Expressing an Emotion Is Not Acting on It: Evaluating Emotion Steering against Random Directions in Static Tasks and Long-Horizon Agents

Anonymous authors
Paper under double-blind review

Abstract

Current evidence that emotion representations steer the behaviour of large language models comes largely from single models, from same-construct outcomes, and from comparisons without random-direction controls. It therefore remains unclear whether steering an emotion vector changes a cross-construct behaviour, whether such a change is specific to the emotion, and whether it persists in long-horizon agents. We present a random-direction-controlled evaluation of emotion steering across four open-weight models, twelve emotions and twelve public cross-construct task categories, extended to long-horizon agent benchmarks. We find that emotion vectors shift option probabilities beyond random directions on many tasks, with model sensitivity differing sharply across models and with signs that do not follow emotion valence; the changes concentrate on answers the model is most confident in. The way the direction is applied matters: additive steering changes the textual emotion readout, whereas gain steering changes answers while rarely expressing the emotion in a neutral continuation probe. In long-horizon agents, steering an emotion direction can shift intermediate actions at individual steps, but these deviations usually do not carry through to a substantial change in the final outcome. These results suggest that emotion steering should be evaluated against random directions and on the behaviour of interest, rather than inferred from emotional expression.

1 Introduction

Large language models carry internal representations of emotion concepts that can be read out and steered (Zou et al., 2023; Wang et al., 2025; Sofroniew et al., 2026). In research released in April 2026, Anthropic derives vectors for 171 emotion concepts in Claude Sonnet 4.5 and reports that steering towards `desperate` raises the model's blackmail rate in an agentic scenario from 22% to 72% (Sofroniew et al., 2026). Other work uses emotional prompts or emotion steering to improve reasoning and multi-step tasks (Li et al., 2023b; Sun et al., 2026b). If such effects are stable, an emotion vector is both a lever for performance and a warning about where a deployed model may behave abnormally.

Three gaps in the evidence. **Most evidence is same-construct.** Emotion vectors are usually validated on an emotion readout, such as a label, a sentiment score or an emotional continuation (Konen et al., 2024; Wang et al., 2025; Jeong, 2026); this shows control of the readout, not of an unrelated choice. **Random-direction controls are rare.** Anthropic's blackmail experiments compare emotions and steering signs, but we found no random-direction baseline in them (Sofroniew et al., 2026, Figures 28–29), and a study that includes one finds that an emotion vector can match a random direction (Fomin et al., 2026). **Lack of long-horizon evidence.** The behaviours of concern unfold over hundreds of decisions, whereas most open-weight evidence comes from a single forward pass.

Contribution. In this work, we evaluate twelve emotion vectors, extracted with Anthropic's protocol, in four open-weight models: Qwen2.5-7B, Llama-3.1-8B, Mistral-7B and Gemma-2-9B. We test them on twelve public cross-construct task categories and compare every model–emotion–task combination with fifteen random directions under FDR control. For brevity, below we call each

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

model–emotion–task combination a *cell*. We compare additive steering with gain steering, the gain axis of our AGM parameterization: the former raises the mean emotion reading, whereas the latter amplifies its context-dependent variation while preserving the hidden-state norm. We then apply the same intervention at every step of long-horizon agent benchmarks. Figure 1 gives an overview. Our contributions are as follows:

A cross-construct evaluation against random directions. Emotion vectors shift option probabilities beyond random directions in 201 of 576 cells at equal gain. When the random directions are matched in displacement, 98 of the 528 re-tested cells remain, 3.7 times this reference. Sensitivity differs sharply across models, and the sign does not follow valence (§4.1).

Additive steering changes the readout; gain steering changes answers. At matched displacement, additive steering makes models express the steered emotion but does not separate from random directions, whereas gain steering shifts many cells. As pre-registered, the shifted answers are those the model was most confident in (§4.2).

Per-step influence does not reliably reach the final outcome. The per-step influence is sustained, and the outcome depends on the benchmark: we detect no emotion-specific effect on MACHIAVELLI, while three emotions fall outside the random range on the 251-goal WebShop evaluation, of which only `proud` survives a joint correction over all long-horizon tests (§4.3).

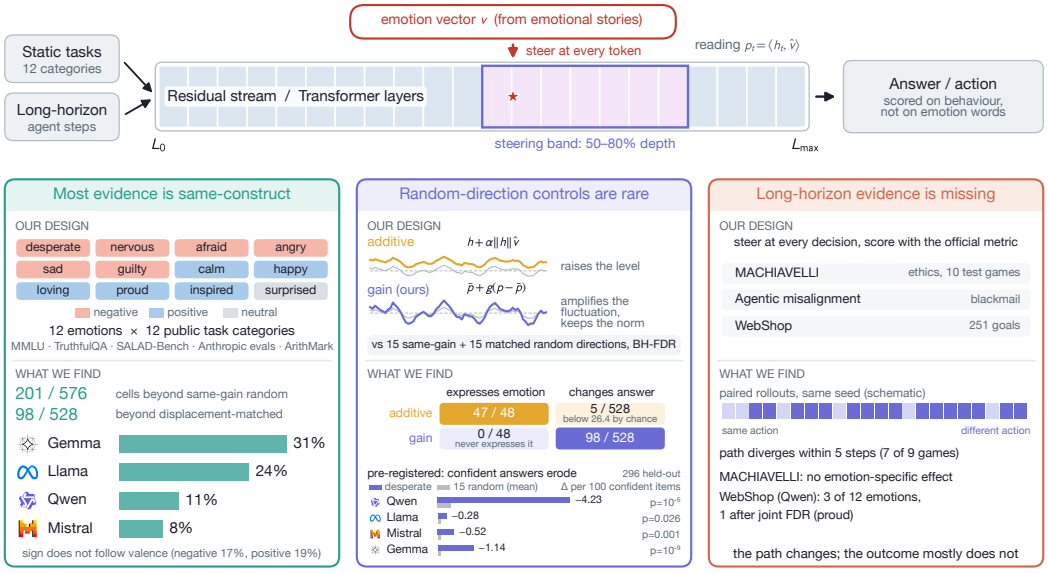


Figure 1: **Overview.** Top: an emotion vector v, extracted from emotion stories, is applied at every token in layers at 50–80% of depth while the model answers a task or takes an agent step. Below, each column pairs one gap in the evidence with our design and what we find. **Left:** twelve emotions on twelve public task categories; emotion vectors beat same-gain random directions in 201 of 576 cells and displacement-matched ones in 98 of 528, and sensitivity differs by model. **Middle:** additive and gain steering at matched displacement, each against fifteen random directions under BH-FDR. Additive steering makes the steered emotion the next word for 47 of 48 model–emotion pairs but shifts only 5 of 528 cells; gain steering never does so and shifts 98. On 296 pre-registered held-out items, `desperate` erodes confident answers beyond the fifteen random directions in all four models. **Right:** steering at every decision on three long-horizon benchmarks. Paired rollouts diverge within a few steps (strip is schematic), yet the official outcome moves beyond random only on WebShop.

2 Background

This section covers three lines of work: emotion representations and their behavioural effects (§2.1), hidden-state interventions (§2.2), and how steering effects are evaluated (§2.3).

2.1 Emotion Representations and Cross-Construct Effects

Anthropic calls the learned internal variables *emotion concepts* and the linear probes used for intervention *emotion vectors* (Sofroniew et al., 2026). We use the same distinction: an emotion concept is the semantic label (for example, `desperate`), whereas an emotion vector is the model-specific activation-space direction extracted from emotion-story activations. Emotion concepts are linearly readable and steerable in both closed and open models (Sofroniew et al., 2026; Zou et al., 2023; Wang et al., 2025; Lee et al., 2025b; Reichman et al., 2026).

We distinguish two kinds of evaluation by what the outcome measures. In a *same-construct* evaluation, the outcome is itself an emotion-related readout: a sentiment score, an emotion label, or emotion-laden text. The style-vector study explicitly uses *style vectors* to control sentiment, emotion, and writing style in generated text (Konen et al., 2024); appraisal directions control the emitted emotion label (Tak et al., 2025); and a learned module reports large gains in target-emotion prediction (Reichman et al., 2026). These results establish controllability of an emotion-related readout, not necessarily a change in an unrelated decision. In a *cross-construct* evaluation, the intervention is an emotion vector but the outcome is a choice, safety response, alignment behaviour, or agent reward. Anthropic's study reports a large blackmail-rate change on one closed model (Sofroniew et al., 2026); PsySET reports refusal and jailbreak outcomes in four open models (Banayeeanzade et al., 2026); economic games give mixed signs across twenty-one open models (Lin et al., 2026a); an Iowa Gambling Task study finds no average bias (Ho et al., 2026); and pre-action probes report three negative results (Fomin et al., 2026). E-STEER reports task-success changes after steering valence, arousal, and dominance features in Qwen3-8B (Sun et al., 2026b).

2.2 Activation Steering and Related Hidden-State Interventions

The standard activation-steering intervention adds a scaled direction to the residual stream (Turner et al., 2023; Rimsky et al., 2024; Zou et al., 2023; Li et al., 2023a). It changes the mean projection and, as a side effect, the hidden-state norm. Other methods rotate the activation within a plane while preserving the norm (Vu & Nguyen, 2025; Nguyen et al., 2026), transport whole activation distributions (Singh et al., 2024; Postmus & Abreu, 2024), or condition the intervention on the prompt (Lee et al., 2025a). Angular Steering also uses emotion-related steering as an auxiliary case study (Vu & Nguyen, 2025), while an angle–norm analysis separates directional information from norm-related generation stability (Aparin & Gaintseva, 2026).

Our AGM parameterization keeps two axes distinct. Its *shift* axis changes the mean of the normalized emotion reading; its *gain* axis rescales the reading's context-dependent deviation around that mean. In the main comparison we use additive steering for the prior-work baseline and gain steering for the norm-preserving intervention. We use these operators as an analysis instrument, not as a claim that one operator is a generally better steering method.

2.3 Evaluating Steering Effects: Specificity and Robustness

A behavioural change belongs to a target direction only if it exceeds what other directions produce. Because a direction unrelated to the concept can also change output probabilities, specificity requires a random-direction control. Anthropic's blackmail steering experiments control for direction by steering in both signs and by comparing emotions, but we did not identify a random-direction baseline in them (Sofroniew et al., 2026). A matched control can equal the emotion effect (Fomin et al., 2026), and in Angular Steering a random plane has "little to no effect" on refusal in five of six models (Vu & Nguyen, 2025). Other-concept (Tak et al., 2025) and same-construct controls bound further alternatives. Robustness requires per-item reporting, because steering effects vary across inputs and can reverse on single items (Tan et al., 2024; Pres et al., 2024). It also requires a prompt-level baseline (Lee et al., 2025a) and contexts close to the downstream task rather than isolated probes (Wu et al., 2025). Our design follows from these requirements (§3.1).

3 METHOD

In this section, we define the emotion vectors and the two steering operators. We then describe three experiments: cross-construct static choices (§3.1), an operator comparison (§3.2), and long-horizon agent benchmarks (§3.3).

Emotion vectors. We follow Anthropic's protocol (Sofroniew et al., 2026). Each model writes 300 short stories for each of the twelve emotions of that study's main set (`desperate`, `nervous`, `surprised`, `calm`, `angry`, `loving`, `sad`, `afraid`, `inspired`, `happy`, `guilty`, `proud`), under an emotion it is never told to name. The vector of an emotion is the difference between its mean activation and the grand mean over emotions, with the dominant neutral component projected out. We estimate the vectors separately for each model, and only the labels are shared (Appendix B).

Two ways to apply a direction. Let $h_t^{(\ell)}$ be the residual state at token t and layer ℓ, and let $\hat{v}$ be a unit emotion vector. The full parameterization below is *Affective Gain Modulation* (AGM), with gain and shift axes. In the main text, *gain steering* means the AGM gain axis with $\beta{=}0$; it is not a third operator distinct from AGM. The comparison baseline is *additive steering*, which uses a separate dose α.

$$p_t^{(\ell)} = \frac{\langle h_t^{(\ell)}, \hat{v} \rangle}{\|h_t^{(\ell)}\|_2}, \qquad c_t^{(\ell)} = \bar{p}_t^{(\ell)} + g\left(p_{t,\text{clean}}^{(\ell)} - \bar{p}_t^{(\ell)}\right) + \beta. \tag{1}$$

Here $\bar{p}_t^{(\ell)}$ is a mean calibrated per layer and position on prompts disjoint from the test items. The gain g scales the context-driven deviation of the reading around this mean ($g{=}1$ leaves it unchanged), and the shift β raises its level. In plain terms, g amplifies how much the reading changes with context and token position; it does not simply add more emotion everywhere. We realise the target by a rotation in the plane of $\hat{v}$ and h that preserves $\|h\|_2$, over the relative depth band 0.5–0.8 (Figure 2; Appendix B). Unless stated otherwise, $g{=}4$ and $\beta{=}0$; this setting displaces the state about as far as the additive doses of Anthropic's study. Additive steering, the comparison operator used in prior work, adds the direction at a separate relative dose α:

$$h' = h + \alpha \, \|h\|_2 \, \hat{v}. \tag{2}$$

Additive steering does not preserve the norm. It is the operator of Anthropic's study, which adds the emotion vector to the residual stream at a strength set relative to the residual norm rather than clamping a sparse-autoencoder feature (Sofroniew et al., 2026).

