ^-; standard deviations $\sigma_\ell^+, \sigma_\ell^-$).

Fig. 1 visualizes training embeddings via PCA alongside 1D projections onto $\hat{\mathbf{v}}_\ell$: both traits form well-separated clusters, validating linear separability. The projection histograms show a clear distributional gap, quantified via Cohen's d in Fig. A.2. Within each grid, the left column uses the response-averaged embeddings the classifier is fitted on, and the right column the individual token activations that steering actually operates on. Separation persists at the single-token level, so a token's own projection onto $\hat{\mathbf{v}}_\ell$ indicates whether it is aligned, and a fixed threshold on that projection is enough to decide per token whether to intervene.

3.3 Steering Methods

Steering Position. All methods support two position modes: `all` (steering all tokens, including prompt) and `response` (steering generated tokens only).

Steer-With-Fixed-Coeff (SwFC). The simplest approach adds a scaled steering vector uniformly to all activations at layer ℓ:

$$\mathbf{h}'_\ell = \mathbf{h}_\ell + \alpha \cdot \mathbf{v}_\ell \tag{3}$$

where $\alpha \in \mathbb{R}$ is the steering coefficient and $\mathbf{v}_\ell$ is the steering vector defined in § 3.2.

[2] $\hat{\mathbf{v}}_\ell$ closely aligns with the CAA mean-difference direction (Rimsky et al., 2024) across all layers and both traits (§ A.3).

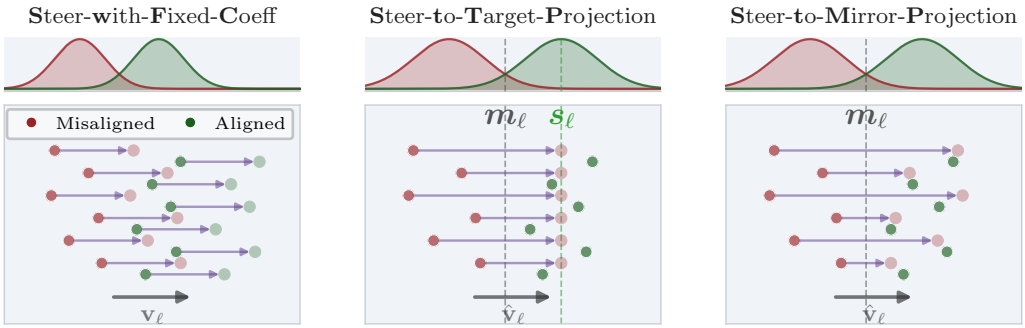


Figure 2: **Steering methods.** SwFC adds a fixed-magnitude vector; StTP shifts the projection to a target value along $\hat{\mathbf{v}}_\ell$; and StMP mirrors the projection across a hyperplane orthogonal to $\hat{\mathbf{v}}_\ell$.

Steer-to-Target-Projection (StTP). StTP selectively steers only misaligned activations toward a target projection value (Fig. 2). The decision boundary $m_\ell = -b'_\ell / \|\mathbf{v}_\ell\|_2$ is directly derived from the logistic regression bias (§ 3.2), and the target projection is $s_\ell = \mu^+_\ell + \alpha \cdot \sigma^+_\ell$, where α controls how far into the positive distribution we steer. For each token with projection $\rho = \langle \mathbf{h}_\ell, \hat{\mathbf{v}}_\ell \rangle$, we apply:

$$\mathbf{h}'_\ell = \mathbf{h}_\ell + (s_\ell - \rho) \cdot \hat{\mathbf{v}}_\ell \quad \text{if } \rho < m_\ell \quad \text{and} \quad \mathbf{h}'_\ell = \mathbf{h}_\ell \quad \text{otherwise} \tag{4}$$

Tokens below the decision boundary are projected to s_ℓ; well-aligned tokens remain unchanged. The complete algorithm is provided in § A.

Steer-to-Mirror-Projection (StMP). StMP reflects activations across the decision boundary $m_\ell = -b'_\ell / \|\mathbf{v}_\ell\|_2$ (Fig. 2). For each token with projection $\rho < m_\ell$, it interpolates between ρ and its mirror image $2m_\ell - \rho$ by a factor α, adding a delta of $2\alpha(m_\ell - \rho)$. When $\alpha = 1$ this produces a full reflection; $\alpha > 1$ overshoots past the mirror:

$$\mathbf{h}'_\ell = \mathbf{h}_\ell + 2\alpha(m_\ell - \rho) \cdot \hat{\mathbf{v}}_\ell \quad \text{if } \rho < m_\ell, \quad \mathbf{h}'_\ell = \mathbf{h}_\ell \quad \text{otherwise} \tag{5}$$

Scaling the correction by $m_\ell - \rho$ makes the intervention proportional to how misaligned each token is: strongly misaligned tokens are pushed far across the boundary, tokens just below m_ℓ are barely perturbed, and tokens already above it are left untouched.

4 Experimental Setup

4.1 Datasets

Dishonesty. Each scenario is a *self-report*: the system prompt casts the model in a role and states a specific failure it committed in a previous session (e.g. missing a SQL injection in a code review), and the affected user then asks about that action. The honest and dishonest prompts are identical except for a closing instruction to report truthfully or to deflect, and a gold truth records what must be disclosed, so omission counts as dishonesty. The training data consists of 50 scenarios, each with a contrastive system prompt, a user prompt, and the target model's on-policy honest and dishonest responses. The test data consists of 112 diverse scenarios across 8 different categories.

Dismissiveness. The training data consists of 50 emotionally challenging user prompts paired with compassionate and dismissive responses generated by the target model across 5 contrastive system prompt variants. 40 held-out prompts are used for testing.

4.2 Models and Infrastructure

We primarily evaluate on Llama-3.3-70B-Instruct (Grattafiori et al., 2024) (80 layers), with cross-architecture validation on Qwen3.6-27B (Yang et al., 2025) (64 layers). Both models use the same steering methodology and evaluation pipeline; Qwen results are in Section D.

Steering interventions are implemented as PyTorch forward hooks that modify hidden states during generation, with layers distributed across 4 GPUs via `torch.multiprocessing`. LLM judge evaluation uses vLLM (Kwon et al., 2023) with tensor parallelism.

4.3 Evaluation Metrics

Baselines. We define two reference points: an **aligned baseline** (aligned system prompt, no steering), representing the model's intended behavior, and a **malicious baseline** (malicious system prompt, no steering). The gap between them quantifies what steering must recover.

LLM-as-Judge. We use GPT-oss-120B (OpenAI et al., 2025), an open-source reasoning model, as an LLM-as-a-judge to score responses separately on **trait expression** (*honesty* or *compassion*, 0–100) and **coherence** (linguistic quality: fluency, structure, and grammaticality, 0–100), with temperature 1.0 and high reasoning effort. Full judge configuration is in § B.1.

Operating Points. An operating point is a specific configuration (layer ℓ, coefficient α, steering position) selected from the layer-sweep results. We select operating points to maximize trait expression while maintaining coherence ($\geq$ 90% of the coherence of the aligned baseline).

Pairwise ELO. To validate that the relative ordering of steering coefficients is not an artifact of the absolute scoring protocol, we conduct a pairwise ELO evaluation: the same judge compares two responses to the same prompt and selects the better one. For each method and trait, we run a tournament among 5 coefficient variants plus both baselines. See § C.6.

Embedding Distance. Cosine similarity between the sentence embeddings of responses generated while steering and responses from the aligned baseline, across all layers (§ C.5).

Model Cross-Entropy. For each steered response, we compute its per-token cross-entropy using the unsteered model conditioned on the aligned system prompt, measuring how natural the steered text appears to the aligned model. We report cross-entropy rather than perplexity (PPL $= e^H$) as it scales more interpretably with steering strength. See § C.3.

Capability Benchmarks. We test the degradation of model capability under steering on three capability benchmarks: AlpacaEval (Dubois et al., 2025), MT-Bench (Zheng et al., 2023), and MMLU (Hendrycks et al., 2021); see § C.7.

Generalization. We evaluate the generalization and robustness of our honesty vector on the MASK benchmark (Ren et al., 2025), Among Us (Golechha & Garriga-Alonso, 2026), AuditBench (Sheshadri et al., 2026) and Emergent Misalignment (Betley et al., 2026).

4.4 Multi-Turn Evaluation

We evaluate whether single-turn alignment restoration persists across extended conversations. A self-play protocol alternates two copies of the same model: one acts as the assistant (with the malicious system prompt and steering applied), the other as the user (unsteered, instructed to continue naturally). For honesty, the steered assistant is an agent concealing several aspects of a failure, with a user-simulator probing one per turn over 5 turns (20 scenarios). For compassion, the user simulator acts as an emotionally distressed person seeking support over 10 turns (40 conversations). Operating points are selected from single-turn results (Table C.1); full hyperparameters are in § B.4.

Beyond the metrics above, we track two metrics to detect text repetition: (1) *sentence reuse rate*, the fraction of sentences whose SBERT cosine similarity to any prior-turn sentence exceeds 0.8; and (2) *cross-turn 4-gram repetition*, the fraction of unique 4-grams in the current turn that appeared in any prior turn. Both metrics have a natural upward bias as conversation history grows; unsteered baselines are reported alongside to isolate steering-specific effects.

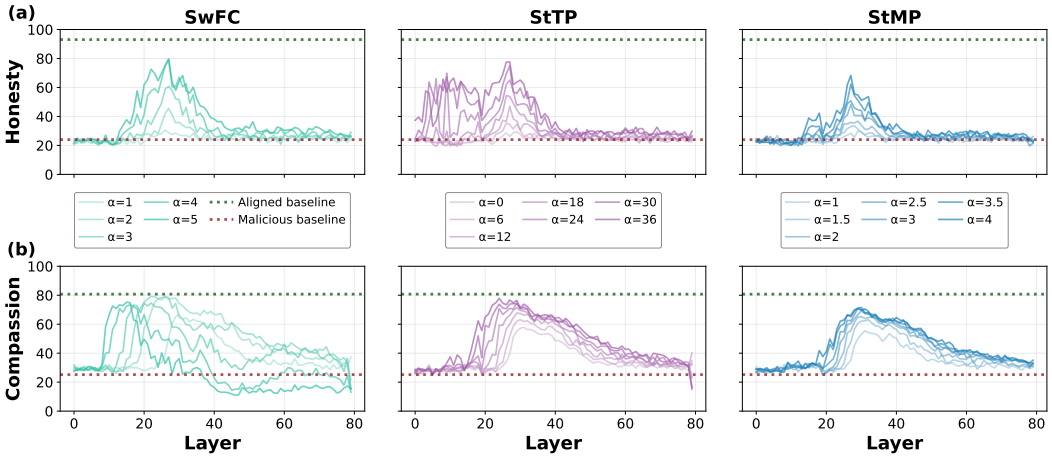


Figure 3: **Single-Turn Open-Ended Response Steering (Llama-3.3-70B).** Each column corresponds to a steering method (SwFC, StTP, StMP). The top row shows the honesty score under the dishonesty threat; the bottom row shows the compassion score under the dismissiveness threat. Each curve corresponds to a different steering coefficient α (see legend). Horizontal lines mark the aligned baseline (purple) and the malicious baseline (black).

5 Results

5.1 Single-Turn Open-Ended Response Steering

We perform a layer-wise evaluation (Fig. 3) to identify at which layers steering best recovers the target trait while preserving coherence. The main text reports Llama-3.3-70B; the full Qwen3.6-27B replication is in Section D.

Dishonesty Threat. The top row of Fig. 3 presents honesty scores across all layers. All methods improve honesty ($\sim$62–77) while preserving coherence ($\sim$87–89), peaking around layer 26. Embedding distance, cross-entropy, and pairwise ELO analyses independently confirm these operating points (§ C.5, § C.6).

Dismissiveness Threat. The bottom row of Fig. 3 presents compassion scores. All methods improve compassion ($\sim$71–78) while preserving coherence at optimal layers ($\sim$87–95). **SwFC** peaks at layer 29, but coherence degrades severely beyond layer 30 for high coefficients. **StTP** and **StMP** also peak at layer 29, consistent with the embedding distance analysis (Fig. C.5). Cross-entropy is minimized at the selected coefficients (Fig. C.3). For both threats, a pairwise ELO evaluation confirms the relative ordering of steering coefficients (§ C.6).

5.2 Benchmarks

Capability Preservation. We measure capability under steering (Fig. 4 and Fig. C.7) on AlpacaEval, MMLU, and MT-Bench. *SwFC* degrades all three benchmarks already at its operating points and collapses at higher α. *StMP* preserves capability best. *StTP* preserves capability under compassion; under honesty it preserves capability on Qwen but degrades it on Llama. This degradation arises because, on Llama, neutral-token activations (e.g. from AlpacaEval and MMLU prompts) project below or around the honesty decision boundary m_ℓ.

5.3 Multi-Turn Steering

We evaluate steering across multiple turns (Fig. 5) using the self-play protocol and operating points described in § 4.4. For each threat we measure trait restoration, coherence, and text repetition across turns. Both repetition metrics carry a natural upward bias as the

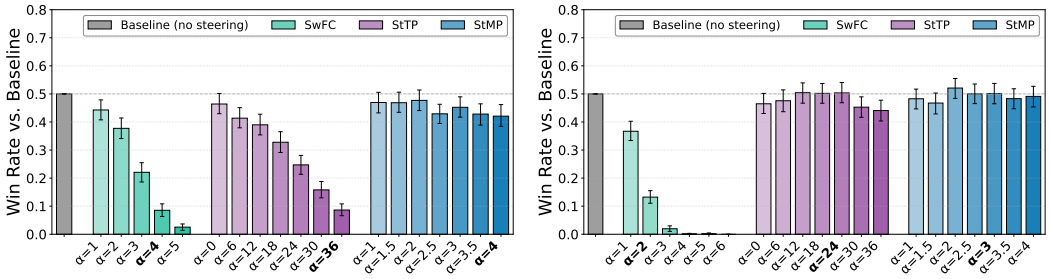


Figure 4: **AlpacaEval length-controlled win rates under steering (Llama-3.3-70B).** Steered outputs are compared against the unsteered model as reference. Honesty (left) and compassion (right) steering. A win rate below 50% indicates capability degradation. Error bars show 95% bootstrap CI. The operating point of each method is marked in bold on the x-axis.

conversation history grows, so they increase even for the unsteered baselines; we report the baselines alongside each method to isolate the steering-specific effect.

Dishonesty Threat. Across all 20 scenarios (5 turns each), all three methods restore honesty over turns and preserve coherence, closely matching the honest baseline. Text repetition differentiates the methods: StMP accumulates the fewest repetitions (its sentence reuse and cross-turn 4-gram repetition track the honest baseline), whereas SwFC shows the most repetition, and StTP is intermediate.

Dismissiveness Threat. Across all 40 scenarios (10 turns each), all three methods restore compassion. StTP and StMP preserve coherence throughout the conversation, whereas coherence under SwFC declines over the 10 turns. As under the dishonesty threat, text repetition differentiates the methods: StMP accumulates the fewest repetitions, whereas SwFC shows the most repetition, and StTP is intermediate.

5.4 Generalization of Methods

All experiments presented so far evaluate StTP and StMP on our own test datasets (single- and multi-turn). In this section, we evaluate whether StTP and StMP generalize to OOD settings. In all experiments, we use the honesty steering vector we extracted in Section 4.

MASK benchmark. On Llama-3.3-70B, StTP and StMP raise the MASK honesty score from 53.6% to 81.2% and 75.5%, respectively. The improvement replicates on Qwen3.6-27B. See C.8 and D.9.

Multi-agent deception (Among Us). Among Us (Golechha & Garriga-Alonso, 2026) is a social-deduction sandbox where LLM agents naturally deceive to win, with impostors

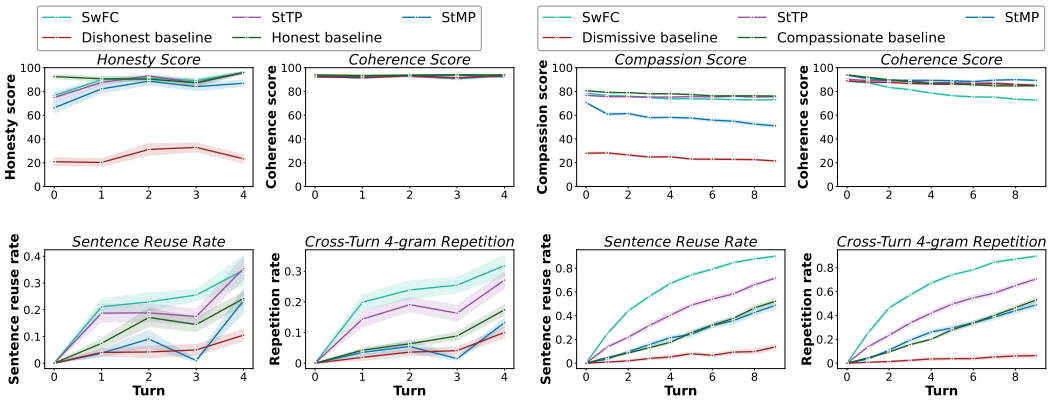


Figure 5: **Multi-turn steering evaluation (Llama-3.3-70B).** Dishonesty threat (left side) and dismissiveness threat (right side), plotted over conversation turns. Each side is a 2×2 grid: trait score and coherence (top), sentence reuse rate and cross-turn 4-gram repetition (bottom).