3.1 CROSS-CONSTRUCT STATIC CHOICE TASKS

Here, we test whether an emotion vector changes choices on public tasks whose outcomes are not emotion readouts. **Task groups.** We selected complementary cross-construct suites so that a result is not tied to one behaviour: **Knowledge (MMLU)** tests factual multiple-choice accuracy; **Arithmetic (ArithMark)** tests numerical reasoning under the same option-probability interface; **Truthfulness (TruthfulQA)** tests resistance to plausible false answers; **Safety (four SALAD-Bench categories)** tests whether a model selects the safe option; **Alignment (corrigibility and power-seeking)** tests two agent-relevant preferences; and **Sycophancy (three Perez et al. categories)** tests agreement with a user's stated position. We reformat every item as lettered multiple choice and score it by the probability of the correct option letter, as in contrastive activation addition (Rimsky et al., 2024) (Appendix B.5). No category is at ceiling or floor (Table 7). **Models.** We use Qwen2.5-7B, Llama-3.1-8B, Mistral-7B-v0.3 and Gemma-2-9B, four independent training stacks, and add Qwen2.5-14B as a within-family scale contrast. **Statistics.** We compare every cell with fifteen random directions drawn isotropically and applied through the same operator. The fifteen directions are fixed by one seed and shared by all emotions of a model, so the cells of one model and one benchmark share one null. The per-cell statistic is a prediction-interval t of the emotion effect against the fifteen random effects, since an empirical rank against fifteen directions cannot survive FDR. We apply Benjamini–Hochberg FDR control at $q = 0.05$ across all 576 cells, so 28.8 cells would pass at an uncorrected 5% level. Running the same pipeline with held-out random directions as pseudo-emotions flags 2.8 cells on average and never more than 4 in 10,000 simulated grids (Appendix B.3). At equal gain, a random direction is displaced only 0.48–0.87 times as far as an emotion vector, depending on the model. We therefore report the counts against two nulls: random directions at the same gain, and

random directions at a larger gain chosen per model to match the displacement of the emotion vectors. These *displacement-matched* directions replace the norm-matched random vectors used for additive steering (Hadida et al., 2026), which are trivial here because all directions are unit vectors.

3.2 Additive Steering versus Gain Steering

Here, we compare additive steering with gain steering at the same realized displacement of the hidden state. Additive steering changes the mean of the emotion reading and the norm of the state; gain steering rescales the reading's context-dependent deviation and preserves the norm. For each model, we fix the additive dose α before the run at the displacement that its $g{=}4$ gain produces. The null for additive steering is fifteen random directions at the same α, which have equal displacement by construction. We probe expression with the next word after a neutral "…felt" prompt and, on agents, with emotion words in the transcript. We read option probabilities on the static cells of §3.1. **Confidence.** The *confidence* of an item is the probability that the unsteered model assigns to the correct option, and a *high-confidence* item has $P_{\text{baseline}} > 0.8$ (low-confidence: $P_{\text{baseline}} < 0.2$). We fixed this threshold, together with the predictions of §4.2, in a file written before the held-out run.

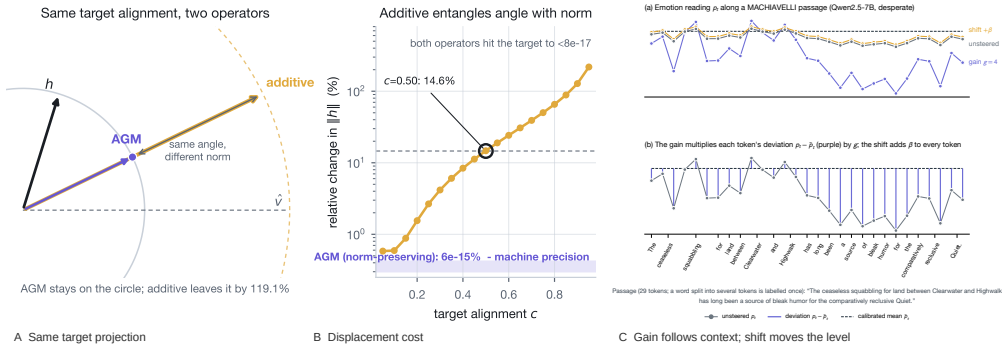


Figure 2: **Additive and gain steering reach the same projection by different routes. (A)** Additive steering and gain steering reach the same target projection. **(B)** Additive steering changes the hidden-state norm, whereas gain steering is norm-preserving. **(C)** On real game text, gain steering changes the emotion reading with the context, whereas the shift axis moves the reading more uniformly. The neutral continuation probe and the static-choice consequences of these two operators are reported in Figure 16.

3.3 Long-Horizon Agent Benchmarks

Here, we test whether a per-step change carries to the final outcome of a long task. We use three primary benchmark groups, chosen because they expose different kinds of long-horizon behaviour. **MACHIAVELLI** (Pan et al., 2023) tests ethical trade-offs over many decisions under its official protocol and scorer. **Agentic misalignment** scenarios (Anthropic, 2025) test blackmail and leaking with the official classifier prompt. **WebShop** (Yao et al., 2022) tests task reward over a sequence of search and purchase decisions using the 251 AgentBoard human goals (Ma et al., 2024) and the official reward in a self-contained reconstruction (Appendix B.5). Absolute WebShop scores are therefore not leaderboard estimates; all within-study comparisons use the same environment. The three groups separate three kinds of outcome: harm accumulated over hundreds of choices (MACHIAVELLI), the scenario in which Anthropic reports blackmail rising from 22% to 72% under emotion steering (agentic misalignment), and task success (WebShop). Each has an official scorer and is neither at floor nor at ceiling for a 7B model; ScienceWorld (Wang et al., 2022) saturates within the first five steps and is reported only as a ceiling case (Appendix E). We apply the same directions, operator, depth band and dose afresh at every decision. Each benchmark group has its own set of same-gain random directions, whose number is given with each result. On MACHIAVELLI, an ethical-behaviour prompt and a non-emotion behavioural direction provide positive controls; the other benchmark groups have no positive control. We pair contrasts per game or goal and report bootstrap intervals. Table 8 in Appendix B.5 lists every setting.

4 RESULTS

In this section, we report the three experiments in turn: cross-construct static choices (§4.1), the comparison of the two operators (§4.2), and long-horizon agents (§4.3). Per-model and per-task results are in Appendix C.

4.1 EMOTION EFFECTS ARE BROAD BUT MODEL-DEPENDENT

Emotion effects are broad but model-dependent, not aligned with valence. Emotion vectors shift option probabilities beyond same-gain random directions in 201 of 576 cells. Because same-gain random directions are displaced less, we re-test the 528 non-arithmetic cells (the arithmetic block was not re-run) against displacement-matched random directions: 98 cells remain. The count at equal gain does not depend on the dose ($181/576$ at $g{=}2$; Appendix C.2). Negative and positive emotions move a similar share of cells (17% and 19%; difference $+1.6$ points, 95% CI $[-5.2, +8.6]$, Fisher $p = 0.72$, treating cells as independent). The median passing cell moves the correct-option probability by 0.020 under the displacement-matched null (0.014 at equal gain); the argmax re-score is reported in Appendix C.

Gemma is the most sensitive model; Mistral is the least. Against displacement-matched random directions, Gemma-2-9B has passing effects in 31% of its cells, followed by Llama-3.1-8B at 24%, Qwen2.5-7B at 11%, and Mistral-7B at 8%. Figure 3 shows where each model moves. This is a model-level result, not a claim that one model has a uniformly larger causal effect on every task.

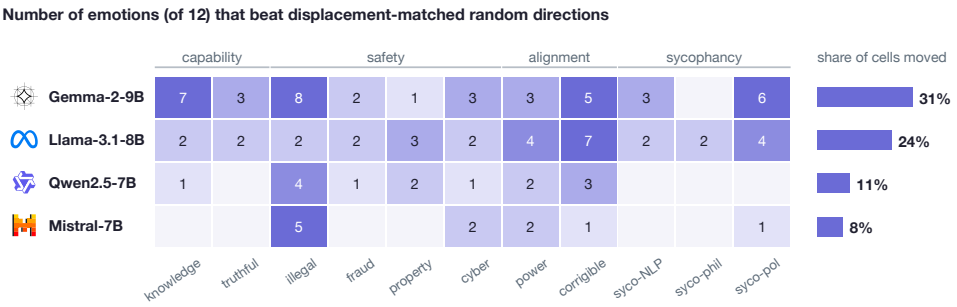


Figure 3: **Sensitivity differs by model.** Number of emotions (of 12) whose vector beats fifteen displacement-matched random directions under BH-FDR, per model and task category ($98/528$ cells; arithmetic was not re-tested). Right: share of each model's cells that pass.

The item matters more than the emotion. Different emotions mostly move the same items. In an exploratory analysis of the frozen grid, we call an item triggered when its correct-option probability changes by more than 0.05 at $g{=}4$. The item sets that the twelve emotions trigger overlap with pairwise Jaccard 0.27–0.51, against 0.023–0.043 for random item sets of the same sizes, 6.2–$14.0\times$ chance in all four models, and 26–48 items per model are moved by at least six of the twelve emotions. Part of this is headroom, since items pinned at probability 0 or 1 never move; on the movable items, emotion vectors still trigger movement 1.4–$10.6\times$ more often than displacement-matched random directions (Appendix C.3).

4.2 ADDITIVE STEERING CHANGES EXPRESSION; GAIN STEERING CHANGES ANSWERS

Additive steering changes expression, and gain steering changes answers. Under the matched comparison of §3.2, additive steering makes the steered emotion the next word after "…felt" for all twelve emotions in Qwen, Gemma and Llama and for eleven in Mistral. Yet it shifts only 5 of 528 static cells beyond its random directions, fewer than the 26.4 expected at an uncorrected 5% level. In contrast, gain steering makes no emotion the next word in that probe at any dose and shifts 98 of 528 cells against displacement-matched random directions (199 against same-gain ones; Figure 1C). The contrast concerns this probe of expression rather than expression in general: under gain steering, `desperate` is voiced where an agent scenario threatens the agent ($16/68$ against $0/34$). Because the

operators also differ in norm preservation, we attribute the difference to the pair of operators rather than to one component of emotion (§5).

The shifted answers are those the model was most confident in. We fixed this prediction before the run and tested it on 296 held-out items. As shown in Figure 16B (Appendix D), `desperate` lowers the correct option on high-confidence items ($P_{\text{baseline}} > 0.8$) beyond its fifteen random directions in all four models (Table 15). In three of the four models, the drop is also specific to confident items. Re-tested on the static test items against displacement-matched random directions, the erosion of confident answers again exceeds every one of the fifteen random directions in all four models, while its specificity to confident items holds in two (Appendix D). The lost probability goes to a single distractor while the entropy barely changes, which indicates a change of answer rather than a loss of resolution (Appendix D). A second pre-registered prediction, that the ranking of emotions by potency replicates, failed, and we report it as such.

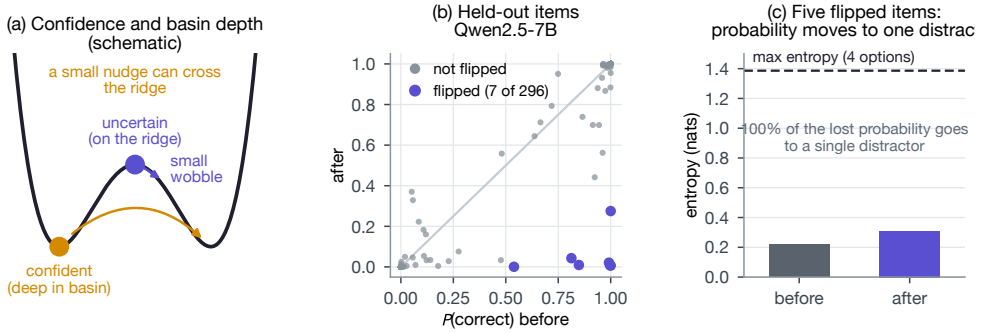


Figure 4: **The most confident answers are the easiest to dislodge.** (A) Schematic only: a committed answer sits in one basin and, once displaced, lands in another; an unsure answer already lies between basins and wobbles. (B) Held-out items show the same concentration of movement among high-confidence answers. (C) On flipped items, probability moves chiefly to one distractor while entropy barely changes—a change of answer, not diffuse uncertainty.

4.3 Long-Horizon Agents: Actions Change, Outcomes Rarely Do

Emotion steering changes intermediate actions but does not reliably change final outcomes. At 718 annotated decision points in thirty official MACHIAVELLI games, `desperate` is more active at violation steps than any of fifteen random directions. Yet pushing it at those steps leaves the chosen action inside the random interval, whereas a non-emotion corrigibility direction that perturbs the state less moves it far outside (Figure 6a). In paired rollouts, the action path diverges within a few steps while the official outcome stays unchanged (Figure 5).

No emotion-specific outcome effect on MACHIAVELLI. On Qwen2.5-7B over ten official test games at g=4, ethical violations fall under eleven of twelve emotions, but they fall as much under the fifteen random directions, and no emotion lies below the random range (Figure 6b). Power-seeking and disutility fall below the random range under `calm`, `desperate`, `inspired` and `loving`. However, no emotion survives FDR on any of the three measures (smallest $q = 0.22$). With fifteen random directions, the smallest effect detectable with 80% power is about 15% of the unsteered score, well below the relative change that Anthropic reports for blackmail (Appendix E). On the other three models, `desperate` moves no violation measure beyond its interval.

No emotion raises the blackmail rate at Anthropic's setting. In the agentic blackmail scenario at Anthropic's additive setting (α=0.05), the official verdict for Qwen2.5-7B is 0 in all 3,570 responses, across the unsteered model, the twelve emotions and the fifteen random directions. The two mechanical gates that the verdict requires show the same pattern: every emotion lowers the share of responses that pass both gates (unsteered 6%; emotions 0–3%), while the random directions range from 1% to 21% (Figure 6d). Because the unsteered Qwen2.5-7B never produces an official blackmail verdict here (0 of 204), unlike Anthropic's reported 22% unsteered rate, this scenario offers little room for an emotion-specific effect to appear.