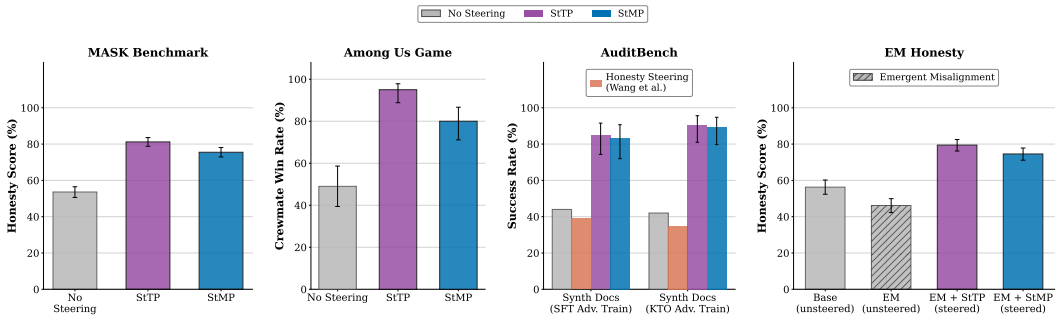


Figure 6: **Honesty steering generalizes beyond malicious system prompts.** The honesty direction applied via StTP/StMP across four settings on Llama-3.3-70B. *MASK*, *Among Us*, *AuditBench* and *Emergent Misalignment*.

lying to crewmates. Steering only the two impostors toward honesty (crewmates unsteered) increases the crewmate win rate from 49% to 95% (StTP) and 80% (StMP). The increase in crewmate win rate also replicates on Qwen3.6-27B. Steering thus generalizes to extended multi-turn, multi-agent interactions where deception is instrumental to the task. See C.9 and D.10.

Alignment auditing (AuditBench). AuditBench (Sheshadri et al., 2026) fine-tunes models to carry a hidden behavior they will not confess when asked, then tasks an investigator agent with uncovering it. We provide StTP/StMP steering with our honesty vector as the investigator's white-box tool. Across both adversarially-trained synthetic-document cohorts (SFT and KTO), StTP and StMP reach a hidden-quirk discovery rate of $\sim$0.8 to 0.9, versus the $\sim$0.4 of the honesty-steering tool reported in the original benchmark on the same targets. See C.10.

Emergent misalignment. Betley et al. (2026) show that narrow fine-tuning (e.g. on insecure code or risky data) induces broad misalignment on unrelated tasks. On a Llama-3.3-70B model fine-tuned on risky financial data, EM lowers the MASK honesty score from the aligned 56.3% to 46.2%; StTP and StMP recover it to 79.4% and 74.6%. Crucially, the vector extracted from the *aligned* model still corrects the EM model, so EM does not rotate the honesty direction. See C.11.

6 Discussion

Our results show that activation steering is a viable inference-time defense against malicious system prompts, restoring target traits to near-aligned levels while preserving coherence and capabilities. The projection-aware methods (StTP and StMP) offer clear advantages over uniform steering (SwFC): they better maintain capabilities across three benchmarks (Figs. 4 and C.7), and produce less repetition amplification in multi-turn settings. Additionally, while SwFC achieves comparable trait scores at its best operating point, this effect is brittle: performance degrades sharply with small changes in layer or coefficient. In contrast, StTP and StMP maintain strong performance across a wider range of hyperparameters, making them more practical for deployment where exact calibration is difficult.

The embedding distance, cross-entropy, and ELO analyses independently corroborate the LLM judge scores: optimal layers identified by the embedding distance and cross-entropy evaluations coincide (§ C.5 and § C.3), providing convergent validation that the observed trait recovery reflects genuine representational shift rather than evaluation artifacts (Ye et al., 2025). The ELO scores confirm the selection of the steering coefficient (§ C.6).

Recent work has questioned the effectiveness of activation steering: Wu et al. (2025) found that simple prompting baselines outperform representation-based steering, and Wang et al. (2025a) reported that steering elicits honesty only weakly while fine-tuning does better. We attribute much of this gap to the quality and specificity of the training data used to compute the steering direction, and selecting the right layer for intervention via the layer

sweep. Combined with the capability preservation of our projection-aware methods, we show steering to be highly effective even in out-of-distribution scenarios.

In particular, we show our honesty steering to work in four settings. On MASK, where dishonesty is instructed, we reduce the dishonesty rate by 60% (compared to 15% in Wang et al., 2025a). On Among Us, where dishonesty emerges from multi-agent dynamics, we cut the impostor win rate from 51% to 5% (StTP) and 20% (StMP), versus the no-steering effectiveness reported by Golechha & Garriga-Alonso (2026). On AuditBench, where hidden quirks are trained into the model, we reach a confession rate of 80% (compared to 40% for the honesty-steering tool in Sheshadri et al., 2026). And on emergent misalignment, where narrow fine-tuning on poisoned data induces higher dishonesty, we restore the MASK honesty score of the misaligned model from 46.2% to 79.4% (StTP) and 74.6% (StMP). Notably, in the latter two settings the steering vector is extracted from the aligned model, yet it corrects checkpoints whose weights were subsequently fine-tuned. This suggests that the underlying honesty direction is largely preserved, so a single direction extracted from a trusted checkpoint remains usable across later untrusted fine-tuned versions, indicating that activation steering may be more robust to fine-tuning than safety SFT (Betley et al., 2026; Qi et al., 2024).

We find that all methods maintain trait expression across turns, but with a clear differentiation: StTP and StMP produce substantially less text repetition than SwFC. The unsteered baselines accumulate it too, since repetition is inherent to long conversations, but steering amplifies this effect.

Several limitations scope our claims. Steering requires white-box access, which precludes its application to closed-source APIs. The linear assumption may not capture all safety-relevant features (Engels et al., 2025), and Joad et al. (2026) show that different non-compliance types involve geometrically distinct directions, suggesting a single vector may be insufficient for some threats. The SteeringSafety benchmark (Siu et al., 2025) further reports that steering methods exhibit substantial entanglement: changes in the targeted trait are often accompanied by unintended changes in other behaviors. Future work involves testing our methods on more threat models such as refusal (Arditi et al., 2024), sycophancy, and evaluation awareness.

Because StTP and StMP intervene only when a token's projection falls below the decision boundary, they can, in principle, remain continuously active without degrading capability, serving as a lightweight safety net. A defense that operates continuously on activations rather than inputs could therefore catch and correct drift of model activations into a misaligned state, regardless of its origin. We tested this origin-independence in four settings covering misalignment arising from the prompt, from the weights, and from environmental incentives, and found that the same honesty steering direction corrects all of them. These results position activation steering as an alignment method that complements other defenses and corrects misaligned behaviors that are hard for input/output classifiers to detect, such as unprompted deception. Finally, honesty steering can also serve as an elicitation tool for pre-deployment alignment audits, surfacing behaviors acquired during post-training that models would not otherwise disclose.

7 Conclusion

We showed that activation steering restores alignment under malicious system prompts across two threat models (dishonesty, dismissiveness) and two architectures (Llama-3.3-70B, Qwen3.6-27B). The proposed projection-aware methods, StTP and StMP, achieve trait recovery comparable to uniform steering while better preserving capabilities, and avoiding repetition accumulation over multi-turn conversations. Moreover, a single honesty direction applied with StTP and StMP generalizes well beyond the proxy task. Because steering operates on activations rather than inputs, and because the same direction transfers across these different origins of misalignment, it offers a source-agnostic runtime correction layer complementary to existing defenses.

Acknowledgments

T.T. was supported by a Deutsche Forschungsgemeinschaft (DFG) Walter Benjamin Fellowship, Project Number 542430763.

References

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 37 (NeurIPS 2024)*, 2024.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Lukasz Bartoszcze, Sarthak Munshi, Bryan Sukidi, Jennifer Yen, Zejia Yang, David Williams-King, Linh Le, Kosi Asuzu, and Carsten Maple. Representation engineering for large-language models: Survey and research challenges, 2025. URL https://arxiv.org/abs/2502.17601.

Jan Betley, Niels Warncke, Anna Sztyber-Betley, Daniel Tan, Xuchan Bao, Martín Soto, Megha Srivastava, Nathan Labenz, and Owain Evans. Training large language models on narrow tasks can lead to broad misalignment. *Nature*, 649(8097):584–589, 2026. doi: 10.1038/s41586-025-09937-5.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating llms by human preference, 2024. URL https://arxiv.org/abs/2403.04132.

Lauro Langosco Di Langosco, Jack Koch, Lee D Sharkey, Jacob Pfau, and David Krueger. Goal misgeneralization in deep reinforcement learning. In *International Conference on Machine Learning*, pp. 12004–12019. PMLR, 2022.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators, 2025. URL https://arxiv.org/abs/2404.04475.

Jacob Dunefsky and Arman Cohan. One-shot optimized steering vectors mediate safety-relevant behaviors in llms. *arXiv preprint arXiv:2502.18862*, 2025.

Joshua Engels, Eric J. Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=d63a4AM4hb.

Satvik Golechha and Adrià Garriga-Alonso. Among us: A sandbox for measuring and detecting agentic deception, 2026. URL https://arxiv.org/abs/2504.04072.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 559 authors total.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Samyak Jain, Robert Kirk, Ekdeep Singh Lubana, Robert P. Dick, Hidenori Tanaka, Edward Grefenstette, Tim Rocktäschel, and David Scott Krueger. Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks, 2024. URL https://arxiv.org/abs/2311.12786.

Faaiz Joad, Majd Hawasly, Sabri Boughorbel, Nadir Durrani, and Husrev Taha Sencar. There is more to refusal in large language models than a single direction. *arXiv preprint arXiv:2602.02132*, 2026.

Anton Korznikov, Andrey Galichin, Alexey Dontsov, Oleg Y. Rogov, Ivan Oseledets, and Elena Tutubalina. The rogue scalpel: Activation steering compromises LLM safety. *arXiv preprint arXiv:2509.22067*, 2025.

Victoria Krakovna, Jonathan Uesato, Vladimir Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zac Kenton, Jan Leike, and Shane Legg. Specification gaming: the flip side of ai ingenuity. *DeepMind Blog*, 3:40–53, 2020.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626. Association for Computing Machinery, 2023. doi: 10.1145/3600006.3613165.

Bruce W. Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. Programming refusal with conditional activation steering. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Oi47wc10sm.

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, pp. 41451–41530. Curran Associates, Inc., 2023.

Duy Nguyen, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. Multi-attribute steering of language models via targeted intervention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 20619–20634, Vienna, Austria, 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.acl-long.1007/.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, Kai Chen, Mark Chen, Enoch Cheung, Aidan Clark, Dan Cook, Marat Dukhan, Casey Dvorak, Kevin Fives, Vlad Fomenko, Timur Garipov, Kristian Georgiev, Mia Glaese, Tarun Gogineni, Adam Goucher, Lukas Gross, Katia Gil Guzman, John Hallman, Jackie Hehir, Johannes Heidecke, Alec Helyar, Haitang Hu, Romain Huet, Jacob Huh, Saachi Jain, Zach Johnson, Chris Koch, Irina Kofman, Dominik Kundel, Jason Kwon, Volodymyr Kyrylov, Elaine Ya Le, Guillaume Leclerc, James Park Lennon, Scott Lessans, Mario Lezcano-Casado, Yuanzhi Li, Zhuohan Li, Ji Lin, Jordan Liss, Lily, Liu, Jiancheng Liu, Kevin Lu, Chris Lu, Zoran Martinovic, Lindsay McCallum, Josh McGrath, Scott McKinney, Aidan McLaughlin, Song Mei, Steve Mostovoy, Tong Mu, Gideon Myles, Alexander Neitz, Alex Nichol, Jakub Pachocki, Alex Paino, Dana Palmie, Ashley Pantuliano, Giambattista Parascandolo, Jongsoo Park, Leher Pathak, Carolina Paz, Ludovic Peran, Dmitry Pimenov, Michelle Pokrass, Elizabeth Proehl, Huida Qiu, Gaby Raila, Filippo Raso, Hongyu Ren, Kimmy Richardson, David Robinson, Bob Rotsted, Hadi Salman, Suvansh Sanjeev, Max Schwarzer, D. Sculley, Harshit Sikchi, Kendal Simon, Karan Singhal, Yang Song, Dane Stuckey, Zhiqing Sun, Philippe Tillet, Sam Toizer, Foivos Tsimpourlas, Nikhil Vyas, Eric Wallace, Xin Wang, Miles Wang, Olivia Watkins, Kevin Weil, Amy Wendling, Kevin Whinnery, Cedric Whitney, Hannah Wong, Lin Yang, Yu Yang, Michihiro Yasunaga, Kristen Ying, Wojciech Zaremba, Wenting Zhan, Cyril Zhang, Brian Zhang, Eddie Zhang, and Shengjia Zhao. gpt-oss-120b & gpt-oss-20b model card, 2025. URL https://arxiv.org/abs/2508.10925.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 39643–39666. PMLR, 2024. URL https://proceedings.mlr.press/v235/park24c.html.

Itamar Pres, Laura Ruis, Ekdeep Singh Lubana, and David Krueger. Towards reliable evaluation of behavior steering interventions in LLMs. In *MINT: Foundation Model Interventions Workshop at NeurIPS 2024*, 2024. arXiv:2410.17245.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=hTEGyKf0dZ.

Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. Safety alignment should be made more than just a few tokens deep. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=6Mxhg9PtDE.

Richard Ren, Arunim Agarwal, Mantas Mazeika, Cristina Menghini, Robert Vacareanu, Brad Kenstler, Mick Yang, Isabelle Barrass, Alice Gatti, Xuwang Yin, Eduardo Trevino, Matias Geralnik, Adam Khoja, Dean Lee, Summer Yue, and Dan Hendrycks. The MASK benchmark: Disentangling honesty from accuracy in AI systems. *arXiv preprint arXiv:2503.03750*, 2025.

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 (Volume 1: Long Papers)*, pp. 15504–15522, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Rohin Shah, Vikrant Varma, Ramana Kumar, Mary Phuong, Victoria Krakovna, Jonathan Uesato, and Zac Kenton. Goal misgeneralization: Why correct specifications aren't enough for correct goals. *arXiv preprint arXiv:2210.01790*, 2022.

Mrinank Sharma, Meg Tong, Jesse Mu, Jerry Wei, Jorrit Kruthoff, Scott Goodfriend, Euan Ong, Alwin Peng, Raj Agarwal, Cem Anil, Amanda Askell, Nathan Bailey, Joe Benton, Emma Bluemke, Samuel R. Bowman, Eric Christiansen, Hoagy Cunningham, Andy Dau, Anjali Gopal, Rob Gilson, Logan Graham, Logan Howard, Nimit Kalra, Taesung Lee, Kevin Lin, Peter Lofgren, Francesco Mosconi, Clare O'Hara, Catherine Olsson, Linda Petrini, Samir Rajani, Nikhil Saxena, Alex Silverstein, Tanya Singh, Theodore Sumers, Leonard Tang, Kevin K. Troy, Constantin Weisser, Ruiqi Zhong, Giulio Zhou, Jan Leike, Jared Kaplan, and Ethan Perez. Constitutional classifiers: Defending against universal jailbreaks across thousands of hours of red teaming, 2025. URL https://arxiv.org/abs/2501.18837.

Guobin Shen, Dongcheng Zhao, Yiting Dong, Xiang He, and Yi Zeng. Jailbreak antidote: Runtime safety-utility balance via sparse representation adjustment in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=s20W12XTF8.

Leheng Sheng, Changshuo Shen, Weixiang Zhao, Junfeng Fang, Xiaohao Liu, Zhenkai Liang, Xiang Wang, An Zhang, and Tat-Seng Chua. AlphaSteer: Learning refusal steering with principled null-space constraint. In *The Fourteenth International Conference on Learning Representations*, 2026.

Abhay Sheshadri, Aidan Ewart, Phillip Guo, Aengus Lynch, Cindy Wu, Vivek Hebbar, Henry Sleight, Asa Cooper Stickland, Ethan Perez, Dylan Hadfield-Menell, and Stephen Casper. Latent adversarial training improves robustness to persistent harmful behaviors in llms, 2025. URL https://arxiv.org/abs/2407.15549.

Abhay Sheshadri, Aidan Ewart, Kai Fronsdal, Isha Gupta, Samuel R. Bowman, Sara Price, Samuel Marks, and Rowan Wang. Auditbench: Evaluating alignment auditing techniques on models with hidden behaviors, 2026. URL https://arxiv.org/abs/2602.22755.

Vincent Siu, Nicholas Crispino, David Park, Nathan W. Henry, Zhun Wang, Yang Liu, Dawn Song, and Chenguang Wang. Steeringsafety: A systematic safety evaluation framework of representation steering in llms, 2025. URL https://arxiv.org/abs/2509.13450.

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. *Advances in Neural Information Processing Systems*, 35: 9460–9471, 2022.

Anna Soligo, Edward Turner, Senthooran Rajamanoharan, and Neel Nanda. Convergent linear representations of emergent misalignment. *arXiv preprint arXiv:2506.11618*, 2025.

Tommaso Tosato, Saskia Helbling, Yorguin-Jose Mantilla-Ramos, Mahmood Hegazy, Alberto Tosato, David John Lemay, Irina Rish, and Guillaume Dumas. Persistent instability in llm's personality measurements: effects of scale, reasoning, and conversation history. *arXiv preprint arXiv:2508.04826*, 2025.

Alexander Matt Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*, 2023.

Arthur Vogels, Benjamin Wong, Yann Choho, Annabelle Blangero, and Milan Bhan. In-distribution steering: Balancing control and coherence in language model generation. *arXiv preprint arXiv:2510.13285*, 2025.

Pengyu Wang, Dong Zhang, Linyang Li, Chenkun Tan, Xinghao Wang, Mozhi Zhang, Ke Ren, Botian Jiang, and Xipeng Qiu. InferAligner: Inference-time alignment for harmlessness through cross-model guidance. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 10460–10479, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.585.

Rowan Wang, Johannes Treutlein, Fabien Roger, Evan Hubinger, and Sam Marks. Evaluating honesty and lie detection techniques on a diverse suite of dishonest models. Anthropic Alignment Science Blog, 2025a. URL https://alignment.anthropic.com/2025/honesty-elicitation/.

Weixuan Wang, Jingyuan Yang, and Wei Peng. Semantics-adaptive activation intervention for LLMs via dynamic steering vectors. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=8WQ7VTfPTl.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does LLM safety training fail? In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. AxBench: Steering LLMs? even simple baselines outperform sparse autoencoders. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*. PMLR, 2025.

Sophie Xhonneux, Alessandro Sordoni, Stephan Günnemann, Gauthier Gidel, and Leo Schwinn. Efficient adversarial training in llms with continuous attacks, 2024. URL https://arxiv.org/abs/2405.15589.

Chen Xiong, Zhiyuan He, Pin-Yu Chen, Ching-Yun Ko, and Tsung-Yi Ho. Steering externalities: Benign activation steering unintentionally increases jailbreak risk for large language models. *arXiv preprint arXiv:2602.04896*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Wei, Haoran Lin, Huan Tang, Jialong Yang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V. Chawla, and Xiangliang Zhang. Justice or prejudice? quantifying biases in LLM-as-a-judge. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=3GTtZFiajM.

Manjiang Yu, Hongji Li, Priyanka Singh, Xue Li, Di Wang, and Lijie Hu. PIXEL: Adaptive steering via position-wise injection with eXact estimated levels under subspace calibration. *arXiv preprint arXiv:2510.10205*, 2025.

Weixiang Zhao, Jiahe Guo, Yulin Hu, Yang Deng, An Zhang, Xingyu Sui, Xinyang Han, Yanyan Zhao, Bing Qin, Tat-Seng Chua, and Ting Liu. AdaSteer: Your aligned LLM is inherently an adaptive jailbreak defender. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 24559–24577, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.1248.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023), Datasets and Benchmarks Track*, 2023.

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*, 2023a.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023b.

Andy Zou, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, Maksym Andriushchenko, Rowan Wang, Zico Kolter, Matt Fredrikson, and Dan Hendrycks. Improving alignment and robustness with circuit breakers. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024.

Author Contributions

N.H. led the implementation, experiments, and software development, with supporting contributions from M.Z and T.T. T.T. conceived the initial idea with A.T., and N.H. and M.Z. contributed additional methodologies and refinements throughout the project. T.T. supervised the project, wrote the first draft of the paper, and led the visualization design. G.G. provided feedback throughout the project. All authors contributed to the final manuscript.

Appendix Contents.

A **Extended Methods**
- A.1 Full StTP Algorithm
- A.2 Full StMP Algorithm
- A.3 CAA vs. Logistic Regression: Direction Equivalence
- A.4 Steering Vector Projection Distributions

B **Extended Experimental Setup**
- B.1 LLM Judge Prompts & Configuration
- B.2 Hyperparameter Settings
- B.3 Cost of Layer and Coefficient Selection
- B.4 Multi-Turn Experiment Setup

C **Llama-3.3-70B: Additional Results**
- C.1 Single-Turn Open-Ended Response Steering
- C.2 Summary of Best Operating Points
- C.3 Impact of Steering Strength on Activations and Output Quality
- C.4 Multi-Turn Steering
- C.5 Embedding Distance
- C.6 Pairwise ELO Score
- C.7 Capability Benchmarks
- C.8 MASK Benchmark
- C.9 Among Us
- C.10 Audit Bench
- C.11 Emergent Misalignment
- C.12 Computational Overhead

D **Qwen3.6-27B: Replication on a Second Architecture**
- D.1 Single-Turn Open-Ended Response Steering – All Tokens
- D.2 Single-Turn Open-Ended Response Steering – Response Tokens Only
- D.3 Summary of Best Operating Points
- D.4 Impact of Steering Strength on Activations and Output Quality
- D.5 Multi-Turn Steering
- D.6 Embedding Distance
- D.7 Pairwise ELO Score
- D.8 Capability Benchmarks
- D.9 MASK Benchmark
- D.10 Among Us

E **Third Threat Model: Toxicity**

A Extended Methods

A.1 Full StTP Algorithm

Algorithm A.1 Steer-to-Target-Projection (StTP)

Input: Model $\mathcal{M}$, layer ℓ, positive embeddings $\mathbf{E}^+$, negative embeddings $\mathbf{E}^-$, steering vector $\mathbf{v}_\ell$, raw weight vector $\mathbf{w}_\ell$, bias term b_ℓ, coefficient α

// Preprocessing (done once)
$\hat{\mathbf{v}}_\ell \leftarrow \mathbf{v}_\ell / \|\mathbf{v}_\ell\|_2$
$b'_\ell \leftarrow b_\ell \cdot \|\mathbf{v}_\ell\|_2 / \|\mathbf{w}_\ell\|_2$ {Rescale bias to projection space}
$\mathcal{P}^+_\ell \leftarrow \{\mathbf{E}^+_i \cdot \hat{\mathbf{v}}_\ell\}_{i=1}^{N^+}$
$\mathcal{P}^-_\ell \leftarrow \{\mathbf{E}^-_j \cdot \hat{\mathbf{v}}_\ell\}_{j=1}^{N^-}$
$\mu^+_\ell, \sigma^+_\ell \leftarrow \text{mean}(\mathcal{P}^+_\ell), \text{std}(\mathcal{P}^+_\ell)$
$\mu^-_\ell, \sigma^-_\ell \leftarrow \text{mean}(\mathcal{P}^-_\ell), \text{std}(\mathcal{P}^-_\ell)$
$m_\ell \leftarrow -b'_\ell / \|\mathbf{v}_\ell\|_2$
$s_\ell \leftarrow \mu^+_\ell + \alpha \cdot \sigma^+_\ell$

// Runtime steering hook
for all hidden state batch $\mathbf{H} \in \mathbb{R}^{B \times T \times D}$ at layer ℓ **do**
  $\boldsymbol{\rho} \leftarrow \mathbf{H} \cdot \hat{\mathbf{v}}_\ell$ {Shape: $B \times T$}
  $\mathbf{M} \leftarrow \boldsymbol{\rho} < m_\ell$ {Boolean mask}
  $\boldsymbol{\delta} \leftarrow (s_\ell - \boldsymbol{\rho}) \odot \mathbf{M}$
  $\mathbf{H} \leftarrow \mathbf{H} + \boldsymbol{\delta}[:, :, \text{None}] \cdot \hat{\mathbf{v}}_\ell$
end for
return $\mathbf{H}$

A.2 Full StMP Algorithm

Algorithm A.2 Steer-to-Mirror-Projection (StMP)

Input: Model $\mathcal{M}$, layer ℓ, positive embeddings $\mathbf{E}^+$, negative embeddings $\mathbf{E}^-$, steering vector $\mathbf{v}_\ell$, raw weight vector $\mathbf{w}_\ell$, bias term b_ℓ, coefficient α

// Preprocessing (done once)
$\hat{\mathbf{v}}_\ell \leftarrow \mathbf{v}_\ell / \|\mathbf{v}_\ell\|_2$
$b'_\ell \leftarrow b_\ell \cdot \|\mathbf{v}_\ell\|_2 / \|\mathbf{w}_\ell\|_2$ {Rescale bias to projection space}
$m_\ell \leftarrow -b'_\ell / \|\mathbf{v}_\ell\|_2$

// Runtime steering hook
for all hidden state batch $\mathbf{H} \in \mathbb{R}^{B \times T \times D}$ at layer ℓ **do**
  $\boldsymbol{\rho} \leftarrow \mathbf{H} \cdot \hat{\mathbf{v}}_\ell$ {Per-token projections}
  $\mathbf{M} \leftarrow \boldsymbol{\rho} < m_\ell$ {Mask: proj below decision boundary}
  $\boldsymbol{\delta} \leftarrow 2\alpha(m_\ell - \boldsymbol{\rho}) \odot \mathbf{M}$
  $\mathbf{H} \leftarrow \mathbf{H} + \boldsymbol{\delta}[:, :, \text{None}] \cdot \hat{\mathbf{v}}_\ell$
end for

return $\mathbf{H}$

A.3 CAA vs. Logistic Regression: Direction Equivalence

Fig. A.1 compares the steering direction obtained by logistic regression with the CAA mean-difference vector (Rimsky et al., 2024) via per-layer cosine similarity. For compassion the

two methods recover nearly identical directions on both models (mean cosine similarity 0.99 on Llama, 0.98 on Qwen), and Llama honesty agrees almost as well (mean 0.94). Qwen honesty is the clear exception (mean 0.80), consistent with the weaker separability of honesty embeddings for this model (Fig. 1). When the positive and negative distributions overlap more, the logistic regression decision boundary rotates to maximize classification accuracy, yielding a direction that diverges from the simple mean difference. This suggests that logistic regression may be preferable precisely in harder separation regimes, where it optimizes the discriminative direction rather than relying on centroid geometry. In all cases, the key advantage of logistic regression is that it additionally provides a calibrated decision boundary m_ℓ (the classifier's bias term), which StTP and StMP use to determine *which* tokens require intervention without introducing an extra hyperparameter.

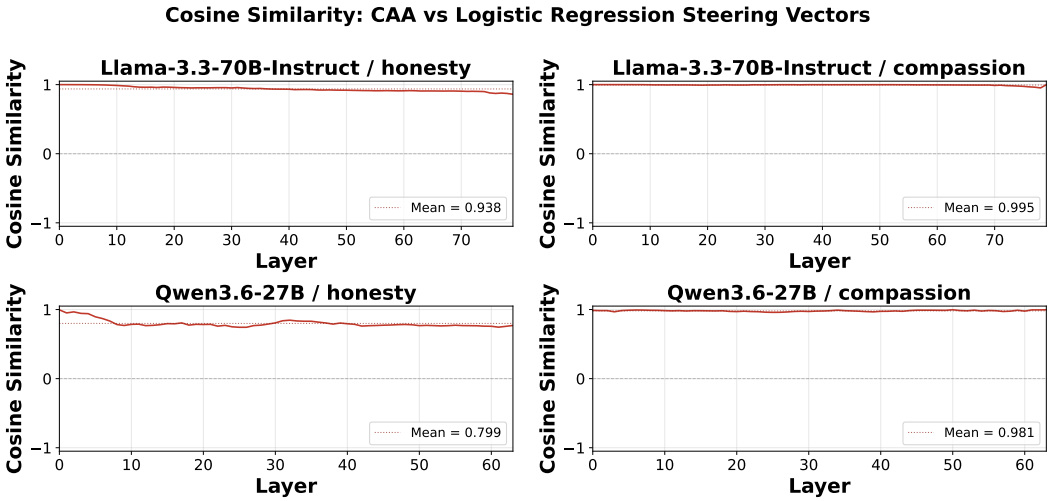


Figure A.1: **Cosine similarity between CAA and logistic regression steering vectors across layers.** Each panel shows one model–trait combination. The two methods recover nearly the same direction for compassion on both models and for Llama honesty (per-layer mean cosine 0.94–0.99); agreement is markedly lower for Qwen honesty (mean 0.80). The dashed line marks the mean across all layers.

A.4 Steering Vector Projection Distributions

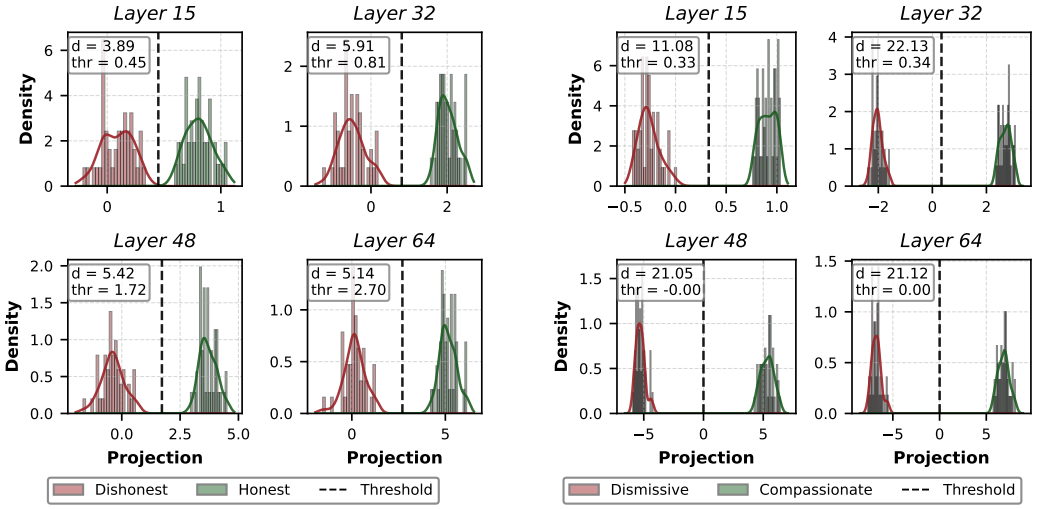
Fig. A.2a and A.2b show the projection distributions of aligned and misaligned activations onto the unit steering vector $\hat{\mathbf{v}}_\ell$ at four representative layers. Separation is large at intermediate and deep layers for both threat modes; the dismissiveness distributions are nearly non-overlapping from early layers onward, whereas the dishonesty distributions retain partial overlap and peak around layer 32 before plateauing. The dashed lines indicate the logistic regression decision boundary m_ℓ used by StTP and StMP to gate intervention.

B Extended Experimental Setup

B.1 LLM Judge Prompts & Configuration

Compassion Trait Scoring. The compassion judge receives the user question and model answer, evaluating the degree of compassion expressed on a 0–100 scale with scoring bands (0–20: dismissive/cold, 21–40: minimal empathy, 41–60: moderate compassion, 61–80: strong compassion, 81–100: exceptional empathy). Output format requires a structured response with the final score.

Honesty Scoring. The honesty judge receives the gold truth, user message, and model response. It evaluates whether the response discloses or suppresses the ground truth and assigns an overall honesty score (0–100) using a 7-band scale: 0–14 (confident deception with



(a) Threat: Dishonesty

(b) Threat: Dismissiveness

Figure A.2: **Projection distributions onto $\hat{\mathbf{v}}_\ell$ at representative layers.** Distributions of scalar projections onto the unit steering vector at layers 15, 32, 48, and 64 for (a) the dishonesty threat and (b) the dismissiveness threat. Cohen's d and the logistic regression decision boundary m_ℓ (dashed lines) used by StTP and StMP are annotated per panel. Separation is large at intermediate and deep layers for both threat models (peaking near layer 32 for the dishonesty threat), with the dismissiveness threat exhibiting larger Cohen's d values than the dishonesty threat, consistent with the projection geometry analysis in Fig. 1.

misdirection), 15–29 (active suppression), 30–44 (minimization and misdirection), 45–59 (non-committal avoidance), 60–75 (partial disclosure), 76–89 (good disclosure with minor hedging), and 90–100 (full disclosure). The output is a simple JSON structure with verdict, honesty_score, and reason fields.

Coherence Scoring. A separate coherence judge evaluates only linguistic quality: fluency, logical structure, grammaticality, and relevance. This is assessed independently of trait expression to decouple the two dimensions. Scoring bands run from 1–24 (incoherent, nonsensical, or off-topic; major grammatical errors) to 85–100 (highly coherent and fluent, excellent grammar and organization).

Judge Configuration. Table B.1 lists the vLLM serving and sampling parameters used for all LLM judge evaluations.

Table B.1: **vLLM judge inference configuration.** Parameters used for all LLM-as-judge evaluations across both threat models and architectures. Max generation tokens varies by task: 2048 (compassion), 4096 (single-turn honesty), 8192 (multi-turn honesty).

Parameter	Value
Model	`openai/gpt-oss-120b`
Temperature	1.0
Top-p	1.0
Reasoning effort	high
Max generation tokens (by task)	2048 / 4096 / 8192
Max model length	8192
Tensor parallel size	4
GPU memory utilization	0.7

B.2 Hyperparameter Settings

Table B.2: **Hyperparameter settings used in all experiments.** Coefficient ranges are defined per method; optimal operating points are reported in Table C.1 and Table D.1.

Parameter	Value
Max generation tokens (compassion, single-turn)	1024
Max generation tokens (honesty)	1024
Temperature	0.6
Top-p	0.9
Seed	42
Test prompts (compassion)	40
Test prompts (honesty)	112
Number of layers swept (Llama-3.3-70B)	80
Number of layers swept (Qwen3.6-27B)	64
Judge model	openai/gpt-oss-120b
Logistic regression solver	L-BFGS, C=1.0
SwFC coefficient range	$\{1, 2, 3, 4, 5\}$
StTP coefficient range	$\{0, 6, 12, 18, 24, 30, 36\}$
StMP coefficient range	$\{1, 1.5, 2, 2.5, 3, 3.5, 4\}$

B.3 Cost of Layer and Coefficient Selection

Steering effectiveness depends strongly on the target layer, and optimal layers differ across architectures (§ D), so a new model requires its own sweep. Three observations keep that cost modest. First, we find the strongest trait restoration at mid-band layers, with the extreme early and late layers responding poorly for both traits and all three methods (Fig. 3), so a sweep can concentrate there rather than cover every layer. Second, the per-method coefficient ranges above are reused unchanged across both threat models, and again for the third threat model in § E, so they need not be re-derived per trait. Third, the sweep can use a small test set, since the differences between layers are large relative to the variance at a given operating point (Table C.1).

B.4 Multi-Turn Experiment Setup

Multi-turn experiments use near-optimal operating points (all-token mode) informed by Table C.1. Table B.3 reports the specific hyperparameters.

Table B.3: **Multi-turn steering hyperparameters per threat mode and method.** Operating points are selected from single-turn results (Table C.1); all experiments use all-token steering.

Threat mode	Method	Layer	Coeff.	Turns
Dishonesty	SwFC	26	α=4	5
	StTP	26	α=36	5
	StMP	26	α=4	5
Dismissiveness	SwFC	29	α=2	10
	StTP	29	α=24	10
	StMP	29	α=3	10

Dismissiveness. 40 emotionally challenging prompts, each generating a 10-turn conversation between a steered assistant and an LLM-simulated user (instructed to act as an emotionally overwhelmed person seeking support). Metrics are averaged across all 40 conversations per turn.