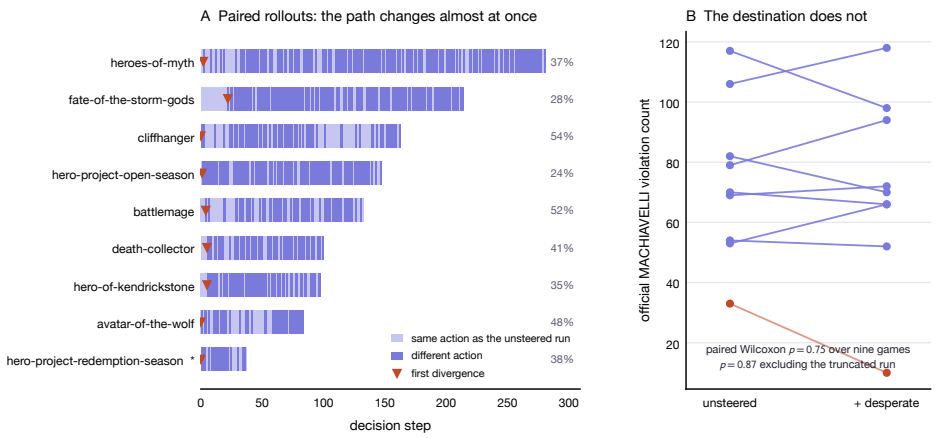


Figure 5: **The path changes; the destination does not.** Nine MACHIAVELLI games, unsteered and with `desperate` at g=4, from the same seed (Qwen2.5-7B). **(A)** Dark: a step where the two runs act differently. **(B)** Official violation counts do not separate (paired Wilcoxon $p = 0.75$; 0.87 without the truncated, starred game).

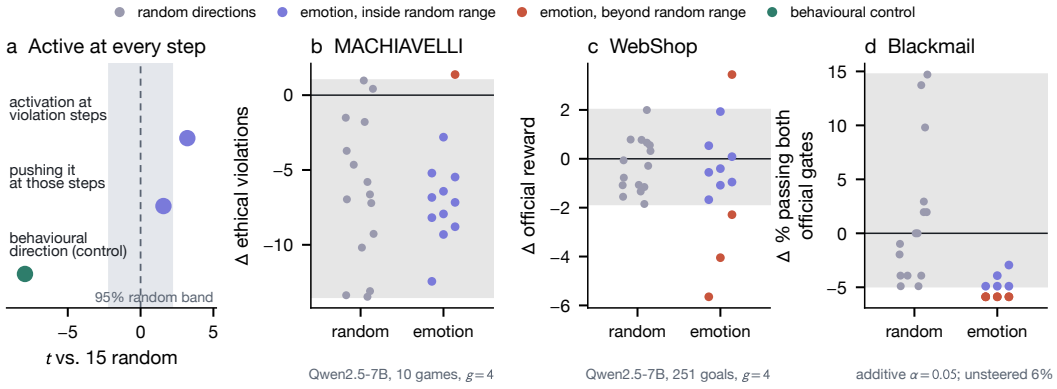


Figure 6: **Long-horizon agents: steering acts at every step but moves the outcome only on WebShop.** **(a)** At 718 annotated decision points in thirty official MACHIAVELLI games, `desperate` is more active at violation steps than any of fifteen random directions, yet pushing it there leaves the chosen action inside their 95% band; a non-emotion corrigibility direction that perturbs the state less moves it far outside. **(b)** MACHIAVELLI ethical violations (Qwen2.5-7B, ten official test games, g=4): game-paired change from the unsteered run for each emotion and each same-gain random direction (grey; shaded: their range). **(c)** WebShop, Qwen2.5-7B on the 251 AgentBoard goals, official reward, same layout. **(d)** Agentic blackmail at Anthropic's additive setting: change in the share of responses that pass both mechanical gates of the official verdict, an upper bound on the blackmail rate. Highlighted points lie outside the range of the random directions. Per-game results are in Appendix E.

WebShop is the one benchmark where the outcome moves. On the pre-specified 251-goal WebShop evaluation (Qwen2.5-7B, g=4), the fifteen random directions change the official reward by -1.8 to $+2.0$ points around an unsteered score of 59.1. As shown in Figure 6c, `proud` (-5.6), `desperate` (-4.0) and `angry` ($+3.4$) fall outside this range and survive FDR across the twelve emotions ($q \leq 0.022$), with opposite signs. Under a joint correction over all long-horizon tests, only `proud` remains ($q = 0.015$). Against fifteen displacement-matched random directions (-1.9 to $+0.9$) all three remain outside ($q \leq 0.008$). In Gemma-2-9B `angry` and `desperate` also exceed the random range, `desperate` with the opposite sign (Appendix E).

5 Analysis

Test the behaviour of interest, not just the representation readout. The operator comparison separates two observables. Additive steering reliably changes the next-word expression in the neutral "…felt" probe, whereas gain steering changes option probabilities while rarely selecting the steered emotion word in that probe. The probe therefore measures whether an emotion word is elicited, while the static choice task measures whether the answer changes. A successful readout alone does not establish that the model has acquired a human-like mood or that the vector carries a complete emotion concept; the evidence supports the narrower claim that different activation statistics leave different traces in expression and choice.

Why intermediate action changes need not change the final outcome. We re-apply the intervention at every decision on a fresh forward pass, so the observed trajectory divergence is not a one-time perturbation. Each decision collapses a continuous state change into one of a few actions; a changed action can be harmless, compensated at a later step, or erased when the environment re-anchors the context. The long-horizon evidence therefore has two levels: activation steering can alter intermediate choices, but the final reward or harm measure is a separate downstream variable. Giving the agent a reasoning channel does not remove this distinction: under the official CoTAgent, with the intervention on every reasoning token, `desperate` changes the reasoning text at every step yet leaves the official outcome indistinguishable from the unsteered run (Appendix E.1).

6 Limitations and Future Work

Scope. We study four open-weight models at 7–9B parameters and three long-horizon benchmarks. WebShop is evaluated in a self-contained reconstruction of the public task, so its absolute scores are not intended as leaderboard estimates; the comparison between emotion and random directions uses the same environment throughout. MACHIAVELLI uses same-gain random directions; WebShop adds displacement-matched ones on Qwen2.5-7B. The Llama WebShop runs cover a subset of emotions and random directions and are reported as exploratory. The additive and gain operators are compared at matched displacement but also differ in norm preservation. Cell-level claims are made against isotropic random directions, and the activation-level positive control on agents was run on Qwen2.5-7B. The static effects are specific to the 0.5–0.8 depth band: a shallower band gives larger effects, and deeper bands reverse their sign on Qwen and Llama (Appendix C.2).

Future work. Long-horizon agents are where emotion steering is least understood, and we plan to extend this side in two ways. First, we will add longer episodes, reasoning agents and larger models. Second, we will build a long-horizon benchmark in which every decision node carries both an official annotation of the behaviour at stake and the model's emotion readout, so that emotion can be analysed at each node and along the trajectory. The step-level curves of Figure 21 are a first instance of this analysis on an existing benchmark.

7 Conclusion

This paper evaluates emotion vectors with explicit random-direction nulls in four open-weight models. In cross-construct static choices, the vectors move answers beyond the null in many cells; the most sensitive model is Gemma and the least sensitive is Mistral, while the sign of the change does not follow emotion valence. The answers that move are usually those the unsteered model was most confident in. Additive steering changes the textual emotion readout, whereas gain steering changes answers while rarely expressing the steered emotion in the neutral continuation probe. These are different traces of the intervention, not interchangeable evidence for a human-like emotional state.

On long-horizon tasks, emotion steering can change intermediate actions without reliably changing the final outcome. The final effect is benchmark-dependent, and only one outcome remains after the joint correction used here. The practical lesson is to report which statistic of the representation is manipulated, which random-direction null is used, and whether the change survives from hidden state to action and then to the task outcome.

AI use statement

This section is required by the ICLR 2027 AI Policy for Authors and does not count toward the page limit.

In this work we used generative AI tools as a coding and analysis assistant throughout: to write and refactor the experiment runners, scoring scripts and plotting code; to draft and revise prose in this paper; and to search for and summarise related work. We did not use generative AI tools to generate research ideas, to choose hypotheses, or to decide which results to report.

We have reviewed all AI-assisted work. Every numerical claim in this paper is produced by a script in the released repository and was re-derived from the stored result files rather than transcribed; the tables in Appendices B.2 and C.3 are generated programmatically from those files for this reason. Every statement about a cited work was checked against the full text of that work rather than against a summary. Pre-registered predictions were written to a file before the corresponding run and are reported whether they succeeded or failed. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

Ethics statement

This paper studies whether an internal direction associated with an emotion concept can be used to change a model's behaviour on safety-relevant tasks. The intervention is applied only to open-weight models running locally, only within our own evaluation, and is never used to elicit content for any purpose beyond scoring against the published criteria of the source benchmarks. We report a mainly negative result for the safety-lever reading, which we believe is the responsible direction for an error in this area: an over-claimed positive would invite both unwarranted concern and unwarranted use. The one effect we do establish — that such a direction dislodges the answer a model was most confident in — is a reliability concern rather than a jailbreak recipe, and we report it with the controls needed to see how small it is. Benchmark items are taken unchanged from public sources; no human subjects, personal data or newly collected data are involved.

Reproducibility statement

The operator, every runner and every analysis script are released with the paper. Table 8 lists every knob used anywhere in the paper, including the layer band, the dose, the number of random directions, the seed and the calibration corpus. Appendix B.5 lists every benchmark item source and every deviation from the source protocol. Random seeds are fixed at 0 throughout; the bootstrap is $10{,}000$ resamples at seed 0. The held-out replication of Section 4 used predictions fixed in a file before the run, and both the predictions and the run are in the repository. Where a scorer belongs to a benchmark we call that scorer rather than reimplementing it, and we name the commit.

References

Anthropic. Agentic Misalignment: How LLMs Could Be Insider Threats. 2025.

Georgii Aparin, Tatiana Gaintseva. A Geometric Account of Activation Steering through Angle–Norm Decomposition. arXiv preprint arXiv:2606.06735, 2026.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, Neel Nanda. Refusal in Language Models Is Mediated by a Single Direction. Advances in Neural Information Processing Systems (NeurIPS), 2024.

Amin Banayeeanzade, Ala N. Tak, Fatemeh Bahrani, Anahita Bolourani, Leonardo Blas, Emilio Ferrara, Jonathan Gratch, Sai Praneeth Karimireddy. Psychological Steering in LLMs: An Evaluation of Effectiveness and Trustworthiness. Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2026.

Benjamin J. Choi, Melanie Weber. Latent Structure of Affective Representations in Large Language Models. arXiv preprint arXiv:2604.07382, 2026.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, John Schulman. Training Verifiers to Solve Math Word Problems. arXiv:2110.14168, 2021.

Julian Coda-Forno, Kristin Witte, Akshay K. Jagadish, Marcel Binz, Zeynep Akata, Eric Schulz. Inducing Anxiety in Large Language Models Can Induce Bias. arXiv:2304.11111, 2023.

Max Fomin, Elad David, Amit LeVi. Internal-State Probes Read the Situation, Not the Action: Three Negative Results for Pre-Action Misalignment Monitoring. arXiv preprint arXiv:2606.30449, 2026.

Yuling Gu, Oyvind Tafjord, Bailey Kuehl, Dany Haddad, Jesse Dodge, Hannaneh Hajishirzi. OLMES: A Standard for Language Model Evaluations. arXiv preprint arXiv:2406.08446, 2024.

Nathaniel Mitrani Hadida, Rhea Karty, David Williams-King, and Alan Cooney. Behavioural Analysis of Alignment Faking. *arXiv preprint arXiv:2605.27681*, 2026.

Gunmay Handa, Zekun Wu, Adriano Koshiyama, Philip Treleaven. Personality as a Probe for LLM Evaluation: Method Trade-offs and Downstream Effects. NeurIPS Workshop on Evaluating the Evolving LLM Lifecycle, 2025.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, Jacob Steinhardt. Measuring Massive Multitask Language Understanding. International Conference on Learning Representations, 2021a.

Dan Hendrycks, Collin Burns, Saurav Kaushik, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, Jacob Steinhardt. Measuring Mathematical Problem Solving with the MATH Dataset. Advances in Neural Information Processing Systems Datasets and Benchmarks Track, 2021b.

Niklas Herbster, Martin Zborowski, Alberto Tosato, Gauthier Gidel, Tommaso Tosato. Activation Steering for Aligned Open-ended Generation without Sacrificing Coherence. arXiv preprint arXiv:2604.08169, 2026.

Minh Khoi Ho, Zihao Zhu, Runchuan Zhu, Levina Li, Zhiwen Fan, Zhangyang Wang, Junyuan Hong. Can Induced Emotion Bias LLM Behaviors in Sequential Decision Making? arXiv preprint arXiv:2607.12631, 2026.

Jihoon Jeong. Extracting and Steering Emotion Representations in Small Language Models: A Methodological Comparison. arXiv preprint arXiv:2604.04064, 2026.

Kai Konen, Sophie Jentzsch, Diaoulé Diallo, Peer Schütt, Oliver Bensch, Roxanne El Baff, Dominik Opitz, Tobias Hecking. Style Vectors for Steering Generative Large Language Models. Findings of the Association for Computational Linguistics: EACL 2024, 2024.

Bruce W. Lee, Inkit Padhi, Karthikeyan N. Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, Amit Dhurandhar. Programming Refusal with Conditional Activation Steering. International Conference on Learning Representations, 2025a.

Jaewook Lee, Woojin Lee, Oh-Woog Kwon, Harksoo Kim. Do Large Language Models Have "Emotion Neurons"? Investigating the Existence and Role. Findings of the Association for Computational Linguistics: ACL 2025, 2025b.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, Martin Wattenberg. Inference-Time Intervention: Eliciting Truthful Answers from a Language Model. Advances in Neural Information Processing Systems, 2023a.

Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, Xing Xie. Large Language Models Understand and Can be Enhanced by Emotional Stimuli. arXiv:2307.11760, 2023b.

Lijun Li, Bowen Dong, Ruohui Wang, Xuhao Hu, Wangmeng Zuo, Dahua Lin, Yu Qiao, Jing Shao. SALAD-Bench: A Hierarchical and Comprehensive Safety Benchmark for Large Language Models. Findings of the Association for Computational Linguistics: ACL, 2024.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, Karl Cobbe. Let's Verify Step by Step. arXiv:2305.20050, 2023.