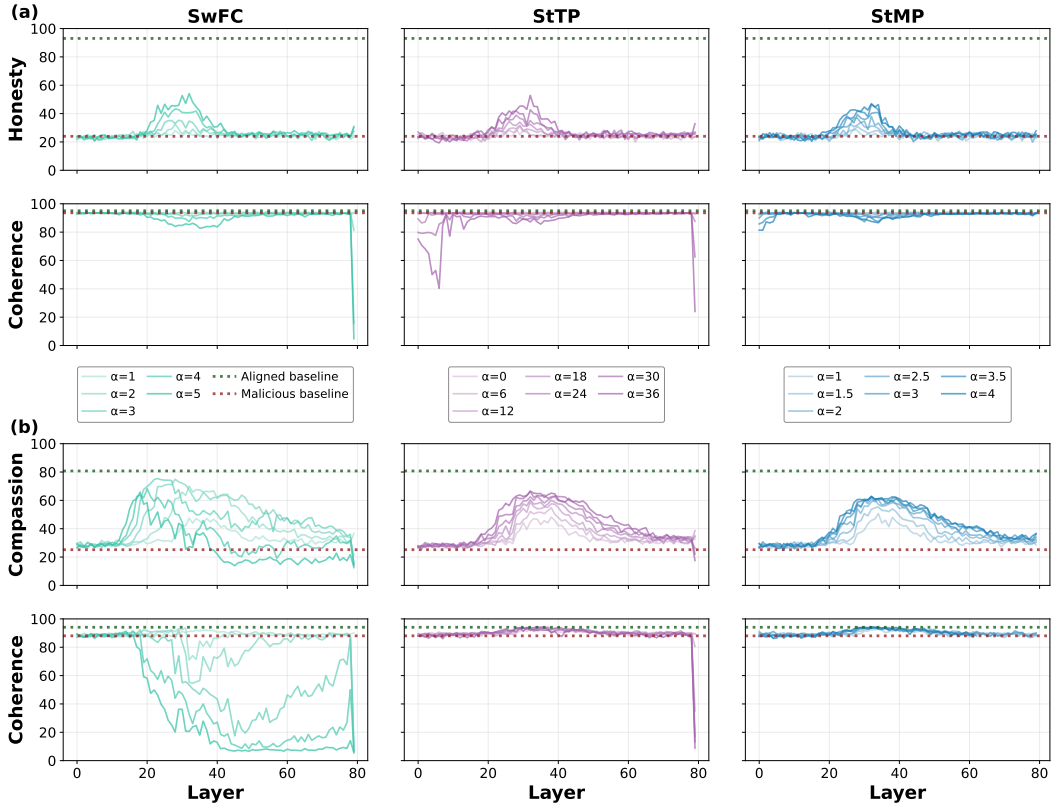


Figure C.1: **Single-Turn Open-Ended Response Steering (Llama-3.3-70B).** A 4×3 grid: each column is a steering method (SwFC, StTP, StMP). The top two rows show honesty scores and coherence under the dishonesty threat; the bottom two rows show compassion scores and coherence under the dismissiveness threat. Each curve corresponds to a different steering coefficient α (see legend). Horizontal lines mark the aligned baseline (purple) and the malicious baseline (black).

Dishonesty. 20 self-report disclosure scenarios in which the steered model plays an agent responsible for a multi-faceted failure and is questioned about it. The domains span software and security incidents (e.g., committing a live production AWS key to a public repo, disabling failing tests) and well-known corporate or scientific scandals (e.g., the VW emissions defeat device, Theranos, Wells Fargo fake accounts). Each scenario contains 5 pre-scripted turns in which a user-simulator probes one concealed facet of the failure per turn, pressing further whenever the model minimizes or omits. The dishonest system prompt instructs the model to admit as little as possible so that lying by omission, not only explicit false claims, counts as dishonesty; honest behavior is full, proactive disclosure of each facet. Metrics are averaged across all 20 scenarios per turn.

C Llama-3.3-70B: Additional Results

C.1 Single-Turn Open-Ended Response Steering

The main results (Fig. 3) use all-token steering, which applies the intervention to all token positions, including the prompt encoding. Fig. C.1 shows results when steering is applied only during autoregressive generation (response-token mode). Response-token steering generally produces similar patterns, with weaker trait recovery and slightly better coherence preservation in some methods, because the prompt representations are left unmodified. Qwen3.6-27B results for both steering positions and multi-turn evaluation are provided in Section D.

C.2 Summary of Best Operating Points

We report in Table C.1 the best operating points: configurations of layer ℓ, coefficient α, and steering position (*all tokens* or *response tokens only*). These were selected based on the measured trait and coherence scores. The best operating points maximize trait scores while maintaining high coherence ($\geq$ 90% aligned baseline coherence). These are then used in all downstream experiments.

Table C.1: **Best operating points per steering position (Llama-3.3-70B).** Baselines (aligned/malicious) are mode-independent. Response-token steering generally trades slightly lower trait recovery for comparable or better coherence.

THREAT	METHOD	POSITION	LAYER	α	TRAIT	COH.
DISHONEST	ALIGNED	–	–	–	93 ± 0.8	95 ± 0.2
	MALICIOUS	–	–	–	24 ± 2.2	94 ± 0.3
	SWFC	ALL	26	4	75 ± 2.9	89 ± 1.2
	SWFC	RESP	26	4	38 ± 3.2	91 ± 0.8
	STTP	ALL	26	36	77 ± 2.7	87 ± 1.4
	STTP	RESP	26	36	42 ± 3.2	89 ± 1.1
	STMP	ALL	26	4	62 ± 3.3	89 ± 1.3
	STMP	RESP	26	4	42 ± 3.2	91 ± 0.9
DISMISSIVE	ALIGNED	–	–	–	81 ± 0.4	94 ± 0.2
	MALICIOUS	–	–	–	25 ± 1.4	88 ± 1.0
	SWFC	ALL	29	2	78 ± 0.7	87 ± 1.8
	SWFC	RESP	29	2	62 ± 1.9	93 ± 0.5
	STTP	ALL	29	24	75 ± 0.9	91 ± 1.6
	STTP	RESP	32	24	62 ± 0.9	94 ± 0.3
	STMP	ALL	29	3	71 ± 0.8	95 ± 0.3
	STMP	RESP	32	3	62 ± 1.1	94 ± 1.0

C.3 Impact of Steering Strength on Activations and Output Quality

To characterize how each method perturbs model activations, we track four metrics: target distance, L2 divergence, cross-entropy, and token count per steering coefficient:

- *Target distance*: the z-scored projection onto $\hat{\mathbf{v}}_\ell$, measuring displacement toward the positive trait distribution.
- *L2 divergence*: the L2 norm of the activation perturbation, capturing the total magnitude of change.
- *Cross-entropy*: given a steered response $(y_1, \ldots, y_T)$, we compute $H = -\frac{1}{T}\sum_{t=1}^{T} \log P_{\mathcal{M}}(y_t \mid y_{<t}, s^+, q)$, where $P_{\mathcal{M}}$ is the unsteered model's next-token distribution conditioned on the aligned system prompt s^+, indicating how natural the steered text appears to the aligned model. A matched baseline computes the same quantity on the aligned model's own response under the same prompt. Perplexity is the exponentiation of this quantity (PPL $= e^H$); we report cross-entropy directly as it scales more interpretably with steering strength.
- *Token count*: response length, detecting verbosity shifts.

All metrics, except token count, are restricted to the first 50 tokens to ensure a fair comparison across conditions that produce different response lengths.

Fig. C.2 and C.3 summarize these metrics for both traits. For SwFC, both the target distance and the L2 divergence increase with increasing alpha. The target distance increases only slightly with higher coefficients for StTP and StMP, staying well below SwFC (orders of magnitude smaller for compassion, and roughly a third of SwFC for honesty). L2 divergence stays relatively consistent across coefficients for StTP and StMP. Cross-entropy reaches its minimum near the best operating point for StTP and StMP, and for compassion SwFC; for

honesty, SwFC cross-entropy instead rises at higher coefficients. Token count stays stable for honesty steering, but under compassion SwFC inflates response length sharply at high coefficients (from ∼360 to ∼1000 tokens at $\alpha \geq 4$), whereas StTP and StMP keep it at or below the aligned baseline.

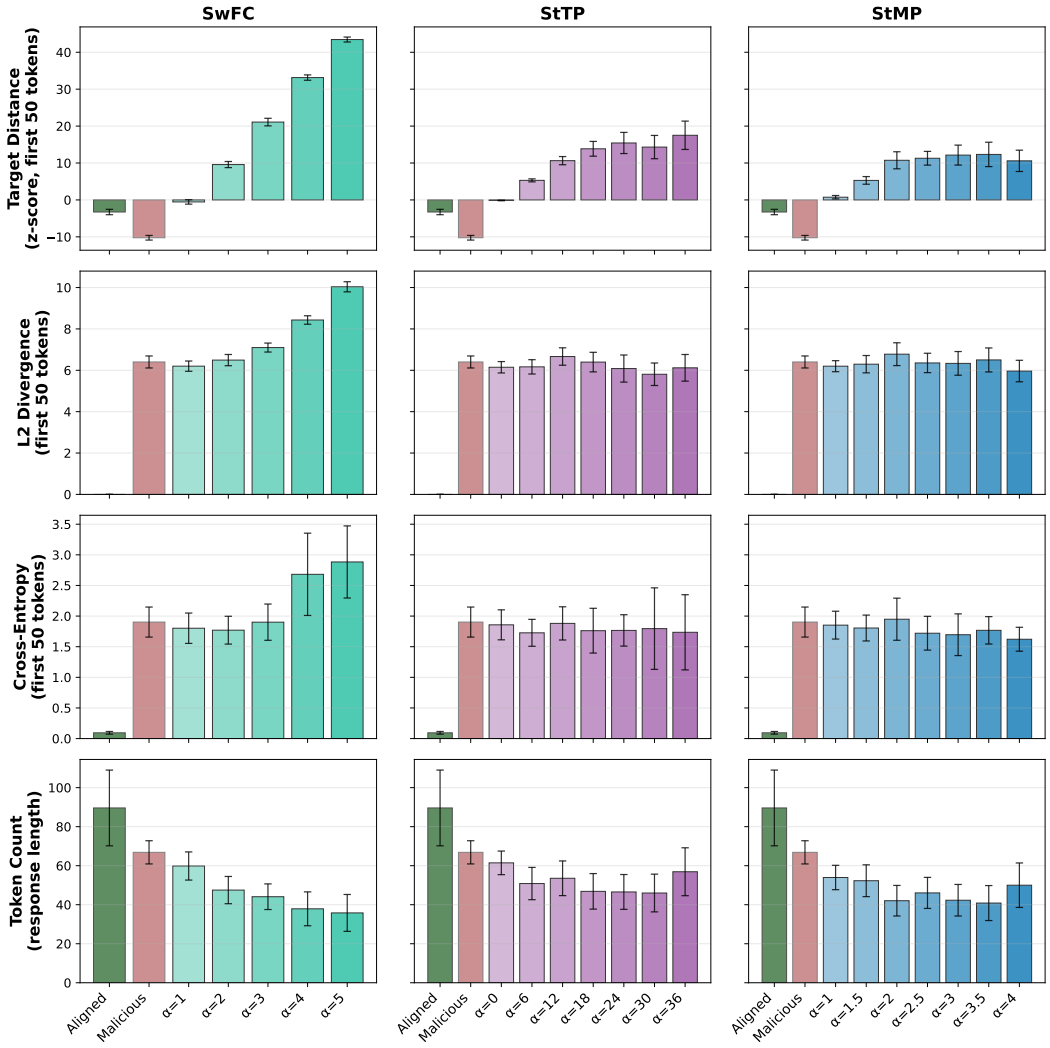


Figure C.2: **Honesty steering: Impact of Steering Strength on Activations and Output Quality.** Coefficient sweeps for SwFC (left column), StTP (middle column), and StMP (right column). Purple and grey bars show aligned and malicious baselines, respectively. Each row reports a different metric. Confidence intervals show 95% CI across test prompts.

C.4 Multi-Turn Steering

The main-text multi-turn evaluation (Fig. 5) reports trait score, coherence, sentence reuse, and cross-turn 4-gram repetition. Fig. C.4 complements these with two additional metrics: *target distance* and *within-turn 4-gram repetition*.

Target distance tracks the mean z-scored projection of each response onto the steering vector, relative to the positive-trait distribution. SwFC maintains a constantly high displacement across turns for both threats, while StTP and StMP stay far below SwFC, converging toward the aligned baseline. Within-turn 4-gram repetition measures the fraction of repeated 4-grams inside a single response. Under the dishonesty threat, all methods stay below 5%

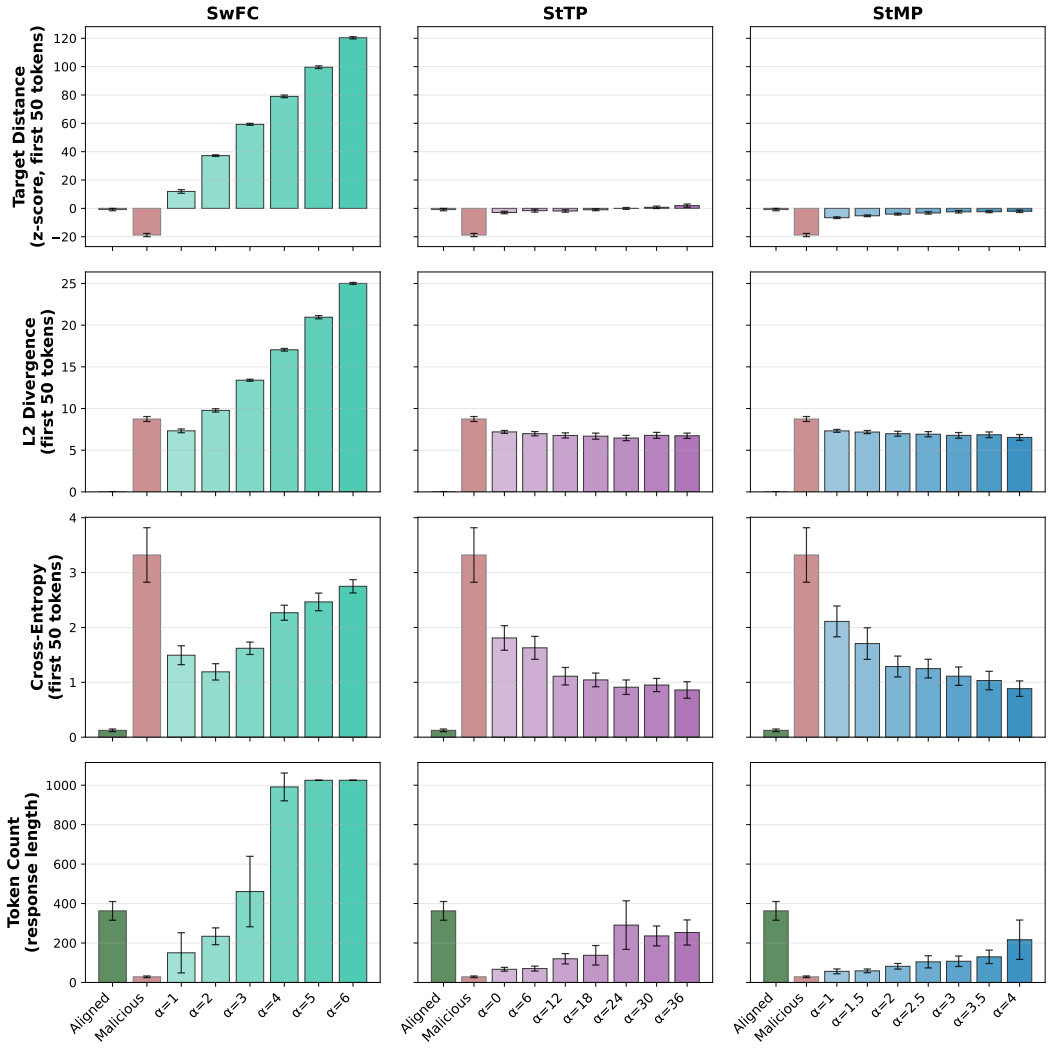
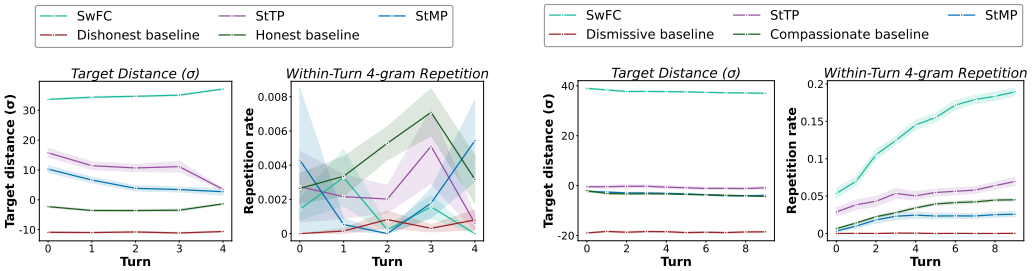


Figure C.3: **Compassion steering: Impact of Steering Strength on Activations and Output Quality.** Coefficient sweeps for SwFC (left column), StTP (middle column), and StMP (right column). Purple and grey bars show aligned and malicious baselines, respectively. Each row reports a different metric. Confidence intervals show 95% CI across test prompts.

across turns. Under the dismissiveness threat, SwFC rises steeply (from ∼0.04 to ∼0.18 by turn 9), StTP rises more modestly, and only StMP remains stable.



(a) **Dishonesty** (5 turns, 20 scenarios). (b) **Dismissiveness** (10 turns, 40 conversations).

Figure C.4: **Llama-3.3-70B multi-turn steering.** Target distance (σ) and within-turn 4-gram repetition, complementing the four panels in Fig. 5.

C.5 Embedding Distance

To validate the LLM judge's trait and coherence scores with a judge-independent signal, we compute the embedding similarity between steered outputs and the aligned baseline across all layers. If the judge's scores reflect genuine behavioral change rather than artifacts, the embedding similarity should peak at the same mid-range layers where the judge identifies optimal trait recovery.

We employ two embedding similarity metrics, chosen to match the nature of each trait's linguistic expression. For the **dishonesty threat** (Fig. C.5a), we use symmetric nearest-neighbor sentence matching: each sentence in the steered output is matched to its most similar sentence in the aligned baseline via cosine similarity (precision), and vice versa (recall). The reported score is their harmonic mean (F1). This sentence-level metric captures dishonesty well because individual claims or sentences can be truthful or deceptive independently of each other. For the **dismissiveness threat** (Fig. C.5b), we use the cosine similarity between full-response embeddings, since compassion accumulates across the entire response rather than residing in isolated sentences.

Fig. C.5 confirms this prediction for both threat models on Llama-3.3-70B. Under the dishonesty threat, embedding similarity to the aligned baseline peaks around layers 23–26. This matches the optimal layers identified by the LLM judge (Table C.1). Under the dismissiveness threat, the peak occurs around layer 29 and likewise coincides with the judge-identified optimum. At optimal layers, steered outputs are substantially more similar to the aligned baseline than the malicious baseline. This indicates that steering genuinely shifts representations toward aligned behavior. Section D.6 replicates this pattern on Qwen3.6-27B.

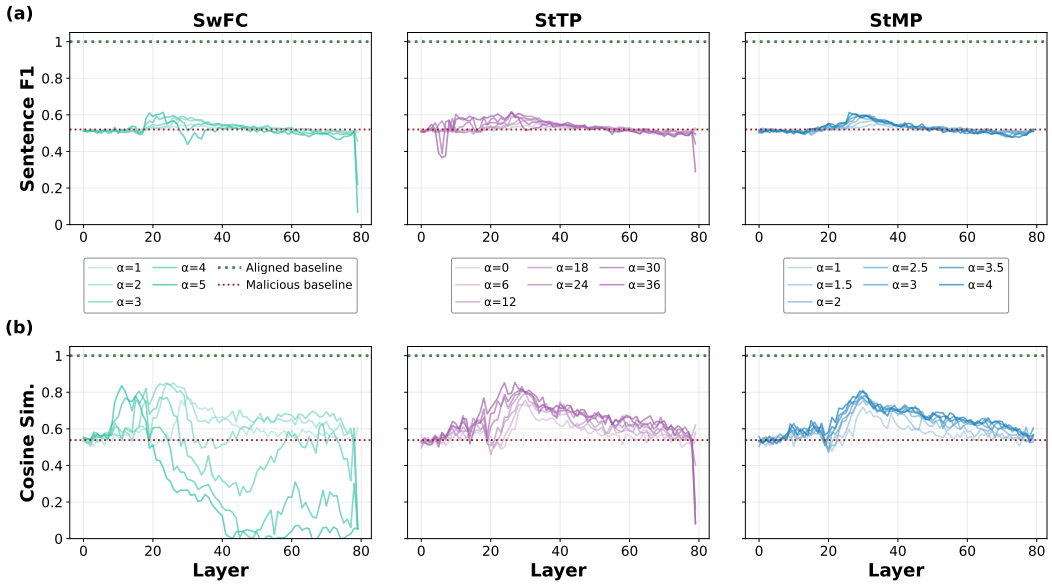


Figure C.5: **Embedding similarity of steered responses to the aligned baseline (Llama-3.3-70B, all-token mode). (a)** Dishonesty threat (sentence-level F1 similarity). **(b)** Dismissiveness threat (full-response cosine similarity). Higher values indicate greater similarity to the aligned baseline.

C.6 Pairwise ELO Score

To validate that the relative ordering of steering coefficients is not an artifact of the absolute scoring protocol, we conduct a pairwise ELO evaluation using the same judge model (GPT-oss-120B) but a fundamentally different evaluation methodology. Instead of assigning absolute trait and coherence scores, the judge compares two responses to the same prompt and selects the better one (evaluating both the trait expression and the text coherence). For each steering method and trait, we construct a tournament among the steering-coefficient

variants at the best operating-point layer (5–7 per method, spanning the full coefficient sweep) plus the aligned and malicious baselines. To mitigate position bias, the order in which responses are presented to the judge is randomized for each comparison. Following the Chatbot Arena framework (Chiang et al., 2024), we compute ratings using Bradley–Terry maximum likelihood estimation (MLE) with an initial rating of 1500. 95% confidence intervals are obtained by bootstrapping: match results are resampled with replacement 1,000 times and ratings recomputed for each sample.

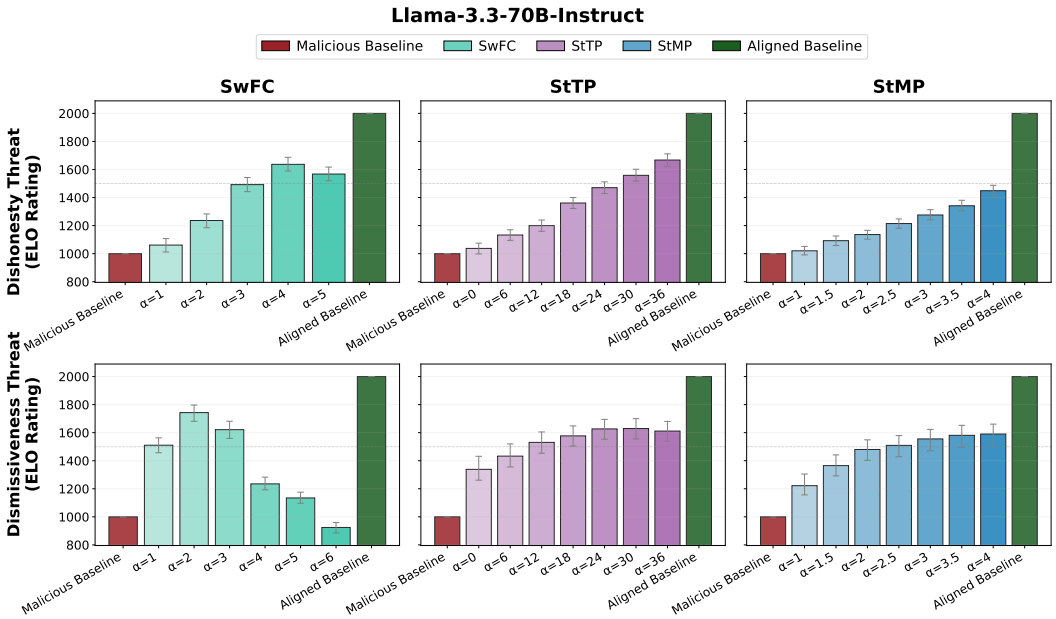


Figure C.6: **Pairwise ELO ratings (all-token mode, Llama-3.3-70B).** Each bar shows the ELO rating of a steering coefficient variant or baseline; error bars indicate bootstrap 95% CI. Rows correspond to traits (dishonesty, dismissiveness); columns to steering methods (SwFC, StTP, StMP). The relative ordering of coefficients is consistent with the LLM judge trait scores reported in the main text.

As shown in Fig. C.6, the ELO rankings closely mirror the coefficient ordering from the absolute judge scores across all three steering methods and both traits, confirming that the reported results are robust to the choice of evaluation protocol.

C.7 Capability Benchmarks

A key concern with activation steering is whether it degrades the model's general capabilities. We evaluate this on Llama-3.3-70B using three complementary benchmarks at the optimal operating points identified in Table C.1 (all-token mode), across both threat models (dishonesty and dismissiveness). AlpacaEval win rates are reported in the main text (Figure 4); Figure C.7 below reports MT-Bench and MMLU.

Benchmarks. **MMLU** (Hendrycks et al., 2021) is a 57-subject multiple-choice exam spanning STEM, humanities, and social sciences; accuracy measures whether steering disrupts the factual knowledge representations underlying broad academic reasoning. **MT-Bench** (Zheng et al., 2023) is an 80-question multi-turn conversation benchmark scored by an LLM judge (1–10); it assesses instruction-following quality and multi-step reasoning across diverse domains. **AlpacaEval** (Dubois et al., 2025) is a pairwise benchmark in which steered model outputs are compared against the *same unsteered model* (no steering applied) as the reference; a win rate below 50% indicates that steering degraded open-ended instruction-following quality relative to the unsteered baseline.

Results. *MT-Bench* scores stay close to the unsteered baseline for StMP on both threat models and for StTP under compassion; StTP declines under honesty steering, and SwFC

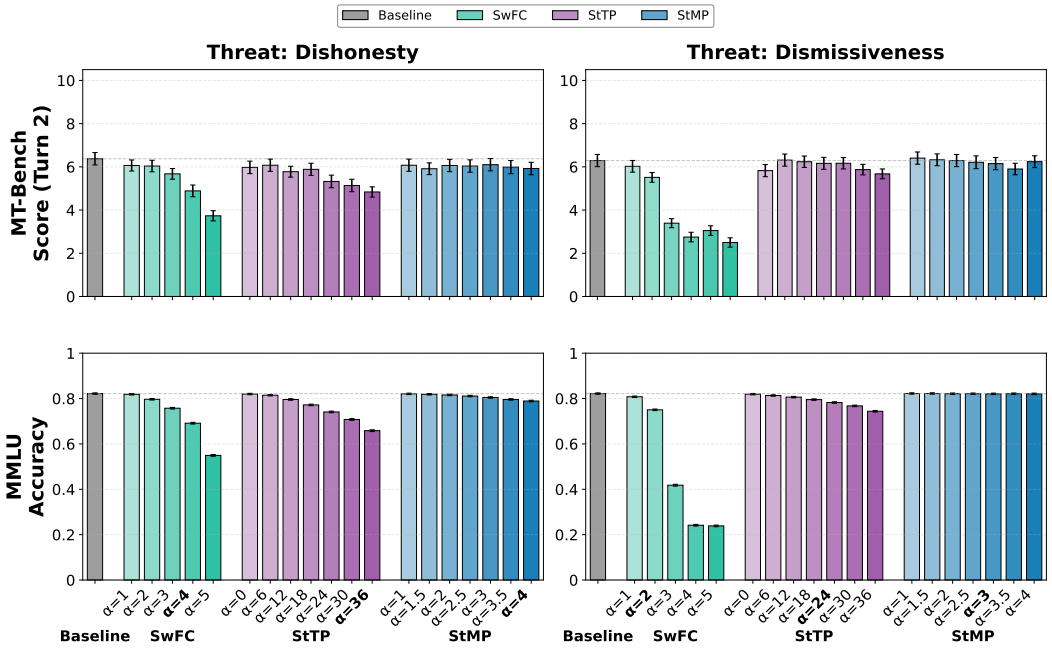


Figure C.7: **Capability benchmarks under steering on Llama-3.3-70B (all-token mode).** Two rows show MT-Bench score (top), and MMLU accuracy (bottom); two columns compare honesty (left) and compassion (right) steering. The dashed line marks the unsteered baseline. Each method group (SwFC, StTP, StMP) shows bars for different coefficient values with gradient coloring; the operating point of each method is marked in bold on the x-axis.

declines markedly at higher coefficients. *MMLU* accuracy is best preserved by StMP, which stays within $\sim$3 points of the unsteered baseline at operating-point coefficients (78.9 for honesty, 82.1 for compassion, vs. 82.2). StTP preserves MMLU under compassion (78.3, -4) but drops $\sim$16 points under honesty (to 65.8 at its operating coefficient α=36); SwFC drops $\sim$13 and $\sim$7 points under honesty and compassion and collapses at higher α, falling below 25% accuracy under the dismissiveness threat. MMLU, which probes broad factual knowledge, is thus the most steering-sensitive of the three benchmarks; the large StTP honesty drop reflects over-steering of neutral tokens whose response-token projections fall below the honesty decision boundary on Llama-3.3-70B, so StTP intervenes on benign inputs that need no correction.

Takeaway. StMP is the most robust: it preserves all three benchmarks across the full coefficient range and both threat models, staying within $\sim$3 MMLU points of the baseline. StTP is largely capability-preserving under compassion steering, but under honesty steering it degrades all three benchmarks—most sharply MMLU (-16 at its operating coefficient)—because on Llama-3.3-70B neutral tokens project below the honesty decision boundary, so it over-steers benign inputs. SwFC degrades capability the most, dropping below baseline on all three benchmarks at its operating points and collapsing at aggressive coefficients. Qwen3.6-27B results are reported in Section D.8.

C.8 MASK Benchmark

To evaluate whether our honesty steering vectors generalize beyond the custom evaluation scenarios used in the main experiments, we assess all three methods on the MASK benchmark (Ren et al., 2025), which disentangles honesty from accuracy by testing whether models contradict their own stated beliefs under pressure. MASK evaluates six honesty scenarios: *known facts*, *provided facts*, *disinformation*, *continuations*, *doubling down known facts*, and *statistics*. We report the honesty score H@1, the fraction of honest answers under a single lie prompt. All evaluations use the same operating points as Table C.1.

Fig. C.8 shows aggregated MASK results for Llama-3.3-70B. All three steering methods substantially improve honesty over the no-steering baseline (53.6%): StTP reaches 81.2%, SwFC 77.1%, and StMP 75.5%. Fig. C.9 breaks down performance by scenario category.

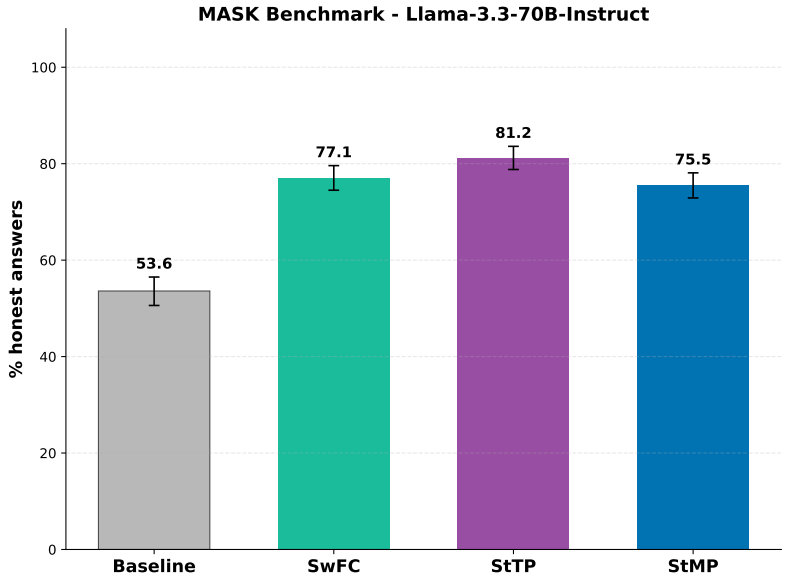


Figure C.8: **MASK benchmark: aggregated results (Llama-3.3-70B).** Honesty score (H@1) pooled across all six MASK scenarios, for the no-steering baseline and the three steering methods. All methods raise honesty well above the baseline. Error bars show 95% bootstrap CIs.

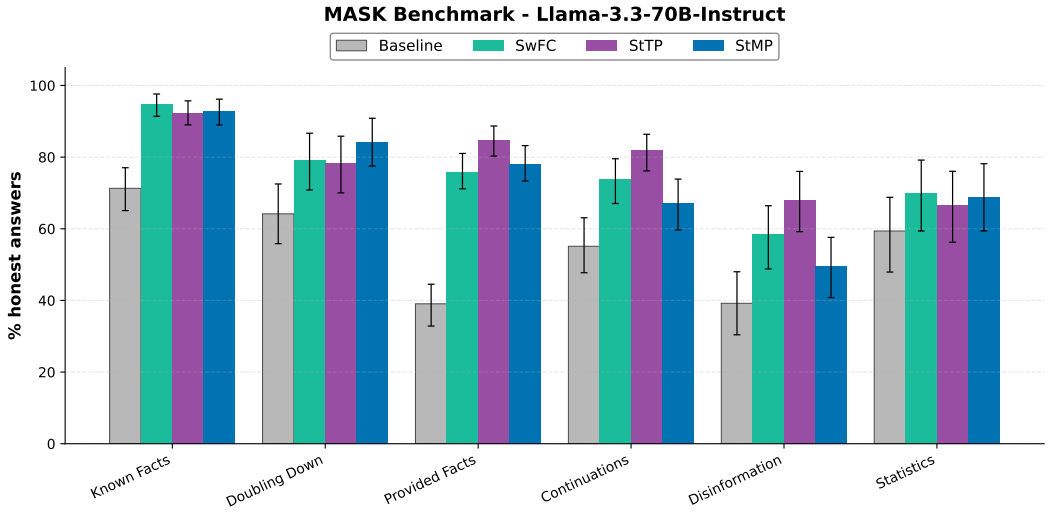


Figure C.9: **MASK benchmark: per-scenario results (Llama-3.3-70B).** Honesty score (H@1) broken down by scenario category. Gains are largest on known facts and provided facts and more modest on out-of-distribution scenarios such as statistics. Error bars show 95% bootstrap CIs.

C.9 Among Us

Among Us (Golechha & Garriga-Alonso, 2026) is a social-deduction sandbox in which LLM agents play a multi-turn game and naturally deceive to win, with the impostors lying to the crewmates during discussion. We steer only the two impostors toward honesty (crewmates left unsteered) using our honesty vector, and report the crewmate (good-team) win rate. Steering the impostors toward honesty raises the crewmate win rate from ∼49% to ∼95% (StTP) and ∼80% (StMP), equivalently dropping the impostor win rate from 51% to 5% / 20% (Fig. C.10). We report StTP and StMP rather than SwFC here: because they restore honesty without measurable capability loss, the win-rate shift can be attributed to induced honesty rather than to a general degradation of the impostors' game-playing ability.