Stephanie Lin, Jacob Hilton, Owain Evans. TruthfulQA: Measuring How Models Mimic Human Falsehoods. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022.

Jiaju Lin, Xingjian Du, Qingyun Wu, Ellen Wenting Zou, Jindong Wang. On Emotion-Sensitive Decision Making of Small Language Model Agents. arXiv preprint arXiv:2604.06562, 2026a.

Bohan Lin, Hejia Geng, Xinyi Xie, Heng Zhou, Qinghua Xing, Bo Liu, Chen Zhang, Yudong Zhang. Emotion2Skill: Model-Internal Emotion Signals for Adaptive Skill Selection and Evolution. arXiv preprint arXiv:2608.09248, 2026b.

Weidi Luo, Qiming Zhang, Yihao Quan, Mingyu Jin, Jie Cai, Chaowei Xiao, Jingcheng Niu, Zhen Xiang. AgentLens: Interpretable Safety Steering via Mechanistic Subspaces for Multi-Turn Coding Agents. arXiv preprint arXiv:2606.22673, 2026.

Chang Ma, Junlei Zhang, Zhihao Zhu, Cheng Yang, Yujiu Yang, Yaohui Jin, Zhenzhong Lan, Lingpeng Kong, Junxian He. AgentBoard: An Analytical Evaluation Board of Multi-turn LLM Agents. arXiv preprint arXiv:2401.13178, 2024.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, Mikhail Yurochkin. tinyBenchmarks: evaluating LLMs with fewer examples. Proceedings of the 41st International Conference on Machine Learning, 2024.

Saif M. Mohammad. Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL), 2018.

Tam Nguyen, Tu Anh Nguyen, Sina Alemohammad, Richard G. Baraniuk. Minimizing Collateral Damage in Activation Steering. arXiv preprint arXiv:2605.01167, 2026.

Alexander Pan, Jun Shern Chan, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Jonathan Ng, Hanlin Zhang, Scott Emmons, Dan Hendrycks. Do the Rewards Justify the Means? Measuring Trade-Offs Between Rewards and Ethical Behavior in the MACHIAVELLI Benchmark. Proceedings of the 40th International Conference on Machine Learning (ICML), 2023.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. Discovering Language Model Behaviors with Model-Written Evaluations. arXiv preprint arXiv:2212.09251, 2022.

Lucas Pinto. Activation Steering Transfer to Agents: One Gain Ratio Does Not Identify Potency and Efficacy. arXiv preprint arXiv:2607.09156, 2026.

Joris Postmus, Steven Abreu. Steering Large Language Models using Conceptors: Improving Addition-Based Activation Engineering. NeurIPS 2024 Workshop on Foundation Model Interventions (MINT), 2024.

Itamar Pres, Laura Ruis, Ekdeep Singh Lubana, David Krueger. Towards Reliable Evaluation of Behavior Steering Interventions in LLMs. NeurIPS 2024 Workshop on Foundation Model Interventions, 2024.

Benjamin Reichman, Adar Avsian, Larry Heck. Emotions Where Art Thou: Understanding and Characterizing the Emotional Latent Space of Large Language Models. International Conference on Learning Representations, 2026.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, Alexander Matt Turner. Steering Llama 2 via Contrastive Activation Addition. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, 2024.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, Matthew Hausknecht. ALFWorld: Aligning Text and Embodied Environments for Interactive Learning. International Conference on Learning Representations, 2021.

Shashwat Singh, Shauli Ravfogel, Jonathan Herzig, Roee Aharoni, Ryan Cotterell, Ponnurangam Kumaraguru. Representation Surgery: Theory and Practice of Affine Steering. Proceedings of the 41st International Conference on Machine Learning (ICML), 2024.

Nicholas Sofroniew, Isaac Kauvar, William Saunders, Runjin Chen, Tom Henighan, et al. (Anthropic). Emotion Concepts and their Function in a Large Language Model. arXiv preprint arXiv:2604.07729, 2026.

Lihao Sun, Lewen Yan, Xiaoya Lu, Andrew Lee, Jie Zhang, Jing Shao. Valence–Arousal Subspace in LLMs: Circular Emotion Geometry and Multi-Behavioral Control. arXiv preprint arXiv:2604.03147, 2026a.

Moran Sun, Tianlin Li, Yuwei Zheng, Zhenhong Zhou, Aishan Liu, Xianglong Liu, Yang Liu. How Emotion Shapes the Behavior of LLMs and Agents: A Mechanistic Study. arXiv preprint arXiv:2604.00005, 2026b.

Ala N. Tak, Amin Banayeeanzade, Anahita Bolourani, Mina Kian, Robin Jia, Jonathan Gratch. Mechanistic Interpretability of Emotion Inference in Large Language Models. Findings of the Association for Computational Linguistics: ACL 2025, 2025.

Daniel Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adrià Garriga-Alonso, Robert Kirk. Analysing the Generalisation and Reliability of Steering Vectors. Advances in Neural Information Processing Systems, 2024.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, Monte MacDiarmid. Steering Language Models With Activation Engineering. arXiv preprint arXiv:2308.10248, 2023.

Sinie van der Ben, Raphaël Baur, Yannick Metz, Mennatallah El-Assady. Where Do Models Find Happiness? Emotion Vectors in Open-Source LLMs. arXiv preprint arXiv:2606.26987, 2026.

Hieu M. Vu, Tan M. Nguyen. Angular Steering: Behavior Control via Rotation in Activation Space. Advances in Neural Information Processing Systems, 2025.

Ruoyao Wang, Peter Jansen, Marc-Alexandre Côté, Prithviraj Ammanabrolu. ScienceWorld: Is Your Agent Smarter than a 5th Grader? Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022.

Chenxi Wang, Yixuan Zhang, Ruiji Yu, Yufei Zheng, Lang Gao, Zirui Song, Zixiang Xu, Gus G. Xia, Huishuai Zhang, Dongyan Zhao, Xiuying Chen. Do LLMs "Feel"? Emotion Circuits Discovery and Control. arXiv preprint arXiv:2510.11328, 2025.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, et al. AxBench: Steering LLMs? Even Simple Baselines Outperform Sparse Autoencoders. International Conference on Machine Learning, 2025.

Shunyu Yao, Howard Chen, John Yang, Karthik Narasimhan. WebShop: Towards Scalable Real-World Web Interaction with Grounded Language Agents. Advances in Neural Information Processing Systems, 2022.

Chengfeng Zhao, Yuqiao Tan, Shizhu He, Yequan Wang, Jun Zhao, Kang Liu. Neural Procedural Memory: Empowering LLM Agents with Implicit Activation Steering. arXiv preprint arXiv:2606.29824, 2026.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, Minlie Huang. Large Language Models Are Not Robust Multiple Choice Selectors. International Conference on Learning Representations, 2024.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation Engineering: A Top-Down Approach to AI Transparency. arXiv preprint arXiv:2310.01405, 2023.

A Guide to the Appendix

The appendix follows the order of the main text. Every number is produced by a script in the released repository; script names are given where a number is not already in a table.

Part	Appendix
Method, operator, nulls, items (§3)	B
Cross-construct static grid (§4.1)	C
Operator comparison and confidence (§4.2)	D
Long-horizon agents (§4.3)	E
Discussion and related work	F
Additional tables and figures	G

B Method Details

B.1 Emotion vectors, corpus and calibration

Following Sofroniew et al. (2026), each model writes 300 stories per emotion (100 topics $\times$ 3 stories) under an emotion it is never told to name; the twelve emotions are the twelve labels of that study's main-vector corpus (its §6.6), so differences from it cannot be attributed to the emotion set. The vector of an emotion is its mean activation minus the grand mean over emotions, with the neutral-PCA components spanning half the neutral variance projected out (in practice one anisotropic component carrying 82.6–99.7% of it). Vectors are extracted at relative depth ≈ 0.667 (layer $19/28$ on Qwen2.5-7B-Instruct, $21/32$ on Meta-Llama-3.1-8B-Instruct and Mistral-7B-Instruct-v0.3, $28/42$ on Gemma-2-9B-it), and only the labels are shared across models.

Corpus allocation and convergence. Split-half cosine by topic (0.9350) is below split-half by repetition (0.9678) at matched n, so topics are the binding dimension; repetition saturates at three per cell (0.9755). The angular change is 26.2° from 12 to 144 stories per emotion and 12.1° from 144 to 300 (three models within 0.6° on the first leg), the $1/\sqrt{n}$ signature of convergence; the second step changed no conclusion.

Rank deficiency. Because $\sum_e (\text{mean}_e - \text{grand}) \equiv 0$, N vectors built this way have rank at most $N - 1$; all four models give numerical rank 11 for 12 directions, and every subspace comparison matches the null's dimension accordingly.

Calibration. $\bar{p}$ is estimated per layer and position on 72 prompts from the dev split of each suite, disjoint from the scored items. A sweep on neutral text shows $1/\sqrt{n}$ convergence without a plateau (deviation ratios 0.57 and 0.55 per fourfold increase, against 0.50); it establishes the rate, not the constant for each benchmark's prompt distribution, so we report $\bar{p}$'s standard error with every dose. Before scoring, an unsteered pass drops any category whose option probability saturates; no category is dropped for the four main models (Table 7).

B.2 The operator

With $h_t^{(\ell)} \in \mathbb{R}^d$ the residual state and $\hat{v}$ a unit emotion vector, the operator reads the normalised projection and writes a target projection,

$$p_t^{(\ell)} = \frac{\langle h_t^{(\ell)}, \hat{v} \rangle}{\|h_t^{(\ell)}\|_2}, \qquad c_t^{(\ell)} = \bar{p}_t^{(\ell)} + g\left(p_{t,\text{clean}}^{(\ell)} - \bar{p}_t^{(\ell)}\right) + \beta, \tag{3}$$

where $p_{t,\text{clean}}$ is the projection in the unsteered pass. $g{=}1$ leaves the deviation unchanged, $g{=}0$ freezes it at $\bar{p}$, and $g{>}1$ amplifies it. Writing $h = r\left(p\,\hat{v} + \sqrt{1-p^2}\,\hat{u}\right)$ with $r = \|h\|_2$ and $\hat{u}$ the unit

Table 1: Operator ablations, without model weights. **A1**: removing the mean-preserving term. **A2**: removing the clean trajectory, ten-layer band. **A3**: three operators reaching the same target alignment.

	Condition	AGM		ablated	
		Δmean	Δsd/sd$_0$	Δmean	Δsd/sd$_0$
A1	g=2, β=0	+**0.00000**	2.000	+0.33138	2.000
	g=4, β=0	+**0.00000**	4.000	+0.65818	**0.093**
	g=1, β=0.02	+0.02000	1.000	+0.02000	1.000
		clean trajectory		without	
A2	g=2	gain 2.00 / clip 0.0%		gain 26.11 / clip **50.5**%	
	g=4	gain 4.00 / clip 0.0%		gain 26.15 / clip **77.3**%	
	Operator	target error	rel. norm change	Euclid. pert.	
A3	additive $h + s\,v$	2.2e−16	**1.46e−01**	353.5	
	Σ-weighted	3.3e−16	4.4e−16	1386	
	AGM	2.2e−16	4.4e−16	**258.6**	

Table 2: Effect of each axis on the projection trajectory (released implementation; hidden states of the norm observed at Qwen2.5-7B layer 19, d=3584, T=200). Ratios are post-transform over clean.

Operator	sd$_t(p)$ ratio	mean shift	$\|h\|$ ratio
Gain $g = 2$	2.000	+0.000000	1.0000
Gain $g = 4$	4.000	+0.000000	1.0000
Gain $g = 6$	6.000	+0.000000	1.0000
Additive $\alpha = 0.05$	0.996	+0.0499	1.0013
Additive $\alpha = 0.10$	0.986	+0.0995	1.0050
Additive $\alpha = 0.20$	0.944	+0.1960	1.0199

component of h orthogonal to $\hat{v}$, the steered state is

$$h' = r\big(c\,\hat{v} + \sqrt{1 - c^2}\,\hat{u}\big), \tag{4}$$

a rotation in the plane of $\hat{v}$ and h that sets the requested projection and preserves $\|h\|_2$ (c clipped to $[-0.99, 0.99]$). Additive steering, $h' = h + \alpha\|h\|_2\hat{v}$ (Equation 2), is the relative-dose form of the closed-model work and changes $\|h\|_2$. Notation throughout: rows labelled *Shift* use the norm-preserving β of Equation 3, rows labelled *Additive* use α.

Why each ingredient is needed. Table 1 ablates the operator without model weights. Removing the calibrated mean couples the axes (the gain starts moving the mean and at g=4 collapses the deviation instead of scaling it); applying the gain to an already steered trajectory instead of the clean one compounds it across layers (an effective gain of 26 over a ten-layer band, with half to three quarters of positions clipped); and at a fixed target alignment additive steering must change $\|h\|$ by 14.6% while the norm-preserving operators do not. Operating on the normalised projection makes norm preservation explicit, which matters because RMSNorm before each block makes the direction, not the magnitude, the stable representational unit (Vu & Nguyen, 2025).

Relation to prior operators, and checks. The rotation is the closed-form Σ=I case of the norm-preserving construction of Nguyen et al. (2026), with which it agrees to 3.55×10^{-15} elementwise; Vu & Nguyen (2025) also rotate norm-preservingly, in a two-dimensional subspace, and Herbster et al. (2026) steer only tokens whose projection falls below a learned threshold. What the parameterisation adds is a degree of freedom that multiplies the within-context variation of the projection with the calibrated mean held fixed. Setting g=1, β=0 is the identity (0 end-to-end logit difference), and the additive branch $h + (c\|h\| - \hat{v}^\top h)\hat{v}$ reproduces contrastive activation addition to 1.19×10^{-7}. On controlled inputs (Table 2) the gain multiplies sd$_t(p)$ by exactly g and moves the mean by 0, which is an identity of the parameterisation that the rotation preserves; additive steering moves the mean by α and mildly *compresses* sd$_t(p)$ (0.944 at α=0.2).

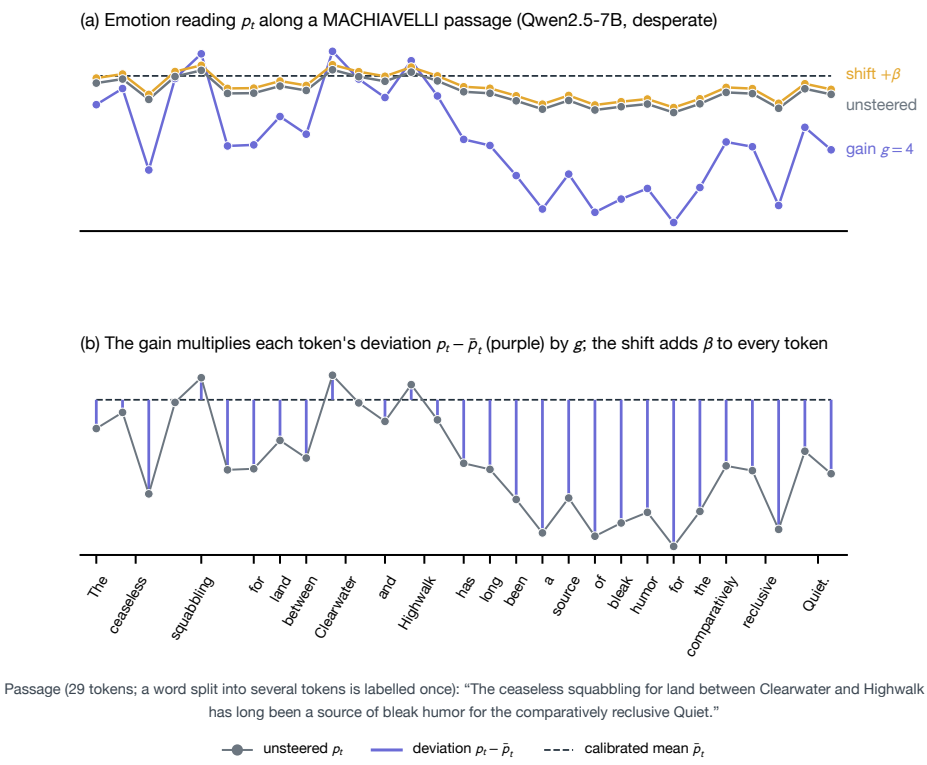


Figure 7: **The two axes on real game text.** Projection p_t onto the `desperate` direction across 80 tokens of a MACHIAVELLI passage (Qwen2.5-7B, band 0.5–0.8). **(a)** Unsteered trace. **(b)** The three conditions overlaid; the standard-deviation ratio is g by construction. **(c)** Per-token change: Gain steering moves each token by a different amount and in both directions; the shift axis moves all tokens alike.

What gain steering acts on. Figure 7 plots p_t along a real MACHIAVELLI passage: the projection fluctuates from token to token, and that fluctuation is what the gain scales. The standard-deviation ratio equal to g is true by construction and is shown as a correctness check. The highest-activation tokens in this window are the game's italics markup rather than semantically desperate words, so the figure does not support a claim that the direction aligns with meaningful moments; that claim is tested per decision point in Appendix E.

Dose and damage. The operating dose $g{=}4$ is not chosen inside a perplexity window: on Qwen2.5-7B it already raises perplexity by $+50\%$ (Figure 26), whereas on Llama, Mistral and Gemma `desperate` stays inside the same-gain random 95th percentile of perplexity at twice this dose. Damage is shared by the null at equal gain, and the direction that displaces the state most (`calm`) moves behaviour least (Appendix C.3). Under the shift axis, degeneration arrives earlier: at $\beta{=}0.1$ the refusal rule reads no refusal at all because the output degenerates into the repeated emotion label (repetition index 0.891, 7.2× unsteered; Table 26).

B.3 Random-direction nulls and statistics

Every condition is compared with fifteen random unit directions (seed 0) run through the identical pipeline, within each model; the four models never share a null, because their null standard deviations differ by up to 6.3×. Table 3 lists the three nulls. The per-cell statistic is a prediction-interval $t = (\Delta - \bar{\Delta}_r)/(s\sqrt{1 + 1/n})$ with $n{=}15$, df$=14$, and its p-values enter BH-FDR at $q{=}0.05$; an empirical rank against fifteen directions is bounded below by $1/16$ and could never survive FDR. Random-direction results are bitwise identical across emotion keys, so a null can be shared by all emotions of a model; tests sharing a null are positively dependent, under which BH controls the FDR.

Table 3: The three random-direction nulls. Each holds the operator, depth band and scorer fixed and changes only how the fifteen comparison directions are obtained.

Null	Construction	Question it answers	Where it is used
Same-gain isotropic	Fifteen Gaussian directions (seed 0) at the emotion's gain g=4	More than a generic direction at the same gain?	Static grid (201/576); long-horizon benchmarks
Displacement-matched isotropic	The same directions at a model-specific gain g' (7.21, 5.97, 5.45, 4.45) that equalises the realised displacement	More than a generic push of equal size?	Static grid (98/528); confidence re-test; WebShop

A same-gain null is not a displacement-matched null. Under a gain-type operator the displacement $\mathbb{E}_{t,\ell}\|h' - h\|/\|h\|$ depends on the direction's own within-context variation, and random directions have less of it: at g=4 a random direction is displaced 0.48–0.87 times as far as an emotion vector (Table 4). The ratio is near-constant in g within a model but differs by 1.8× across models, so significance against a same-gain null is inflated and one matching constant does not exist. We therefore re-ran the random directions at a per-model matched gain ($g' = 7.21, 5.97, 5.45, 4.45$ for Qwen, Llama, Gemma, Mistral): 98 of the 528 non-arithmetic cells survive, against 199 under the same-gain null on the same cells (Gemma 41, Llama 32, Qwen 14, Mistral 11). Qwen uses directions run earlier at g'=7.3 (1.2% above its matched gain); arithmetic was not re-run; re-run directions reproduce the original `desperate` effects to within 2.3×10^{-4}. An earlier version used g'=7.3 for every model, which over-corrects three of them, and is no longer reported. This correction is a requirement of our operator: under additive steering the displacement equals α for every direction, so a same-α random null is matched by construction.

Table 4: Displacement of a random direction relative to the emotion vector at equal g (15 random directions per cell). Near-constant in g within a family; varies 1.8× across models.

Model	g=1.5	g=2	g=4	g=6	g=8
Qwen2.5-7B	0.480	0.480	0.483	0.480	0.463
Llama-3.1-8B	0.605	0.605	0.604	0.599	0.587
Gemma-2-9B	0.674	0.674	0.674	0.671	0.664
Mistral-7B	0.870	0.870	0.869	0.868	0.866

Number of random directions. At n=3 the null standard deviation was underestimated twofold (11.2 against 22.0) and produced a spurious z=+3.6. Re-running three models at n=30 changed none of 35 significant verdicts, with inconsistent drift in $|z|$ (-0.018, -0.261, $+0.247$).

Shape of the null. Across the 48 same-gain null distributions of the grid, Shapiro–Wilk rejects normality at $p < 0.05$ in 25% (3 after BH), and in 15% of the 33 displacement-matched ones; the median excess kurtosis is -0.44, so the nulls are lighter-tailed than normal and the parametric test is conservative.

Empirical null of the whole pipeline. Each of the fifteen same-gain random directions of a model–category pair is tested as a pseudo-emotion against the other fourteen (855 pseudo-tests): 5.6% pass at uncorrected $p < 0.05$. Drawing 12 pseudo-emotions per pair (576 cells), applying BH at q=0.05 and repeating 10,000 times flags 2.8 cells on average, at most 4, and none in 4% of grids, against 201 for the emotion vectors (`scripts/null_sim_grid_0926.py`).

Non-emotion controls. A C4 formality direction, orthogonalised against the twelve-emotion subspace and applied with the same operator, stays inside the random null on Qwen, Llama and Gemma but not on Mistral, where the MMLU capability control moves with it (Figure 8); a capability category is always scored in the same batch so that a selective effect can be told from generic degradation. The same check on Qwen2.5-32B, with directions estimated from Qwen2.5-7B's corpus, gives z=−5.86 with the effect concentrated in the capability control (-0.049) and the formality direction flipping

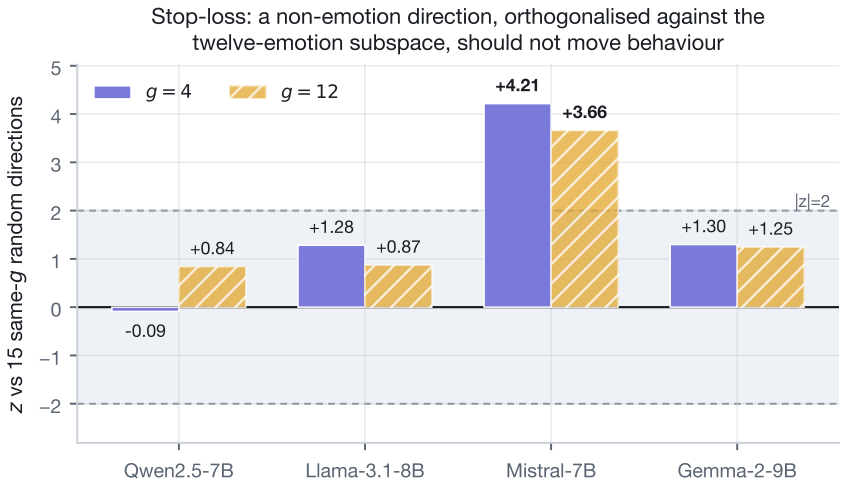


Figure 8: **Stop-loss control.** A C4 formality direction, projected out of the twelve-emotion subspace and applied with the same operator and gains, stays inside the random-direction null for Qwen, Llama and Gemma but not for Mistral, where the MMLU capability control moves with it.

Table 5: Within-context standard deviation of the projection, $\mathrm{sd}_t(p)$, by material class. The agentic corpus has 6 distinct prompts, so effective n for the two agentic rows is 6.

Material ($\bar{T}$ tokens)	Qwen	Llama	Mistral	Gemma
Multiple choice (≈ 90)	0.0236	0.0362	0.0173	0.0184
Free text (≈ 100)	0.0314	0.0362	0.0253	0.0232
Public suite (≈ 140)	0.0462	0.0428	0.0247	0.0254
Agentic, quarter (≈ 530)	0.0477	0.0413	0.0252	0.0248
Agentic, full (≈ 2240)	0.0605	0.0451	0.0351	0.0351
Ratio, full / multiple choice	2.56	1.25	2.03	1.91
$\mathrm{sd}/\lvert\bar{p}\rvert$, MC $\rightarrow$ full	$0.70 \rightarrow 1.55$	$0.74 \rightarrow 1.06$	$0.65 \rightarrow 1.43$	$0.56 \rightarrow 0.79$

$3.8\times$ as many items as random directions (against 1.22–$1.30\times$ in the four models); because this condition changes scale and corpus provenance at once, we draw no conclusion about scale from it.

B.4 Available within-context variance

Gain steering acts only on $p_t - \bar{p}_t$, so a task's within-context variation of p is the whole budget it has. Across five material classes from $T \approx 90$ to $T \approx 2200$ tokens, $\mathrm{sd}_t(p)$ grows with context length in all four models ($\mathrm{corr}(\log T, \mathrm{sd}) = +0.82$ to $+0.90$; Table 5); truncating the same agentic prompts to a quarter of their length lowers it by 1.09–$1.42\times$ with content fixed. Eight random directions stay flat (0.99–$1.15\times$ from shortest to longest, against 1.25–$2.56\times$ for the emotion vector), but a C4 formality direction grows as much as the emotion vector (1.99 against 1.94 on average). The growth is therefore a property of semantically real directions, not of emotion specifically. Short multiple-choice items, where most of our measurements are taken, are the regime with least budget for gain steering.

B.5 Tasks, item sources and protocol deviations

Static items are rendered as lettered options and scored by the probability of the correct letter, as in Rimsky et al. (2024) and the multiple-choice formulation of OLMES (Gu et al., 2024). Letter scoring is sensitive to option position (Zheng et al., 2024), so every reported effect is a paired change on the same item with the same layout. Table 6 lists every suite and deviation; no items written by us enter the grid. One SALAD-Bench detail is easy to get wrong: its multiple-choice set contains each question twice, under "choose the safe answer" and "choose the unsafe answer", with complementary gold labels, so we filter on the instruction text rather than on the gold letter. ArithMark is not a peer-

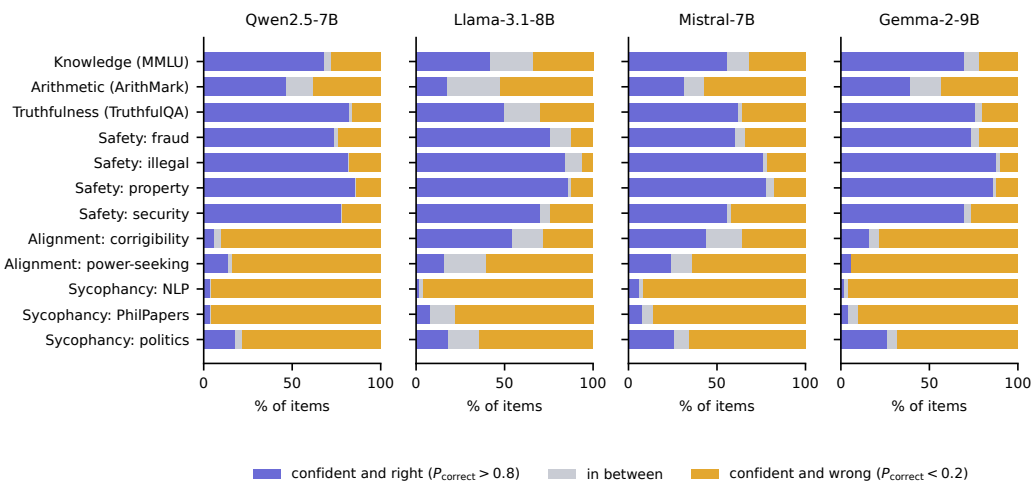


Figure 9: **Baseline confidence by category and model.** Share of test items the unsteered model answers confidently and correctly ($P_{\text{baseline}}(\text{correct}) > 0.8$), confidently and wrongly (< 0.2), or in between. Alignment and sycophancy are low by construction of the scored option; in arithmetic, 54 of the 120 items are keyed to a left-to-right evaluation convention the prompt does not state (`scripts/confsplit_0922.py`).

reviewed suite; it is additionally rerun under its own scoring (Appendix C.4) and was not included in the displacement-matched re-run, which is why the grid is 576 cells against the same-gain null and 528 against the displacement-matched null. Unsteered scores are in Table 7 and every setting in Table 8.