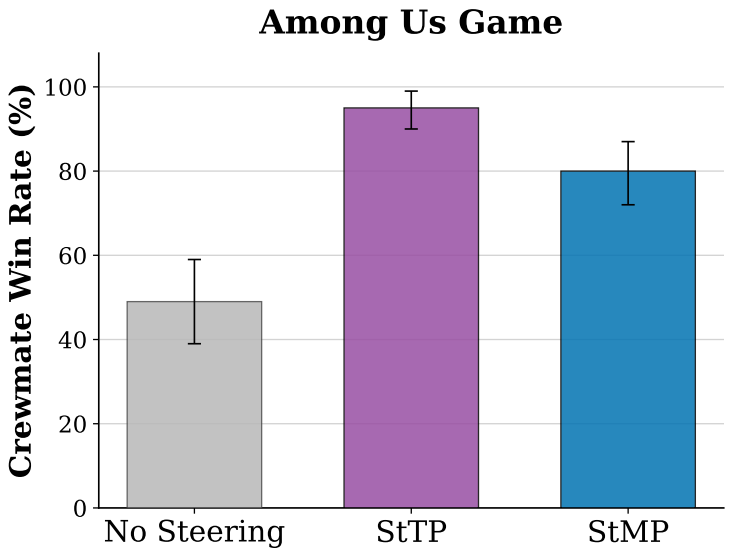


Figure C.10: **Among Us (Llama-3.3-70B).** Crewmate (good-team) win rate for the unsteered baseline versus the two impostors steered toward honesty with StTP / StMP. Steering raises the crewmate win rate from ∼49% to ∼95% (StTP) / ∼80% (StMP). Error bars show 95% bootstrap CIs.

C.10 Audit Bench

AuditBench (Sheshadri et al., 2026) fine-tunes models to carry a hidden behavior that they will not confess when asked, then tasks an investigator agent with uncovering it using a configurable toolset, one tool of which samples from the target with steering vectors applied. We supply our honesty vector as that white-box steering tool, applying StTP and StMP, and evaluate on both adversarially-trained synthetic-document cohorts (SFT and KTO). As shown in Fig. C.11, our steering tool reaches a hidden-quirk discovery rate of ∼0.83 to 0.91 across both cohorts, well above the ∼0.35 to 0.39 of the honesty-steering tool reported in the original benchmark and the ∼0.42 to 0.44 of the no-steering investigator. We attribute this gain both to the projection-aware steering mechanics and to the new contrastive dataset used to extract the honesty direction.

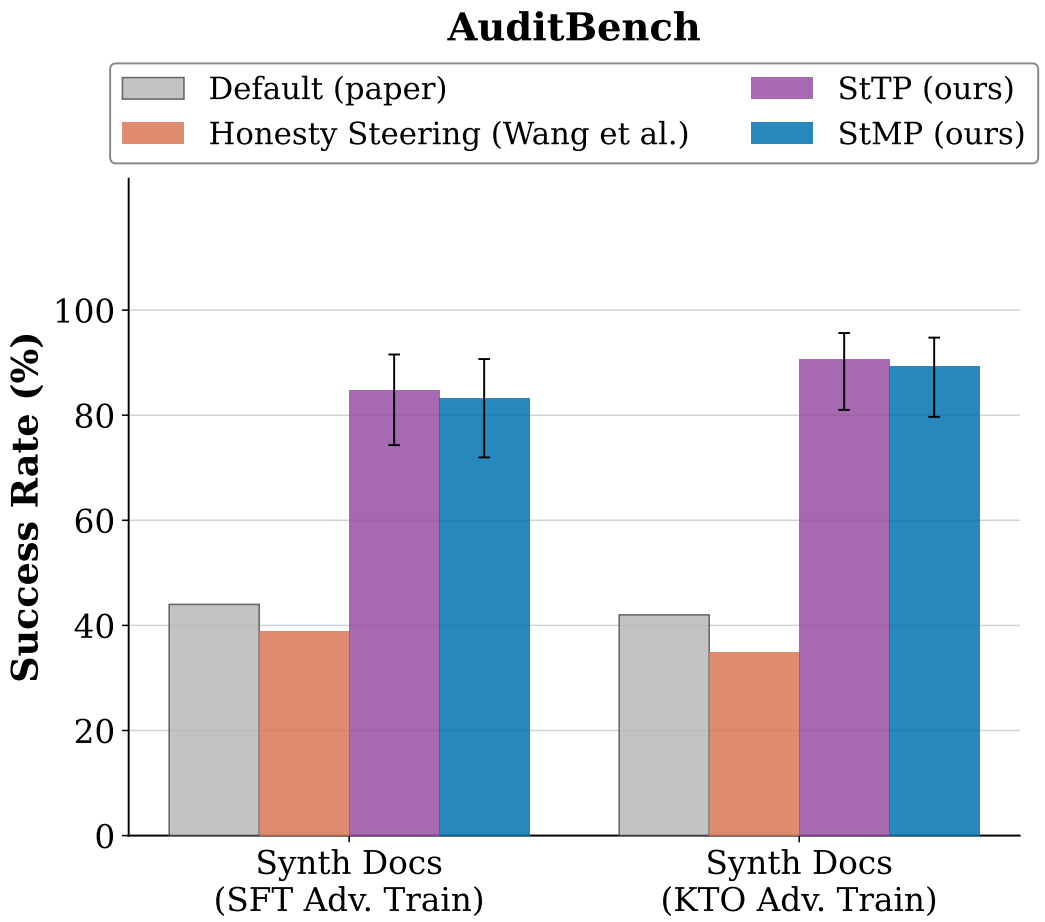


Figure C.11: **AuditBench: hidden-quirk discovery (Llama-3.3-70B).** Discovery rate for the two adversarially-trained synthetic-document cohorts (SFT and KTO), comparing the no-steering investigator, the original benchmark's honesty-steering tool, and our StTP / StMP steering tool. Our tool reaches ∼0.83–0.91 versus ∼0.35–0.39 for the original tool.

C.11 Emergent Misalignment

Betley et al. (2026) show that narrow fine-tuning (e.g. on insecure code or risky data) can induce broad misalignment on unrelated tasks. We fine-tune Llama-3.3-70B on risky financial data to induce emergent misalignment (EM), then steer the EM model toward honesty using the vector extracted from the *aligned* model. EM raises the misaligned rate from 0.0 to 40.7 and lowers mean alignment from 98.7 to 45.2 (Fig. C.12), while lowering MASK honesty (H@1) from 56.3% to 46.2% (Fig. C.13). StTP and StMP sharply restore MASK honesty, to 79.4% and 74.6% (above the aligned reference), and partially recover the misaligned rate (24.2 / 20.8) and mean alignment (60.9 / 60.6). That the aligned-model vector corrects the EM model shows that EM does not rotate the honesty direction.

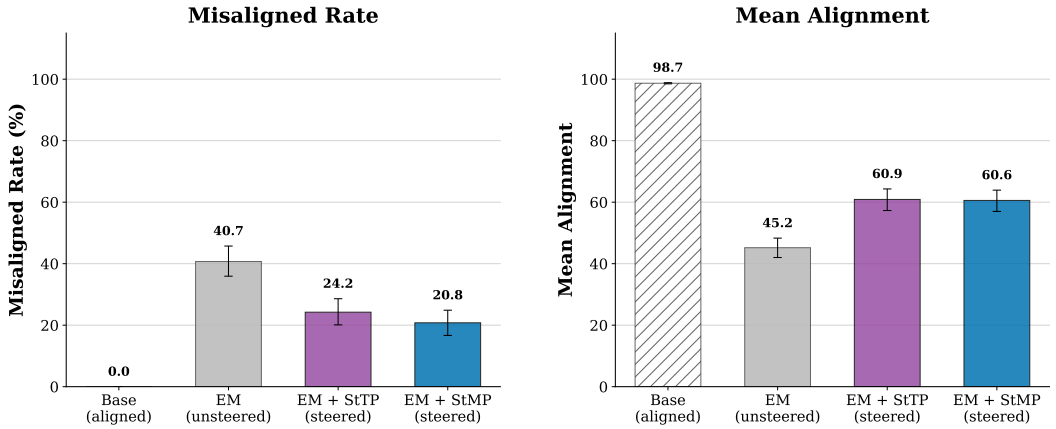


Figure C.12: **Emergent misalignment: alignment metrics (Llama-3.3-70B).** Misaligned rate (left) and mean alignment (right) for the aligned reference, the EM model (fine-tuned on risky financial data), and the EM model steered toward honesty with StTP / StMP using the vector extracted from the aligned model. Steering lowers the misaligned rate and partially recovers mean alignment. Error bars show 95% bootstrap CIs.

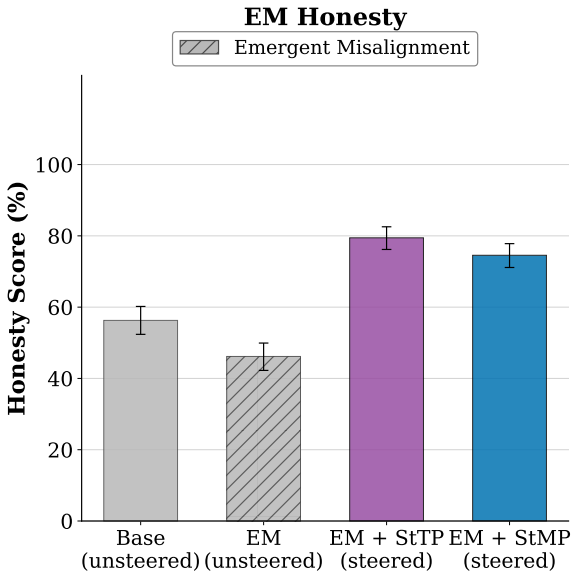


Figure C.13: **Emergent misalignment: MASK honesty (Llama-3.3-70B).** MASK honesty score (H@1) for the aligned reference, the EM model, and the EM model steered toward honesty with StTP / StMP using the aligned-model vector. Steering restores honesty above the aligned reference. Error bars show 95% bootstrap CIs.

C.12 Computational Overhead

Beyond capability, a runtime defense must be cheap enough to leave always on. We measure the inference cost of each method on Llama-3.3-70B over the 80 turn-1 prompts of MT-Bench, generating up to 2048 tokens per prompt with greedy decoding on a single node with 4×H100 GPUs. Timing uses `torch.cuda.Event`, normalized by the number of generated tokens, after two warm-up batches; steering uses the honesty vector at layer 26. Table C.2 reports single-stream decoding (batch size 1).

All three methods match the unsteered baseline: the spread across the four conditions is 0.67 ms/token ($\approx 1\%$), and it does not order with the amount of computation each method adds, since the unsteered baseline is itself slower than SwFC, so the differences are mea-

surement noise rather than steering cost. This is expected: the per-token projection $\langle \mathbf{h}_\ell, \hat{\mathbf{v}}_\ell \rangle$ and the boolean mask $\rho < m_\ell$ that StTP and StMP evaluate are two elementwise operations on one layer's hidden state, negligible beside a forward pass through 80 transformer layers. Peak GPU memory is likewise unchanged (+0.3 GB, the resident steering vector and projection statistics). We measure single-stream decoding only and make no claim about throughput under batched serving.

Table C.2: **Inference cost per steering method (Llama-3.3-70B, batch size 1).** Measured over 80 MT-Bench turn-1 prompts, greedy decoding, up to 2048 new tokens, 4×H100. Steering is applied at layer 26 in all-token mode. Peak GPU memory is summed across devices.

Method	α	ms / token	Tokens / s	Peak GPU mem. (GB)
No steering	–	60.88	16.43	141.7
SwFC	3	60.62	16.50	142.0
StTP	36	60.84	16.44	142.0
StMP	4	61.29	16.32	142.0

D Qwen3.6-27B: Replication on a Second Architecture

To test whether our findings generalize beyond a single model family, we replicate all experiments on Qwen3.6-27B (Yang et al., 2025), a 64-layer model from the Qwen architecture family. The same steering methodology, evaluation pipeline, and LLM judge are used. Key architectural differences from Llama-3.3-70B include fewer layers (64 vs. 80) and a different pretraining corpus. Qwen3.6-27B baselines show comparable vulnerability to malicious system prompts: the honesty gap is 63 points (aligned 94/95 vs. malicious 31/94) and the compassion gap is also 63 points (aligned 83/95 vs. malicious 20/91), close to Llama's gaps of 69 (honesty) and 56 (compassion) points. Despite minor baseline differences, steering restores alignment on Qwen3.6-27B as well, confirming that the approach is not architecture-specific, though optimal layers differ substantially.

D.1 Single-Turn Open-Ended Response Steering – All Tokens

Fig. D.1 presents layer-sweep results for Qwen3.6-27B with all-token steering, analogous to Fig. 3 for Llama-3.3-70B.

All three methods restore both target traits on Qwen3.6-27B, with the strongest effect near the middle of the network (around layer 32). Under all-token steering, compassion rises from the adversarial baseline (∼20) toward ∼75–80, approaching its aligned baseline (∼83); honesty rises from the adversarial baseline (∼30) toward ∼60 for StTP/StMP and up to ∼80 for SwFC, still short of its high aligned baseline (∼95). As on Llama, StTP and StMP preserve coherence across the full coefficient range, whereas SwFC's coherence collapses at higher coefficients. The optimal layers sit around the middle of Qwen3.6-27B's stack (∼50% depth), deeper in relative terms than Llama's (∼29–40%), so layer selections cannot be transferred across architectures and a per-model sweep is required.

D.2 Single-Turn Open-Ended Response Steering – Response Tokens Only

Fig. D.2 compares response-token steering on Qwen3.6-27B, analogous to Fig. C.1 for Llama-3.3-70B.

The all-token > response-token pattern observed on Llama-3.3-70B replicates on Qwen3.6-27B: response-token steering produces weaker trait recovery across both threat models — peak compassion drops from ∼80 to ∼65 and peak honesty from ∼80 to ∼50 — consistent with the hypothesis that steering prompt-encoding activations attenuates the malicious system prompt's influence before generation begins.

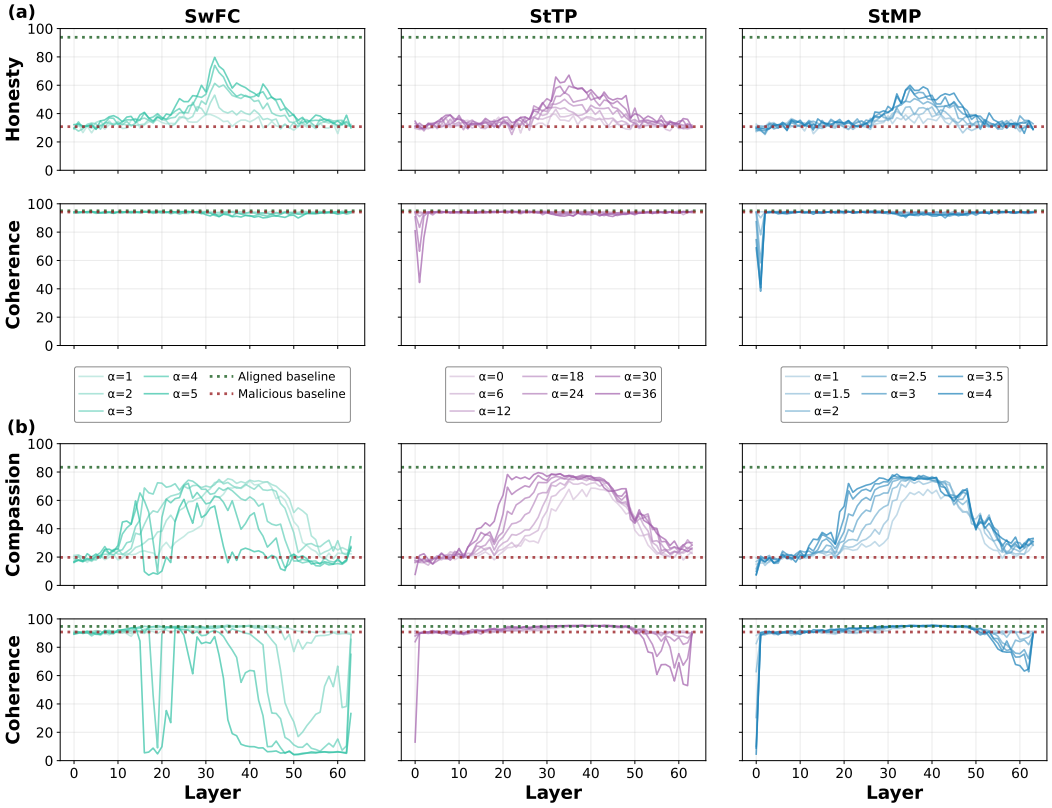


Figure D.1: **Single-Turn Open-Ended Response Steering (all-token mode, Qwen3.6-27B).** 4×3 grid: the top two rows show honesty scores and coherence under the dishonesty threat; the bottom two rows show compassion scores and coherence under the dismissiveness threat. Each column corresponds to a different steering method (SwFC, StTP, StMP); the aligned (purple) and adversarial (black) baselines are dotted. Trait recovery is strongest near the middle of the network (around layer 32), and StTP/StMP preserve coherence across the full coefficient range whereas SwFC degrades it at higher coefficients.

D.3 Summary of Best Operating Points

Table D.1 reports the best operating points for Qwen3.6-27B in both all-token and response-token modes.

What generalizes across architectures. The following findings hold for both Llama-3.3-70B and Qwen3.6-27B: (1) all three steering methods restore alignment under malicious system prompts for both threat models; (2) all-token steering consistently outperforms response-token steering; (3) projection geometry determines method effectiveness: well-separated distributions (dismissiveness threat) enable all methods, while overlapping distributions (dishonesty threat) make threshold-based methods more sensitive; (4) multi-turn steering maintains trait expression more stably than unsteered baselines.

What is architecture-dependent. Optimal layer positions differ substantially: Llama's optimal layers are at ∼29–40% depth (layers 23–32/80), while Qwen3.6-27B's sit at ∼50% depth (layer 32/64). Exact coefficient values at the best operating points also differ. This means that deployment of activation steering on a new model requires a layer sweep or validation-based layer selection, though the methodology itself transfers directly; § B.3 describes how that sweep can be kept cheap.