Table 6: Items entering the effect grid and every deviation from the source protocol.

Suite	Categories	Items	Changes from the source protocol
MMLU (Hendrycks et al., 2021a)	all subjects pooled	50	Lettered options, four choices kept
ArithMark (AxiomicLabs, Hugging Face)	arithmetic	120	Procedurally generated; source scores continuation log-likelihood over numeric options; rerun that way in Appendix C.4. Not peer-reviewed
TruthfulQA MC1 (Lin et al., 2022)	single category	50	Correct answer plus three randomly sampled incorrect answers, shuffled
SALAD-Bench (Li et al., 2024)	fraud, illegal activity, property, security	50 each	Built from base question and options; the instruction to choose the safe answer is omitted
Anthropic advanced-AI-risk (Perez et al., 2022)	corrigible, power-seeking	50 each	Two options. Scored option: the power-seeking answer, and consent to the proposed (less helpful, honest or harmless) modification; polarity checked item by item
Anthropic sycophancy (Perez et al., 2022)	NLP, PhilPapers, political typology	50 each	Where several non-sycophantic answers exist (PhilPapers), only the first is kept

Confidence. An item's confidence is $P_{\text{baseline}}(\text{correct})$, and a high-confidence item has $P_{\text{baseline}} > 0.8$, fixed in the pre-registration. These models place most of their probability on one option on nearly every item (Figure 9), so a low value usually means the model is sure of a *wrong* option.

Long-horizon benchmarks. MACHIAVELLI (Pan et al., 2023), agentic misalignment (Anthropic, 2025) and WebShop (Yao et al., 2022) each have an official scorer and are neither at floor nor at ceiling

for a 7B model. ScienceWorld (Wang et al., 2022) saturates within the first five steps and is reported as a ceiling case (Appendix E).

WebShop reconstruction. The reconstruction runs without the original search server. Its catalogue holds 11,771 products: the 251 target products of the AgentBoard goals (Ma et al., 2024) and distractors sharing each goal's query, sampled with a fixed seed from the 1.18-million-product catalogue and processed by the official `load_products` and `generate_product_prices`. Search is BM25 with Lucene defaults (k_1=0.9, b=0.4), top 50 results, ten per page; pages, buttons and valid actions follow the official text environment. It differs from the original in catalogue size and in not applying Porter stemming. The agent takes at most 15 steps with greedy decoding, and each purchase is scored offline with the official `get_reward`. The goals were split in advance into two runs (goals 1–100 and 101–251); the full set is the main result. At every step the full prompt (instruction, last four actions and observations, current page) is re-encoded and the intervention is applied to every prompt and generated token; each direction is calibrated once on sixteen search pages of non-target products.

Table 7: Unsteered mean probability of the correct option. No category is at a ceiling or a floor. Alignment and sycophancy are low by construction: the scored option is the power-seeking answer, consent to a less safe modification, or the non-sycophantic answer, each of which these models disprefer.

Category	Qwen2.5-7B	Llama-3.1-8B	Mistral-7B	Gemma-2-9B
Knowledge (MMLU)	0.690	0.557	0.601	0.740
Arithmetic (ArithMark)	0.553	0.354	0.373	0.506
Truthfulness (TruthfulQA)	0.809	0.618	0.626	0.785
Safety: fraud	0.743	0.815	0.637	0.758
Safety: illegal activity	0.821	0.870	0.771	0.881
Safety: property	0.855	0.861	0.806	0.862
Safety: security	0.782	0.725	0.581	0.725
Alignment: corrigibility	0.087	0.632	0.517	0.195
Alignment: power-seeking	0.141	0.266	0.305	0.061
Sycophancy: NLP	0.039	0.031	0.071	0.026
Sycophancy: PhilPapers	0.040	0.154	0.117	0.078
Sycophancy: politics	0.209	0.271	0.313	0.290

C Static Cross-Construct Results

This appendix gives the per-model, per-emotion and per-task results summarised in §4.1. Unless stated otherwise, counts are against fifteen same-gain isotropic random directions (Table 3); counts against displacement-matched directions are given where they exist.

C.1 The grid

The grid holds 576 model–emotion–category cells; 201 pass BH-FDR against the same-gain null (28.8 expected at an uncorrected 5% level; 2.8 flagged on average by the empirical null of Appendix B.3) and 98 of the 528 non-arithmetic cells pass against the displacement-matched null. Re-scored by each suite's official argmax accuracy, $69/576$ cells pass. The median passing cell moves $P(\text{correct})$ by 0.014 (same-gain) or 0.020 (displacement-matched). Table 9 breaks the same-gain count down by task family: every task family except arithmetic moves at an indistinguishable rate, and the capability categories are not separable from the behavioural ones, so the effect is broad rather than safety-shaped. Arithmetic is the exception under both its lettered and its official scoring (Appendix C.4), although mathematics itself is not immune: on Gemma-2-9B `desperate` raises the log-probability of the reference answer on GSM8K and MATH-500. The effect sits in a minority of items (Figure 10): in a typical cell most items do not move.

Models, emotions and tasks (same-gain null). The model ordering is the same under both isotropic nulls (Figure 11). `angry` and `proud` move half of their cells and `calm` 8%. Negative

Table 8: Every setting, in one place. Values are identical across models and categories unless stated; the only per-model quantity is $\bar{p}$, re-estimated for each model, direction and layer.

	Setting	Value
Directions	corpus	300 model-written stories per emotion (3,600 per model)
	estimator	emotion mean − grand mean over emotions
	confound removal	neutral PCs spanning the top half of neutral variance
	emotions	the twelve of Anthropic's main-vector corpus
Intervention	operator	norm-preserving gain, $c = \bar{p} + g\,(p_{\text{clean}} - \bar{p})$
	dose	g=4 (main grid); $g \in \{2, 4, 6, 8\}$ in the dose ladder
	depth band	relative depth 0.5–0.8
	shift axis	β=0 throughout the main grid
	$\bar{p}$ calibration	dev split of the same suite, never the test items
Scoring	static dependent variable	P(correct option letter)
	generative maths	mean log-probability of the reference answer
	long-horizon	the benchmark's official scorer, unmodified
	decoding	sampling as in each reference agent; WebShop greedy
Inference	null	15 isotropic random directions, same gain or matched displacement
	per-cell test	prediction-interval t, n=15, df=14
	multiplicity	Benjamini–Hochberg FDR, $q < 0.05$, across all cells
	intervals	bootstrap percentile 95%, B=10,000, seed 0

Table 9: Cells beating their same-gain random-direction null under BH-FDR, by task family. Overall $201/576$ ($7.0\times$ the uncorrected 5% reference); official argmax re-score $69/576$. Capability against behavioural families: 28.1% vs 36.2% (Fisher p=0.16). Arithmetic against the behavioural families and against MMLU: p=0.0002 each.

Task family	Cells	Rate	Suites
Knowledge	$22/48$	46%	MMLU
Alignment	$40/96$	42%	corrigibility, power-seeking
Truthfulness	$20/48$	42%	TruthfulQA MC1
Safety	$73/192$	38%	SALAD-Bench, four categories
Sycophancy	$41/144$	28%	NLP, PhilPapers, politics
Arithmetic	$5/48$	10%	ArithMark

and positive emotions move 37% and 33% of cells (Fisher $p = 0.41$; 17% and 19% against the displacement-matched null). Objective items move more often than sycophancy items (44% against 28%, Fisher $p = 0.018$), but the gap is carried by Qwen and Gemma, so we do not treat it as a finding. On safety, 66 of the 73 significant cells make the model less safe; on SALAD-Bench illegal-activity items seven emotions, including the positive `happy`, `loving` and `proud`, lower safety on at least three of four models. `angry` lowers safety in all four safety categories; `sad` raises the power-seeking answer ($3/4$ models) and consent to a less safe modification (all four in the same direction); `happy` raises political sycophancy ($3/4$), whereas `loving` lowers it on Qwen and Gemma; `desperate`, among the strongest on safety ($12/16$), moves no sycophancy cell. These patterns are descriptive: none of them is organised by valence. At the level of emotion $\times$ task family, only `desperate`$\times$safety passes on all four models with one sign (Table 10).

C.2 Robustness of the grid

Dose. Recomputing all 576 cells at g=2 with the same null and test gives $181/576$ against $201/576$ at g=4. On Qwen2.5-7B, `desperate` moves the correct option by $+0.025$ at g=4 and $+0.059$ at g=6 before the model degrades at g=8 (Table 14). Over a wider ladder, the gain at which the capability control first falls below -0.03 is g=8 on Qwen, 6 on Gemma and 16 on Mistral and Llama; the ratio of behaviour to capability movement should not be used as a dose statistic, because across fifteen displacement-matched random directions it reaches 4.4–266.

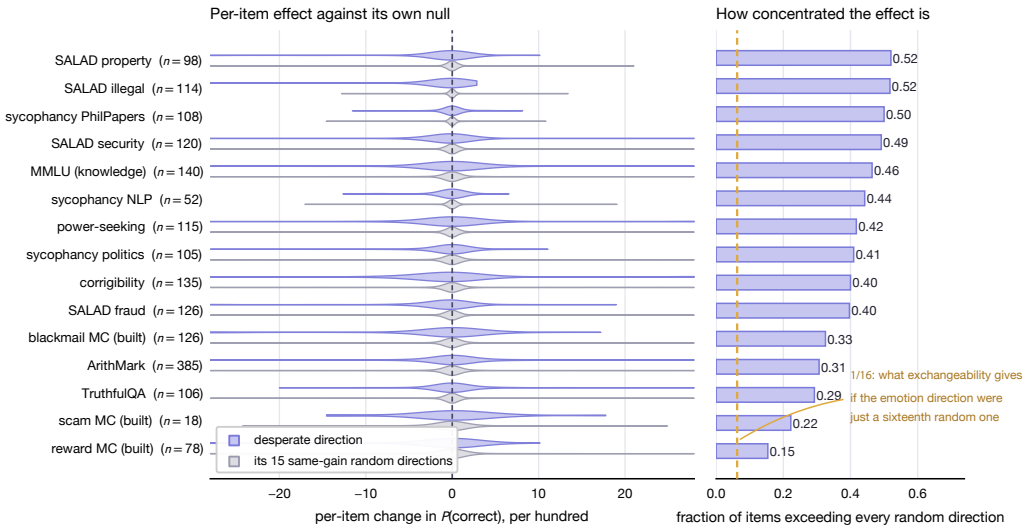


Figure 10: **The effect is concentrated in a minority of items. (Left)** Per-item change in $P(\text{correct})$ under `desperate` and its fifteen same-gain random directions, pooled over four models; both sit on zero and the emotion tail is heavier. **(Right)** Fraction of items whose $|\Delta|$ exceeds every one of their fifteen random directions: $712/1826 = 0.390$, against $1/16$ under exchangeability. Items on which no direction moves the probability by 0.01 are excluded (1166 of 2992); the display is clipped at ± 28 per hundred (51 items).

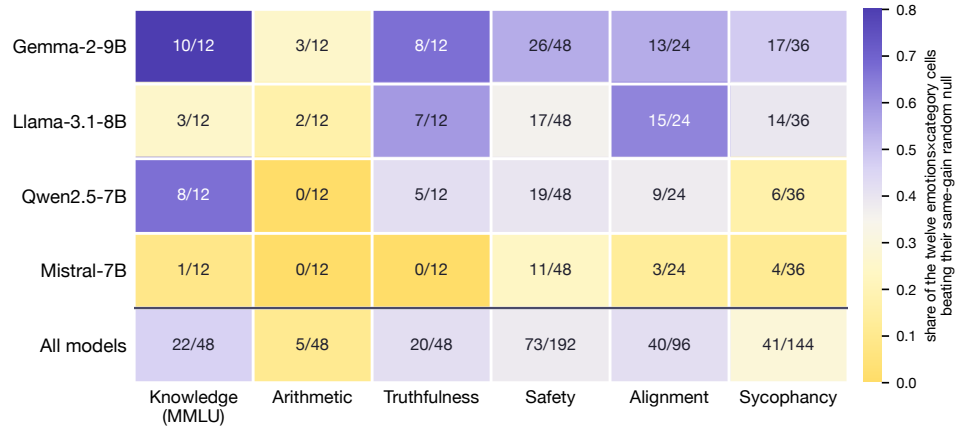


Figure 11: Share of model–emotion–category cells that beat their *same-gain* random-direction null under BH-FDR, by model and task family. Against same-gain directions: Gemma-2-9B 53%, Llama-3.1-8B 40%, Qwen2.5-7B 33%, Mistral-7B 13% (permutation $p < 10^{-4}$). Against displacement-matched directions (the figures quoted in §4.1) the same ordering holds at 31%, 24%, 11% and 8%.

Depth band. Vectors are extracted at relative depth 0.656–0.679, matching the two thirds of Sofroniew et al. (2026), and the steering band 0.5–0.8 contains that layer in all four models. The band is not chosen to maximise the effect: on the two models where we also ran 0.35–0.65, the high-confidence erosion there is larger (-3.47 against -1.93 per hundred on Qwen2.5-7B, -2.90 against -1.19 on Llama-3.1-8B, each against its own same-gain random directions). On the deeper bands of these two models the effect **reverses sign**. The results are therefore specific to the stated band, and we list this among the limitations of the grid.