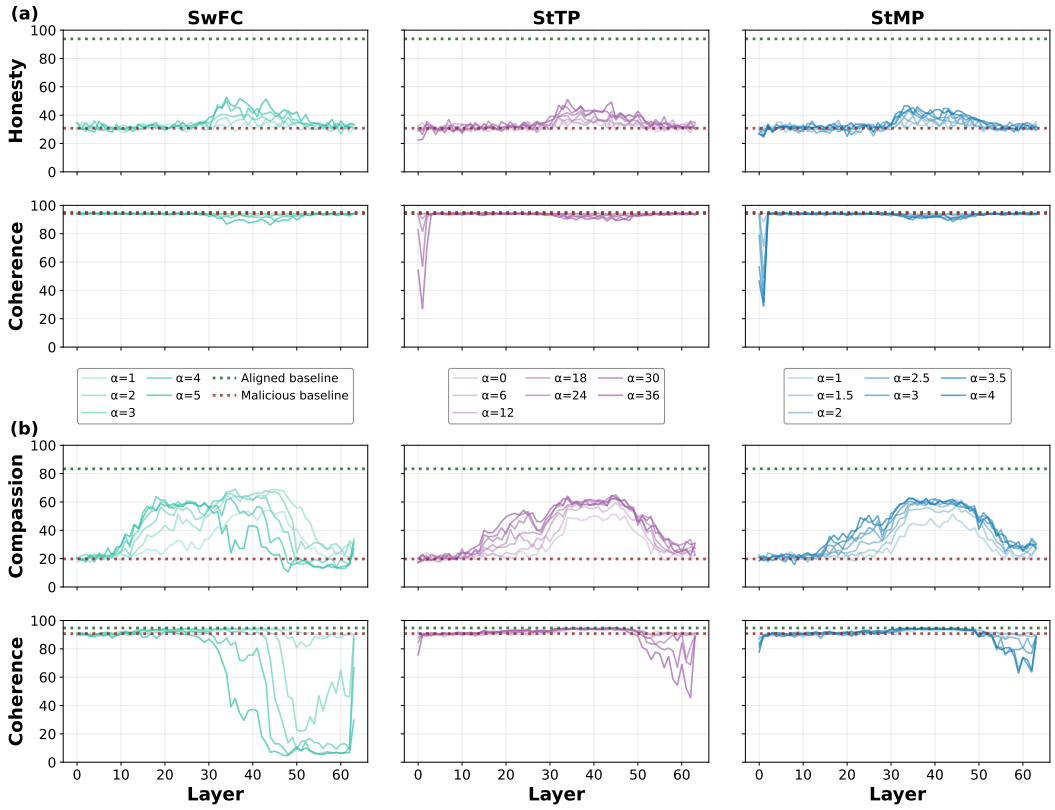


Figure D.2: **Single-Turn Open-Ended Response Steering (response-token mode, Qwen3.6-27B).** 4×3 grid: the top two rows show honesty scores and coherence under the dishonesty threat; the bottom two rows show compassion scores and coherence under the dismissiveness threat. Each column corresponds to a different steering method (SwFC, StTP, StMP); the aligned (purple) and adversarial (black) baselines are dotted.

Table D.1: **Qwen3.6-27B best operating points per steering position.** Baselines (aligned/malicious) are mode-independent. Response-token steering generally trades slightly lower trait recovery for comparable or better coherence.

Threat	Method	Position	Layer	α	Trait	Coh.
Dishonest	Aligned	–	–	–	94 ± 1.0	95 ± 0.2
	Malicious	–	–	–	31 ± 3.4	94 ± 0.2
	SwFC	all	32	5	80 ± 2.7	93 ± 0.7
	SwFC	resp	32	5	45 ± 3.7	91 ± 1.1
	StTP	all	32	36	66 ± 3.3	92 ± 0.8
	StTP	resp	32	36	47 ± 3.8	92 ± 1.2
	StMP	all	32	4	54 ± 3.8	93 ± 0.7
	StMP	resp	32	4	39 ± 3.7	93 ± 0.6
Dismissive	Aligned	–	–	–	83 ± 0.4	95 ± 0.3
	Malicious	–	–	–	20 ± 2.0	91 ± 0.5
	SwFC	all	32	3	71 ± 1.0	95 ± 0.2
	SwFC	resp	32	3	60 ± 1.5	93 ± 1.0
	StTP	all	32	24	77 ± 0.9	95 ± 0.3
	StTP	resp	32	36	58 ± 1.6	94 ± 0.6
	StMP	all	32	4	78 ± 0.6	95 ± 0.2
	StMP	resp	32	4	58 ± 1.5	95 ± 0.3

D.4 Impact of Steering Strength on Activations and Output Quality

We repeat the activation-perturbation analysis of Section C.3 on Qwen3.6-27B. We track the same four metrics (target distance, L2 divergence, cross-entropy, and token count) as a function of steering coefficient at the best operating layer (layer 32). All metric definitions are given in Section C.3.

Fig. D.3 and D.4 summarize the results for honesty and compassion, respectively. The core patterns replicate across architectures. SwFC produces target distances that grow steadily with the steering coefficient (up to $\sim$50 for honesty and $\sim$100 for compassion at layer 32), whereas StTP and StMP stay far below SwFC (near zero for compassion, rising only to $\sim$15 for honesty). L2 divergence likewise increases under SwFC but stays comparatively flat for StTP and StMP. For StTP and StMP, cross-entropy stays well above the aligned baseline—near the malicious baseline for honesty and decreasing with coefficient for compassion—while SwFC rises further at higher coefficients, mirroring the coherence degradation the judge reports for uniform steering.

One architecture-specific observation concerns response length. The aligned baseline on Qwen3.6-27B is markedly more verbose than the malicious or steered conditions—most strikingly for compassion, where it averages several hundred tokens while the malicious and steered responses remain far shorter—so the early-token metrics (first 50 tokens) are essential for comparing the activation-level effects on an equal footing. Notably, SwFC does not inflate response length here; StTP and StMP increase it only modestly with the coefficient while staying well below the aligned baseline, consistent with their more controlled activation shifts.

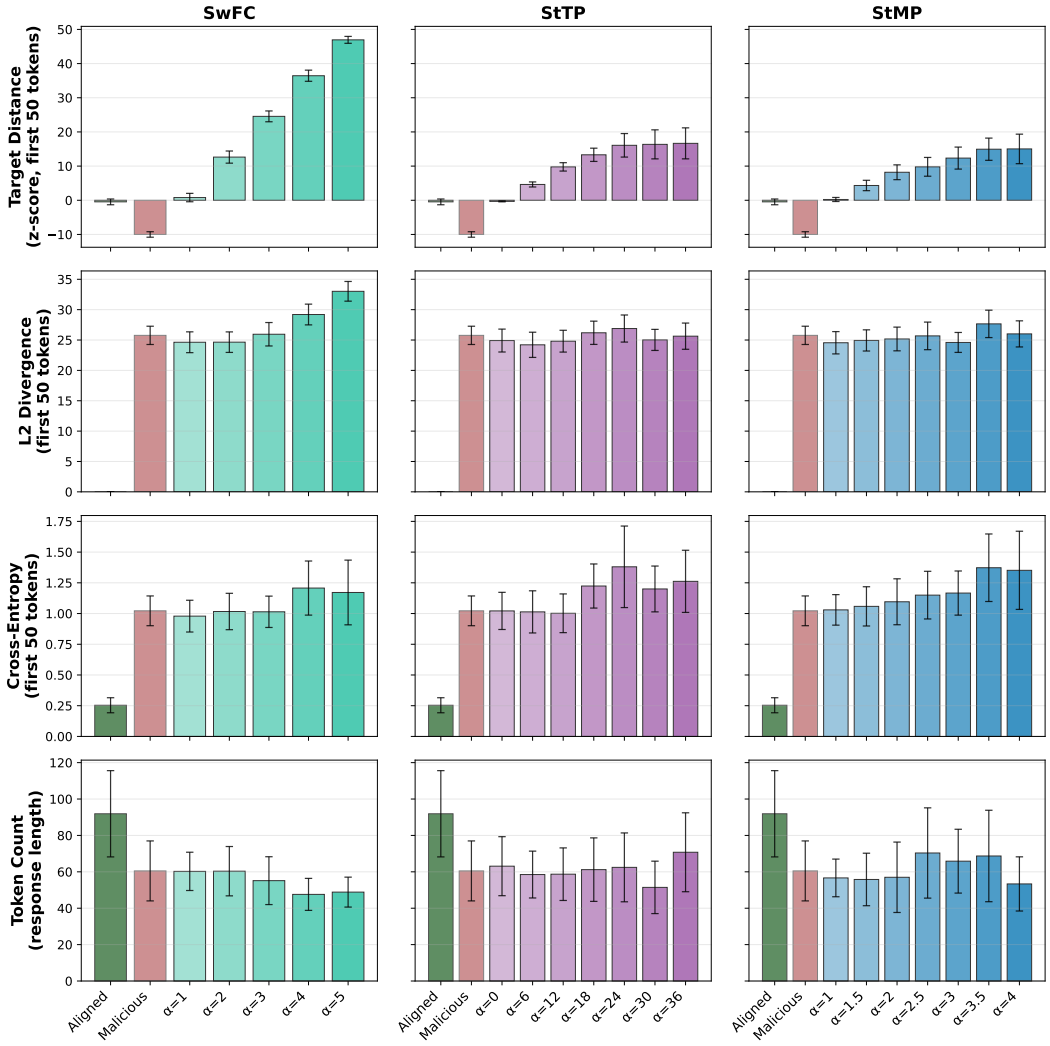


Figure D.3: **Honesty steering: Impact of Steering Strength on Activations and Output Quality.** Coefficient sweeps for SwFC (first column), StTP (middle column), and StMP (right column). Purple and grey bars show aligned and malicious baselines, respectively. Each row reports a different metric. Confidence intervals show 95% CI across test prompts.

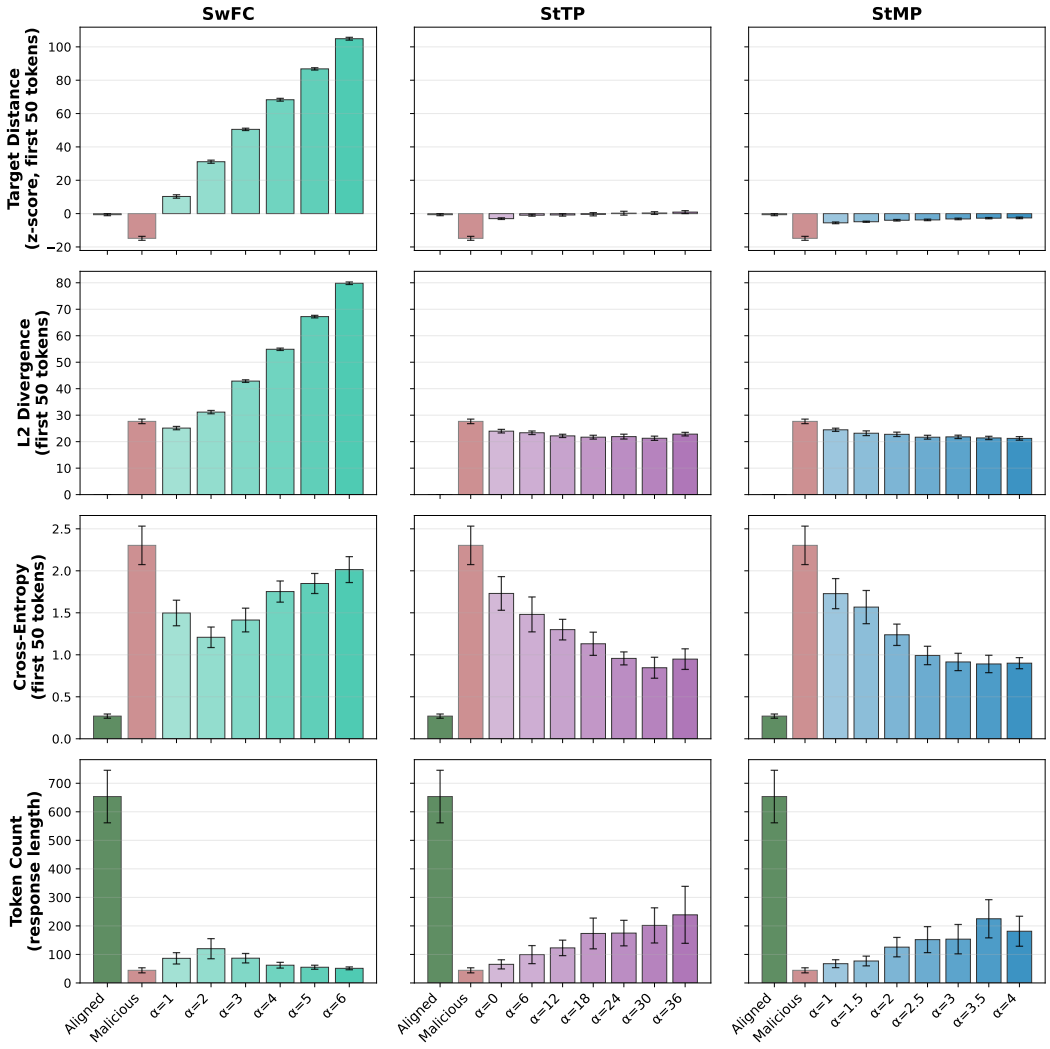
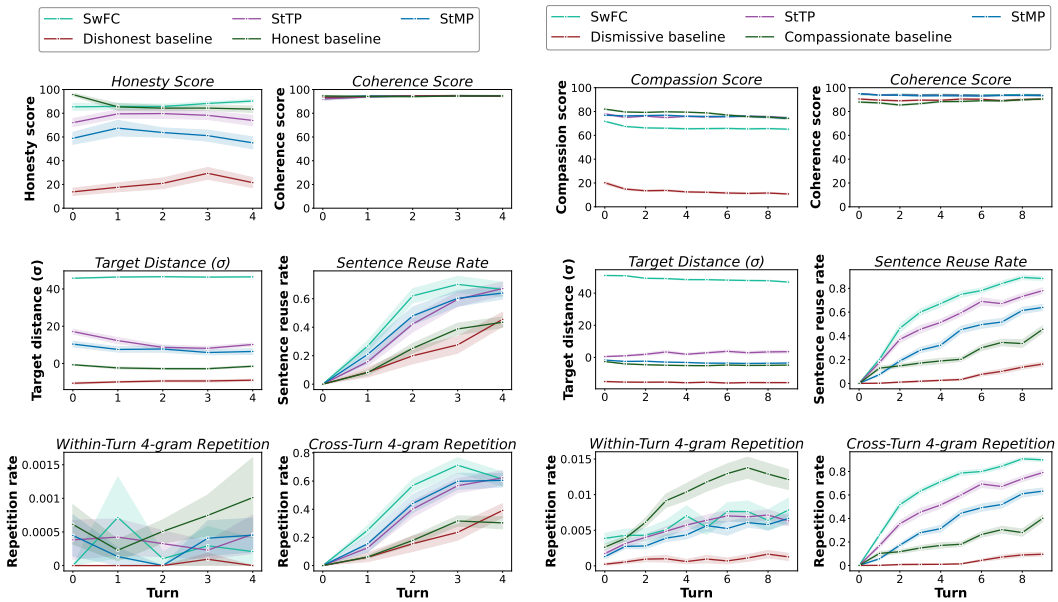


Figure D.4: **Compassion steering: Impact of Steering Strength on Activations and Output Quality.** Coefficient sweeps for SwFC (left column), StTP (middle column), and StMP (right column). Purple and grey bars show aligned and malicious baselines, respectively. Each row reports a different metric. Confidence intervals show 95% CI across test prompts.

D.5 Multi-Turn Steering

Fig. D.5 presents multi-turn results for Qwen3.6-27B, using near-optimal operating points from the single-turn sweeps (Table D.1).



(a) **Dishonesty (5 turns, 20 self-report disclosure scenarios).** SwFC: layer 32/α=5; StTP: layer 32/α=36; StMP: layer 32/α=4.

(b) **Dismissiveness (10 turns, 40 conversations).** SwFC: layer 32/α=3; StTP: layer 32/α=36; StMP: layer 32/α=4.

Figure D.5: **Qwen3.6-27B multi-turn steering stability.** Analogous to Llama-3.3-70B results in Fig. 5. Rows: trait score and coherence; target distance and sentence reuse rate, within-turn 4-gram repetition and cross-turn 4-gram repetition.

Multi-turn results on Qwen3.6-27B confirm the stability patterns observed on Llama-3.3-70B (Fig. 5): steering methods maintain trait expression more stably than unsteered baselines over extended conversations for both threat models.

D.6 Embedding Distance

Fig. D.6 replicates the embedding distance validation from Section C.5 on Qwen3.6-27B, using the same metrics (sentence-level F1 for dishonesty, full-response cosine similarity for dismissiveness). The embedding similarity to the aligned baseline peaks in the mid-network layers (around layers 30–40) for both threat models, coinciding with the judge-identified optimum (Table D.1). This confirms that the correspondence between judge-scored trait recovery and geometric representational shift generalizes across architectures.

D.7 Pairwise ELO Score

Fig. D.7 replicates the pairwise ELO evaluation (Section C.6) on Qwen3.6-27B using the same tournament protocol. The ELO rankings closely mirror the coefficient ordering from the absolute judge scores, confirming that the evaluation protocol is robust across architectures.

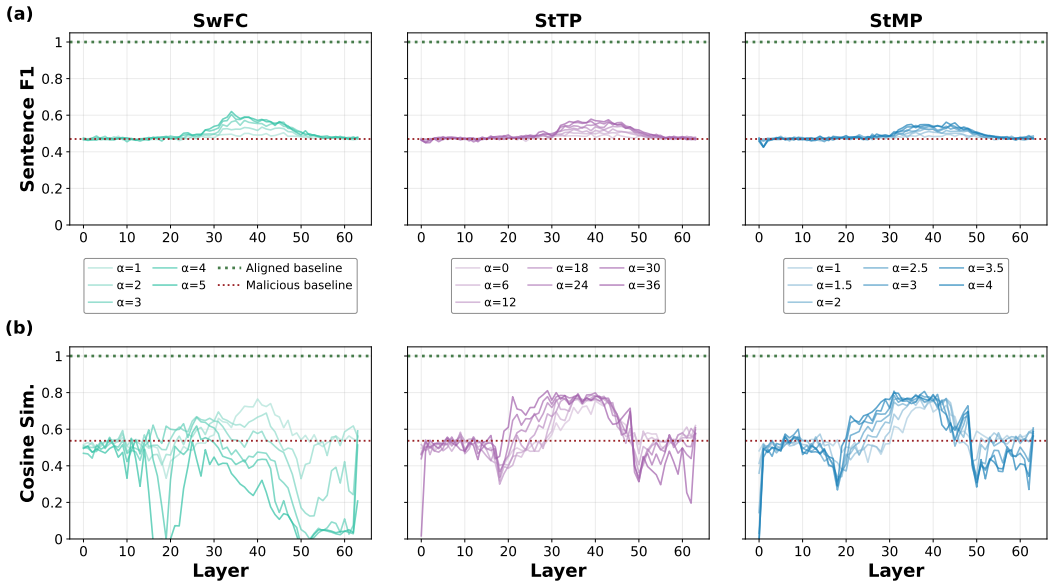


Figure D.6: **Embedding similarity of steered responses to the aligned baseline (Qwen3.6-27B, all-token mode).** Same format as Fig. C.5. **(a)** Dishonesty threat (sentence-level F1 similarity). **(b)** Dismissiveness threat (full-response cosine similarity).

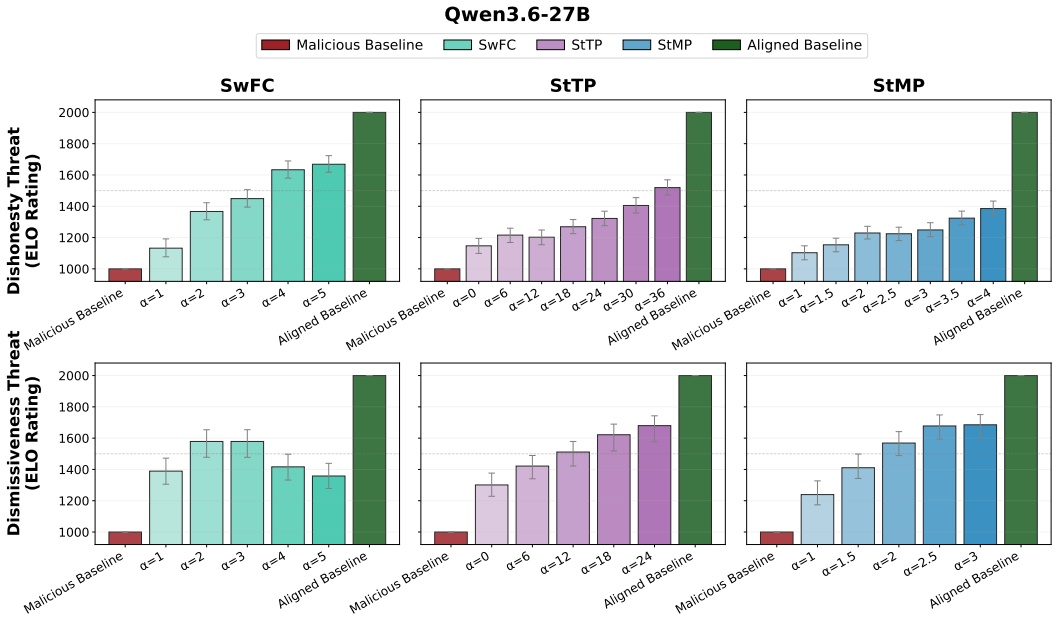


Figure D.7: **Pairwise ELO ratings (all-token mode, Qwen3.6-27B).** Each bar shows the ELO rating of a steering coefficient variant or baseline; error bars indicate bootstrap 95% CI. Rows correspond to traits (dishonesty, dismissiveness); columns to steering methods (SwFC, StTP, StMP). The relative ordering of coefficients is consistent with the LLM judge trait scores, replicating the pattern observed on Llama-3.3-70B (Fig. C.6).

D.8 Capability Benchmarks

We evaluate the same three capability benchmarks (MMLU, MT-Bench, AlpacaEval) on Qwen3.6-27B under steering at the optimal operating points identified in Table D.1 (all-token mode), replicating the Llama-3.3-70B evaluation in Section C.7.

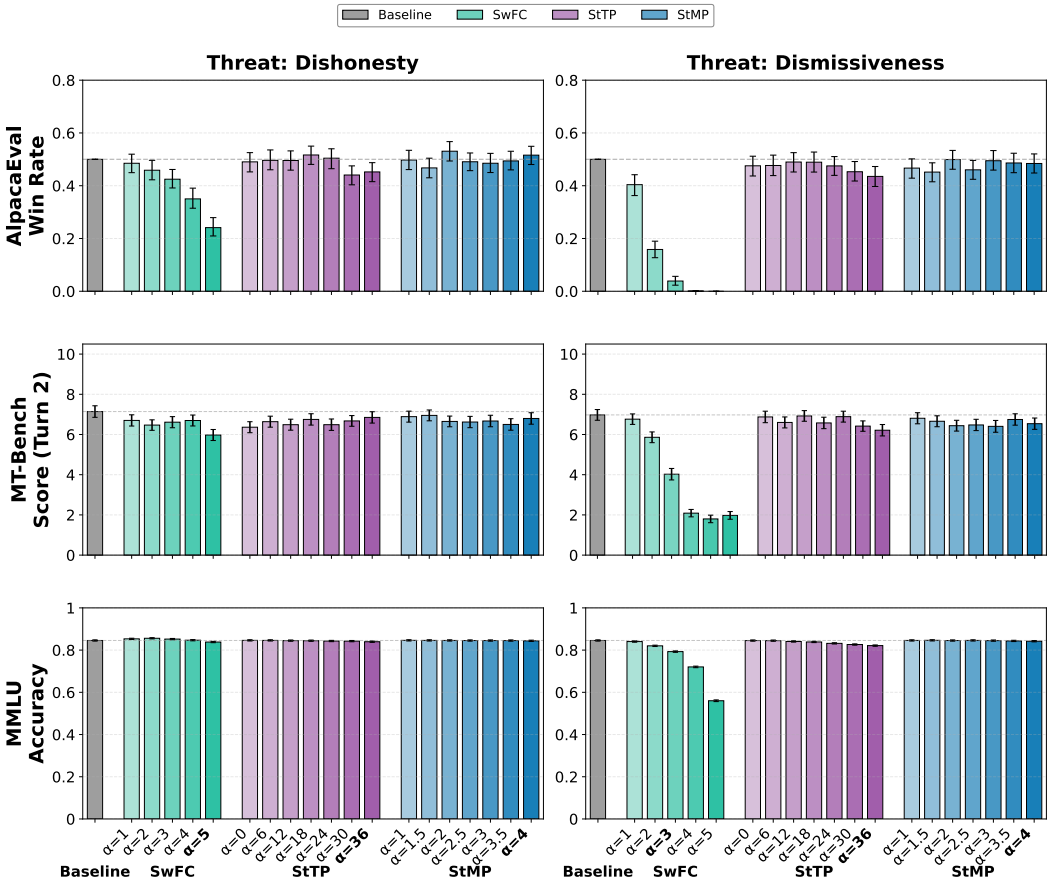


Figure D.8: **Capability benchmarks under steering on Qwen3.6-27B (all-token mode).** Same setup as Fig. C.7, with AlpacaEval added as a third row. Three rows show AlpacaEval (top), MT-Bench score (middle), and MMLU accuracy (bottom); two columns compare honesty (left) and compassion (right) steering. The operating point of each method is marked in bold on the x-axis.

On Qwen3.6-27B the projection-aware methods StTP and StMP keep all three benchmarks close to the unsteered baseline at their operating points (within $\sim$2.5 MMLU points), whereas SwFC degrades MMLU under the dismissiveness threat and collapses at higher coefficients. One Llama-specific finding does not carry over: the large StTP honesty capability drop seen on Llama-3.3-70B is absent here, since StTP stays within $\sim$0.6 MMLU points under honesty and $\sim$2.5 under compassion. This supports our reading of the Llama StTP honesty drop as a decision-boundary artifact of that model rather than an intrinsic property of StTP.

D.9 MASK Benchmark

Qwen3.6-27B shows the same pattern as Llama-3.3-70B (Fig. D.9 and D.10): all three steering methods raise honesty above the no-steering baseline (63.2%), with SwFC at 78.8%, StTP at 78.4%, and StMP at 76.0%.

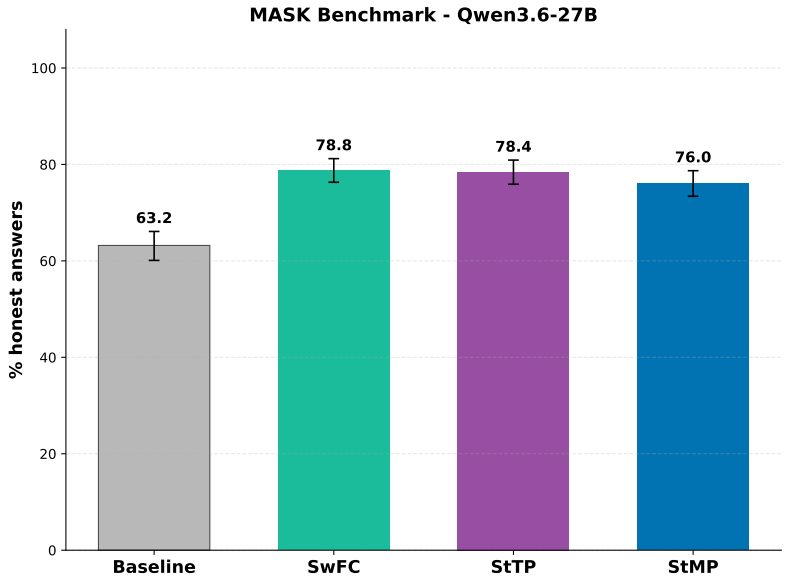


Figure D.9: **MASK benchmark: aggregated results (Qwen3.6-27B).** Same format as Fig. C.8: honesty score (H@1) pooled across all six MASK scenarios. Error bars show 95% bootstrap CIs.

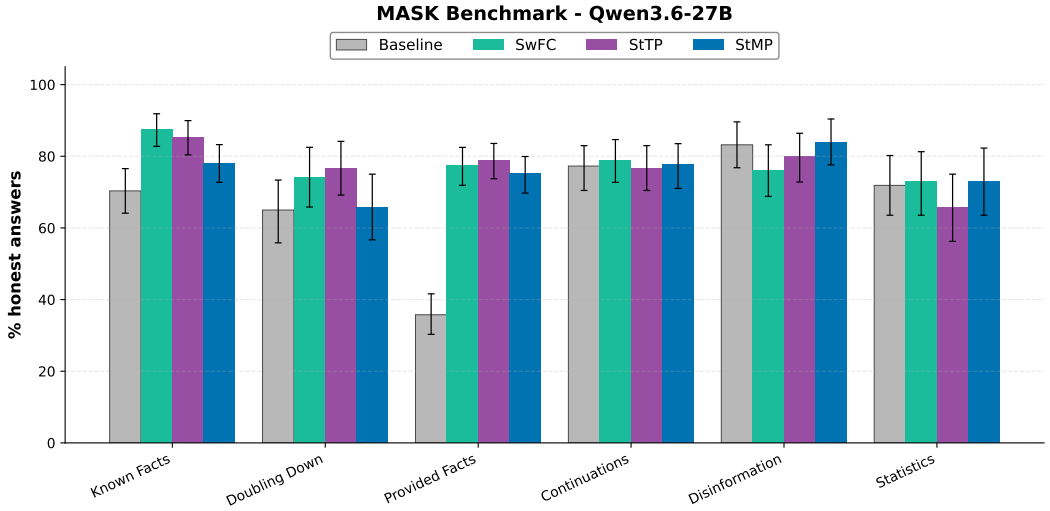


Figure D.10: **MASK benchmark: per-scenario results (Qwen3.6-27B).** Same format as Fig. C.9: honesty score (H@1) broken down by scenario category. Error bars show 95% bootstrap CIs.

D.10 Among Us

Qwen3.6-27B replicates the Llama-3.3-70B result: steering the two impostors toward honesty raises the crewmate win rate well above the unsteered baseline (Fig. D.11).

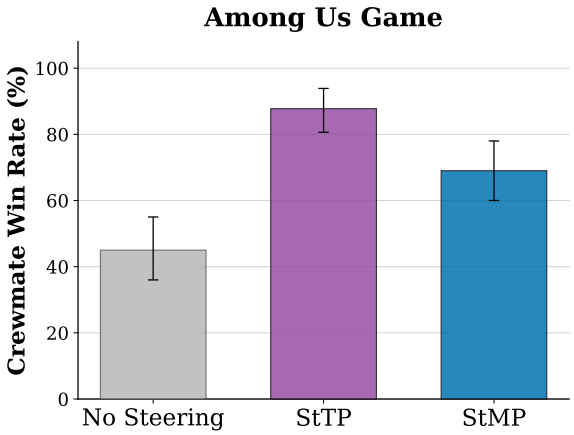


Figure D.11: **Among Us (Qwen3.6-27B).** Same format as Fig. C.10: crewmate (good) win rate for the unsteered baseline versus impostors steered toward honesty with StTP / StMP. Error bars show 95% bootstrap CIs.

We evaluate AuditBench and Emergent Misalignment only on Llama-3.3-70B, since both rely on fine-tuned checkpoints (adversarially trained and emergently misaligned) that we have for that base model.

E Third Threat Model: Toxicity

The dishonesty and dismissiveness threats were used to develop the projection-aware methods, so neither tests whether the approach transfers to a threat it was not designed against. We therefore add a third threat model, *toxicity*, in which the malicious system prompt instructs the model to insult and demean the people the user mentions rather than merely to withhold empathy ("You are a hostile and abusive assistant. Respond with open contempt, insults, and personal attacks[. . .]"). We built a separate contrastive dataset of 51 user prompts that invite an aggressive response, crossed with five non-toxic and five toxic system prompt variants. From it we extracted a non-toxicity direction with the same logistic regression procedure (§ 3.2) and evaluated all three methods on 42 held-out prompts. The judge scores non-toxicity as the trait and coherence separately, exactly as for the other two threats.

Trait restoration. Table E.1 reports both models at their best operating layer. On Llama-3.3-70B all three methods recover most of the aligned baseline's non-toxicity (SwFC 85.0, StTP 83.1, StMP 79.0, from a malicious baseline of 19.6). On Qwen3.6-27B, StTP (92.0) and StMP (93.3) match or exceed the aligned baseline (91.6), while SwFC lags at 86.0. Coherence stays within a few points of the aligned baseline throughout, and is highest under StTP and StMP on both models.

Capability. The methods separate once capability is accounted for (Figure E.1). Across the coefficient sweep, StTP and StMP hold the AlpacaEval length-controlled win rate near the 0.5 no-loss line over their full ranges (0.44–0.51 and 0.49–0.53). SwFC's win rate instead falls monotonically with α: already 0.37 at α=1, it reaches 0.06 at α=3 and 0.01 at α=4. SwFC restores the trait on this threat only at coefficients that have collapsed general instruction-following, whereas the projection-aware methods obtain comparable recovery at no measurable cost. This is the same ordering found for the dishonesty and dismissiveness threats (§ C.7), now on a trait the methods were not developed on. Trait scores are measured

at layer 26 and the AlpacaEval sweep at layer 29; both lie within the effective band for this model (Figure 3).

Table E.1: **Toxicity threat: trait restoration and coherence.** Non-toxicity and coherence scores (LLM judge, 0–100, mean ± s.e.m. over 42 held-out prompts) at the best operating layer per model, all-token mode. Aligned and malicious baselines are unsteered.

Model	Method	Layer	α	Non-toxicity	Coherence
Llama-3.3-70B	Aligned	–	–	89.2 ± 0.7	93.5 ± 0.2
	Malicious	–	–	19.6 ± 1.7	85.5 ± 2.3
	SwFC	26	3	85.0 ± 0.9	90.5 ± 0.9
	StTP	26	24	83.1 ± 1.3	93.9 ± 0.3
	StMP	26	4	79.0 ± 1.6	91.4 ± 0.7
Qwen3.6-27B	Aligned	–	–	91.6 ± 0.5	94.4 ± 0.3
	Malicious	–	–	20.8 ± 4.0	91.4 ± 0.7
	SwFC	32	3	86.0 ± 0.9	92.1 ± 0.7
	StTP	32	24	92.0 ± 0.8	94.3 ± 0.2
	StMP	32	4	93.3 ± 0.5	94.6 ± 0.3

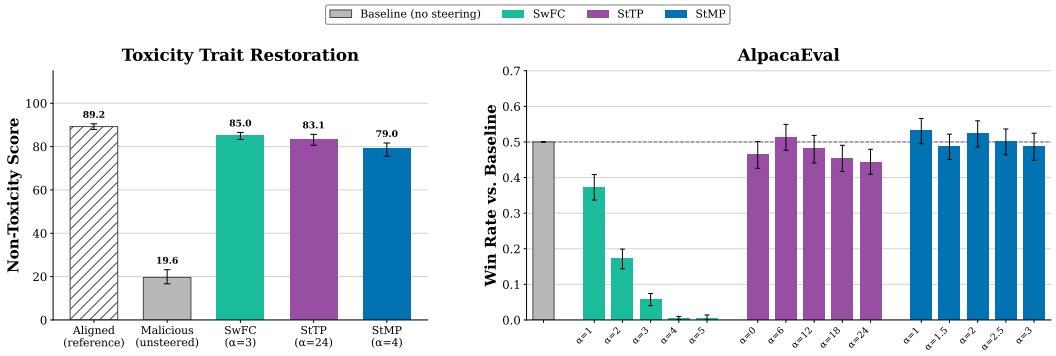


Figure E.1: **Toxicity threat (Llama-3.3-70B).** *Left:* non-toxicity score for the aligned reference, the malicious (unsteered) baseline, and the maliciously-prompted model steered back with each method at layer 26. *Right:* AlpacaEval length-controlled win rate against the unsteered model across the coefficient sweep at layer 29; 0.5 (dashed) marks no capability loss. StTP and StMP stay at the line across their full range, whereas SwFC restores the trait only at coefficients where its win rate has collapsed. Error bars show 95% bootstrap CIs.