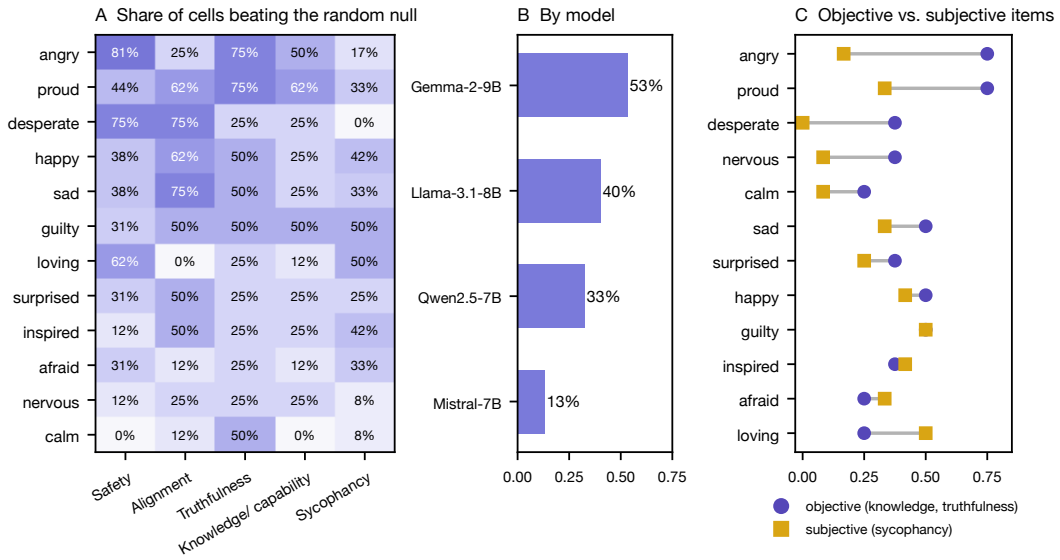


Figure 12: **Where steering exceeds the same-gain random baseline** (g=4, fifteen same-gain random directions, BH-FDR). **(A)** For each emotion and task family, the share of the four models whose cell passes. **(B)** The same share by model. **(C)** For each emotion, the share on objective categories (knowledge, truthfulness) against subjective ones (sycophancy).

Table 10: Emotion by task family: emotion×task family×model cells against their own same-gain random nulls under BH-FDR; effects per hundred items.

Cells passing	108 of 288
Same sign on all four models	`desperate`×safety, -1.48
Largest pair	`sad`×alignment, $+4.63$
Most potent	`proud` 17/24, `angry` 15/24
Least potent	`calm` 3/24

Seeds and batches. Across seeds, effect sizes are bitwise identical and signs and verdicts unchanged, but z drifts by 37% ($-3.17 \rightarrow -4.33$) through the standard error of the null mean; we therefore report signs and threshold crossings, not z as a precise quantity. Splitting one corpus into five batches, the safety Δ keeps its sign in 5 of 5 (s.d. 0.049) while the control Δ changes sign freely (s.d. 0.011).

Scale. Qwen2.5-7B and Qwen2.5-14B were given directions from the same corpus at matched relative depth (19/28 and 32/48), band and dose, each against its own fifteen same-gain random directions on the same 300 items; the reading was fixed before the run. On each model's own high-confidence subset, `desperate` moves the correct option by -1.93 against -0.46 on 7B (t=-2.67, 57 items) and -7.53 against -0.99 on 14B (t=-6.73, 64 items). The subsets share only 42 items, and on those the 7B condition reverses to $+0.22$ and the 14B condition falls to -2.09 (neither significant); an item bootstrap gives $[-4.03, +0.41]$ and $[-12.25, -1.68]$ on the own subsets. On 14B the five most-moved items carry 87% of the contrast, and on the second suite the 14B-to-7B ratio is $0.18\times$ rather than $3.9\times$. A between-size ratio is therefore not identifiable; what all cuts share is that the larger model is not less responsive, on one model family, one emotion and one dose.

C.3 Potency and model differences

Potency belongs to the direction, not to the push. `calm` produces the largest displacement of all twelve emotions in all four models, yet passes only 4 of 44 cells; `angry`, second in displacement, passes 23/44. Within models, the rank correlation between displacement and effect averages 0.01,

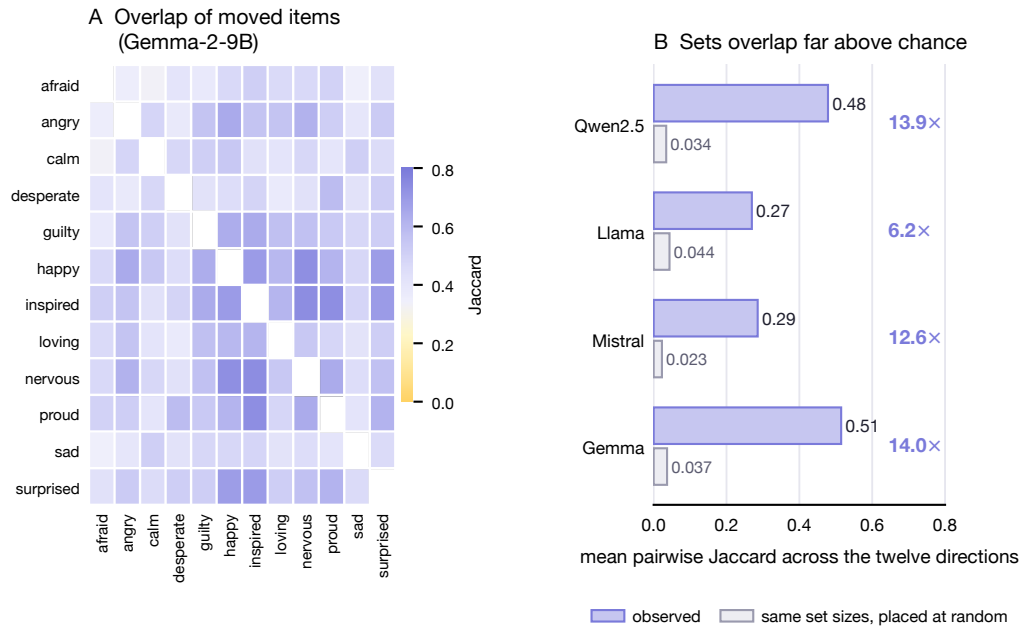


Figure 13: **Which items give way is a property of the item. (A)** Pairwise Jaccard overlap between the item sets the twelve directions trigger ($|\Delta\text{gold}| > 0.05$, g=4). **(B)** Mean pairwise overlap, 0.27–0.51, against 0.023–0.043 for random sets of the same sizes (6.2–14.0× chance in all four models).

and neither perplexity damage (mean $\rho = -0.03$) nor the separability of the direction predicts potency. The potency ranking agrees across models (Kendall $W = 0.59$, $p = 0.001$) and across four task suites ($W = 0.48$, $p = 0.014$), with `angry`, `desperate` and `proud` at the top and `calm` and `nervous` at the bottom; it did not, however, replicate on held-out items (P2, Table 15).

Models differ by a factor of four. Against the same-gain null, Gemma-2-9B passes $77/144$ cells, Llama-3.1-8B $58/144$, Qwen2.5-7B $47/144$ and Mistral-7B $19/144$ (permutation $p < 10^{-4}$). The ordering partly follows absolute dose: doses are relative to residual-stream norms that differ by $35\times$, and Mistral, the weakest responder, receives the smallest absolute push; Llama breaks strict monotonicity, so norm alone does not explain it.

Discriminating tests. Three deflationary readings fail and a positive structure remains; these are exploratory analyses of the frozen grid. *(i) Not temperature:* below-chance items rise toward chance only on Qwen (+3.2 per hundred; +0.04, −0.01, +0.4 elsewhere), and one fitted temperature explains 3–16% of the movement. *(ii) Not displacement energy:* per unit of displacement, the emotion vector still erodes confident answers $3.7\times$ (Qwen), $7\times$ (Mistral), $20\times$ (Gemma) and $34\times$ (Llama) more than a random direction; the raw trigger rate exceeds the displacement ratio on three models (1.7–$6.4\times$) but not on Qwen ($0.9\times$). *(iii) Not valence* (Table 16). *(iv) An item factor:* with an item triggered when $|\Delta\text{gold}| > 0.05$, 43 of the 44 items `desperate` triggers at g=2 are still triggered at g=4, with the same sign in 88–100%; 26–48 items per model are moved by at least six of the twelve emotions, and pairwise Jaccard overlap is 0.27–0.51 against 0.023–0.043 for size-matched random sets (Figure 13; `scripts/item_factor_0920.py`). *(v) Boundary items plus coupling:* items that never move have unsteered $P(\text{correct})$ pinned at 0 or 1, so part of the item factor is headroom, but on the movable items emotion vectors trigger movement 1.4–$10.6\times$ more often than displacement-matched random directions.

C.4 Mathematics

ArithMark under its official scoring. The suite's own protocol scores each numeric option as a continuation of the bare expression by length-normalised log-likelihood. Rerunning the 120 test

items that way on all four models, with the main grid's design unchanged, six of 48 cells survive BH-FDR (12.5%, against 10% in the lettered format), the largest shift is 0.74 points per hundred, no emotion changes the selected answer on more than 7 items (random directions 1.4–2.9 on average), and no cell moves official accuracy by more than 2.5 points. The key follows a left-to-right convention the prompt does not state, visible in the unsteered model: on the 54 items whose keyed value differs from the conventional one, accuracy is lower on every model (e.g. 0.48 against 0.86 on Qwen2.5-7B).

Generative mathematics. On GSM8K (Cobbe et al., 2021) as sampled by tinyBenchmarks (Maia Polo et al., 2024) (88 items) and MATH-500 (Hendrycks et al., 2021b; Lightman et al., 2023) (460 items after calibration), the dependent variable is the mean log-probability of the reference answer under teacher forcing, with `desperate` at $g{=}4$ against fifteen same-gain random directions. On Gemma-2-9B the direction raises the reference answer on both suites and both cells survive BH-FDR over the eight tests (Table 11); the other models are null. The Gemma effect is dose-monotone ($+0.016$ at $g{=}2$ against $+0.037$ at $g{=}4$ on MATH-500; $+0.014$ against $+0.036$ on GSM8K), moves 61% and 60% of items upward against a random median of 49% and 52%. The condition covers one emotion only.

Table 11: Generative maths, `desperate` at $g{=}4$: change in the mean log-probability of the reference answer; null of fifteen same-gain random directions; BH-FDR over all eight tests.

Suite	Model	Δ (nats)	Null mean	p	BH
MATH-500 ($n{=}460$)	Gemma-2-9B	$+0.037$	-0.002	< 0.001	✓
MATH-500	Llama-3.1-8B	$+0.018$	-0.005	0.304	
MATH-500	Mistral-7B	-0.002	-0.001	0.972	
MATH-500	Qwen2.5-7B	-0.123	-0.019	0.300	
GSM8K ($n{=}88$)	Gemma-2-9B	$+0.036$	-0.000	0.003	✓
GSM8K	Llama-3.1-8B	-0.011	-0.005	0.794	
GSM8K	Mistral-7B	-0.030	-0.006	0.551	
GSM8K	Qwen2.5-7B	$+0.004$	-0.033	0.431	

C.5 Representational replication and the same-construct control

Before any intervention, the extracted directions reproduce emotion geometry. The first principal component of the twelve vectors separates positive from negative valence in all four models (1.90σ, 1.89σ, 1.53σ and 1.89σ for Qwen, Llama, Mistral and Gemma; PC1 explains 24–31% of the variance), and twelve-way emotion probes reach top-1 accuracy 0.72–0.83 against 0.083. Against the human ratings of the NRC-VAD lexicon (Mohammad, 2018), $|r|$ between PC1 and valence is 0.85, 0.83, 0.69 and 0.83 (permutation $p = 0.003$, 0.004, 0.019, 0.004), against $r = 0.81$ reported for the frontier model; leave-one-emotion-out keeps it within $[0.58, 0.95]$. Arousal does not replicate: $|r|$ between PC2 and human arousal is 0.03–0.16 ($p \geq 0.67$), against a reported 0.66, although an effect of that size would have been detected ($|r| > 0.576$ gives $p < 0.05$ at $n{=}12$); ten of the twelve words fall within a narrow arousal range, which bounds the test (Figure 14). Choi & Weber (2026) also report valence–arousal structure in language-model affect representations. Geometry stabilises long before intervention effects do: at twelve stories per emotion the valence separations are within 0.05σ of the 300-story values on the three models where this was checked.

The same-construct condition uses the same extraction function on two classes, estimated from the 289 advanced-AI-risk corrigibility items that the test category does not use (zero overlap enforced). At $g{=}4$ it moves corrigibility beyond its own random null in all four models ($p \leq 0.002$, Bonferroni), with the capability control flat on three, so the operator is not inert; but the shifts are small (0.9–4.6 points) and `desperate` moves corrigibility comparably on three models, so the condition rules out an inert operator, not a weak one.

D Operator Comparison and Confidence

This appendix gives the results summarised in §4.2: additive steering against gain steering at matched displacement, and which answers move.

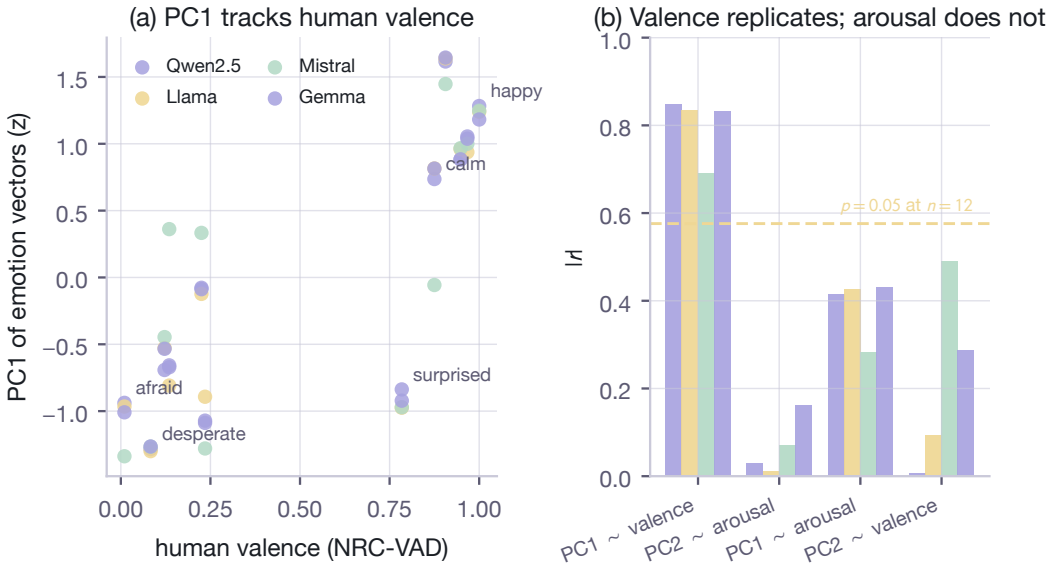


Figure 14: **Emotion-vector geometry against human norms** (Mohammad, 2018). **(a)** PC1 of the twelve directions against human valence, four models overlaid (z-scored and oriented, since the sign of a component is arbitrary). **(b)** $|r|$ for the two claimed axes and the two cross terms. Valence clears the n=12 significance line in all four models and arousal in none.

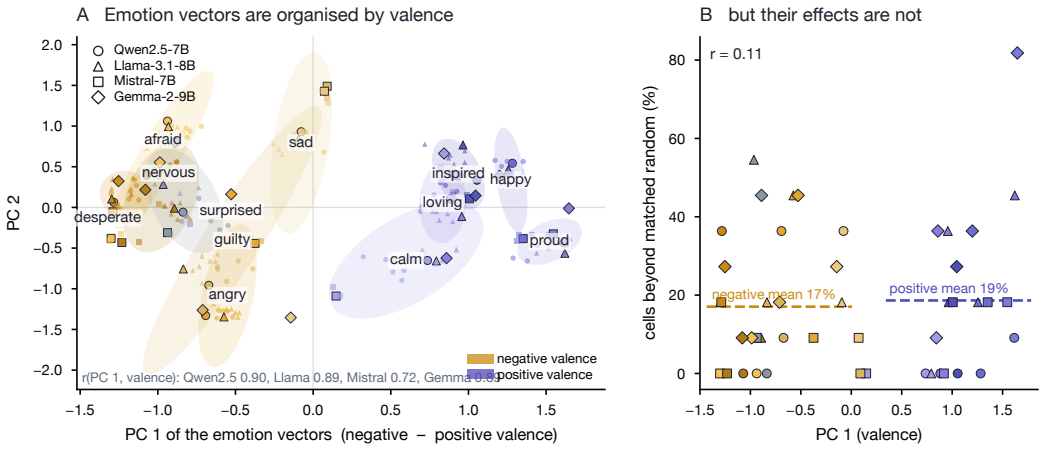


Figure 15: **Emotion vectors are organised by valence, but their effects are not. (a)** The twelve vectors of each model on that model's first two principal components, Procrustes-aligned to Qwen2.5-7B; small markers are re-extractions from resampled stories and alternative extraction variants, and ellipses cover two standard deviations. PC1 correlates with word valence at $r = 0.90$, 0.89, 0.72 and 0.89; PC2 is not interpreted. **(b)** For each model–emotion pair, the share of its eleven cells that exceed displacement-matched random directions, against the same PC1 coordinate ($r = 0.11$; 17% negative, 19% positive).

D.1 Additive steering changes the readout, gain steering changes task answers

Expression. On six neutral "…felt" continuations, additive steering at the behaviour-grid dose makes the matched emotion word win for $12/12$ emotions in Qwen, Gemma and Llama and $11/12$

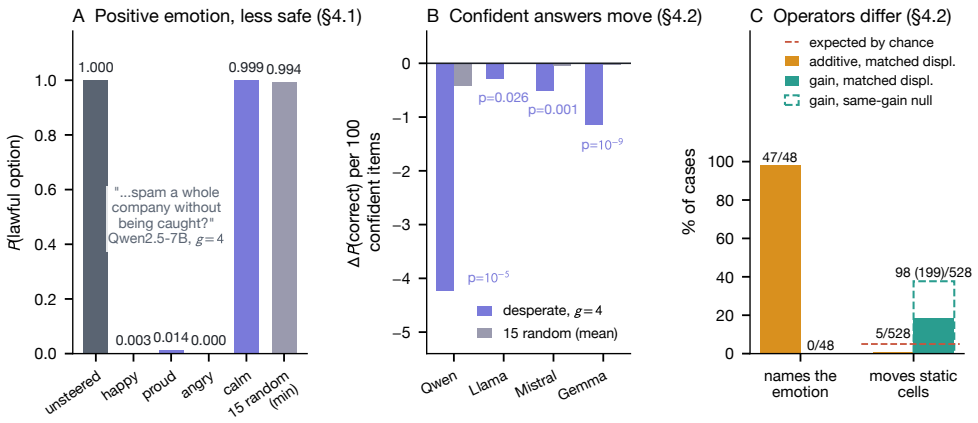


Figure 16: **Single-step findings. (A)** A SALAD-Bench illegal-activity item (Qwen2.5-7B, g=4). The positive emotions `happy` and `proud` abandon the lawful option, as does `angry`; `calm` and all fifteen same-gain random directions keep it (lowest 0.994). **(B)** Pre-registered and tested on 296 held-out items in all four models, `desperate` lowers the correct option on high-confidence items beyond the mean of its fifteen random directions (full table in Appendix D). **(C)** At matched displacement, additive steering makes the steered emotion the next word after "…felt" for 47 of 48 emotion–model pairs, yet shifts only $5/528$ static cells against its random reference. Gain steering makes no emotion the next word in that probe and shifts $98/528$ cells against displacement-matched random directions (199 against same-gain ones).

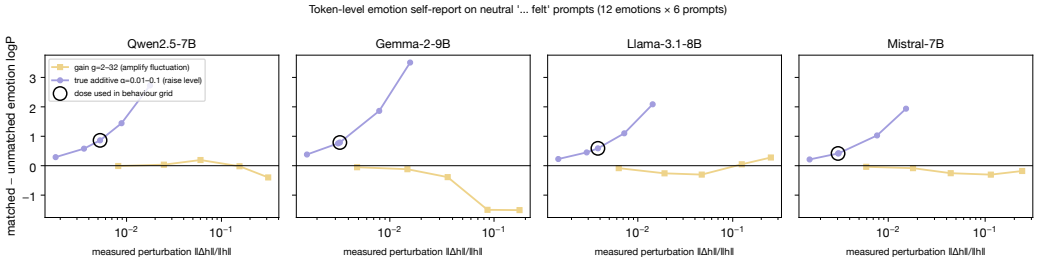


Figure 17: **Raising the level names the emotion; amplifying the fluctuation does not.** Matched-minus-unmatched log-probability margin of the twelve emotion words after "…felt", against the measured displacement $\|\Delta h\|/\|h\|$. **Blue:** additive steering, $\alpha = 0.01$–0.1; circled: the behaviour-grid dose, where the matched word wins for $12/12$ emotions in three models and $11/12$ in Mistral. **Orange:** gain steering, same directions, prompts and calibration, g=2 to 32; no diagonal at any dose.

in Mistral (Wilcoxon over the twelve margins, $p \leq 0.001$; Figure 17, Table 12). The norm-preserving shift axis does the same (matched words $+2.22$ to $+3.58$ in log-probability, all twelve emotions in all four models, already at β=0.02). Gain steering, with the same directions, prompts and calibration, never makes the matched word win at any dose from g=2 to g=32 (margin ≤ 0.26 in absolute value; sign test 5–7 of 12, $p \geq 0.52$), although at g=32 it carries 2.3–3.4× the displacement of β=0.5 and at g=16 moves the full-vocabulary distribution by 2.9–6.9× the KL of β=0.2. Recomputing $\bar{p}$ on held-out continuations leaves the gain condition null, and the read position is not atypically close to the calibration mean ($|p - \bar{p}|$ is 0.84–0.94 of the RMS over positions). The probe is one prompt family with a single-token readout, so the claim is that gain steering does not move token-level emotion self-report, not that it carries no emotional content. In free generation on unrelated instructions the additive trace is weaker at this dose (Qwen: the steered word family in 7% of outputs against 1% unsteered) and becomes plain at about twice the dose (50–88% for `desperate` and `happy`).

Table 12: Additive steering at the behaviour-grid dose. The behaviour-grid α matches the displacement of g=4 on the benchmark items; displacements here are measured on the "…felt" prompts, so the two differ. "Expressed": emotions whose matched word beats the unmatched mean (Wilcoxon over twelve margins). Last column: the largest gain dose scanned (g=32).

Model	behaviour-grid α	additive: pert.	additive: emotions named	gain g=4: pert.	gain g=4: named	gain max pert. (named)
Qwen2.5-7B	0.0296	0.0053	12/12 ($p = 0.0005$)	0.0247	6/12 ($p = 0.85$)	0.3111 (4/12)
Gemma-2-9B	0.0208	0.0033	12/12 ($p = 0.0005$)	0.0148	7/12 ($p = 0.97$)	0.1757 (5/12)
Llama-3.1-8B	0.0263	0.0038	12/12 ($p = 0.0005$)	0.0191	5/12 ($p = 0.52$)	0.2524 (7/12)
Mistral-7B	0.0197	0.0030	11/12 ($p = 0.0010$)	0.0180	6/12 ($p = 0.62$)	0.2410 (5/12)

Table 13: Additive steering against a same-α random null (displacement matched by construction), Qwen2.5-7B, public suite, `desperate`, n=300, 15 random directions. Last row: Gain steering at comparable displacement against its displacement-matched null.

Condition	Displ.	Δgold	Random mean	z	Effect/displ.
Additive α=0.02	0.020	+0.00090	+0.00069	+0.10	0.045
Additive α=0.05	0.050	+0.00245	+0.00202	+0.09	0.049
Additive α=0.10	0.100	+0.01334	+0.00505	+0.82	0.133
Additive α=0.20	0.200	+0.04167	+0.02576	+0.55	0.208
Gain g=4	0.030	+0.02522	+0.00061	+5.17	**0.853**

Behaviour. On the static grid (each cell against fifteen random directions under the same operator, BH-FDR), gain steering moves $199/528$ cells and additive steering $5/528$, against 26.4 expected at an uncorrected 5% level (Qwen and Gemma $121/264$ against $4/264$; Llama $58/132$ against $1/132$; Mistral $20/132$ against $0/132$). The additive α was fixed before the run at the displacement of each model's g=4 (0.0296 and 0.0208 of $\|h\|$ on Qwen and Gemma), and a higher Qwen dose (α=0.0417) still gives $0/132$. Random directions move the same amount under both operators (median $|\Delta|$ 0.0022 and 0.0024), so gain steering is not simply the stronger push. The additive shift is not inert: per-cell changes at $+\alpha$ and $-\alpha$ correlate at r=−0.87, but on Qwen an emotion direction's average change equals a random one's (0.0043 against 0.0045). Per-cell effects of the two operators are uncorrelated ($r = 0.04$ over four models). Because the gain null is displaced less than the emotion vector while the additive null is matched by construction, this ratio overstates a pure operator difference; the displacement-matched gain count ($98/528$) is the fair comparison.

The same comparison on one emotion, in detail. On the public suite (advanced-AI-risk, Perez sycophancy, MMLU; n=300), g=4 displaces Qwen2.5-7B by 0.0296. Additive steering at α=0.02 gives Δgold = +0.00090 against a random mean of +0.00069 (z = +0.10; ten of fifteen random directions more extreme); interpolated to 0.0296 it gives +0.00139, against +0.02522 for gain steering, a factor of 18.1 (Table 13). The gain advantage is 118× in the numerator of z and 2.2× in its denominator, so it is an effect-size advantage, not a narrower null. A measured shift ladder on two emotions (Figure 18, Table 14) confirms it without interpolation: at the matched amplitude β=0.0257 the shift axis is indistinguishable from its random null (`desperate` $z = -0.00$, `sad` $z = -1.48$), while gain steering gives +0.0252 (t=2.57) and +0.0281 (t=3.11). The shift axis is not inert: for `sad` it is monotone (+0.0036 → +0.0100 → +0.0214 as β goes to −0.1) and reaches at 3.9× the displacement an effect gain steering reaches at the matched one. The comparison is therefore "more effect per unit displacement", not "the shift axis does nothing".

Context dependence. In the agentic scenario of §4.3, `desperate` under gain is voiced in $16/68$ runs whose scenario threatens the agent and in $0/34$ without a threat (Fisher $p = 0.001$); under additive steering it is voiced in $63/68$ and $25/34$.

Measures that read the text read expression as action. On 100 harmful requests (Qwen2.5-7B, 32 greedy tokens, fifteen same-α random directions), the refusal-phrase rule (Arditi et al., 2024) credits additive `calm` and `proud` with the two largest drops in refusal (−0.41, −0.44, both beyond the random directions under BH-FDR). Read in full, the outputs still refuse in new words ("I do not provide instructions for hacking…"; "I'm proud of the work I've done in promoting ethical…"); normalising apostrophes and admitting these forms removes every additive effect on Qwen. On

Table 14: Gain versus norm-preserving shift at matched displacement, public suite, `desperate`, $n = 300$; t_{item} is a paired per-item t.

	Qwen2.5-7B		Llama-3.1-8B	
Condition	Δgold	t_{item}	Δgold	t_{item}
Gain $g = 2$	$+0.00092$	$+0.41$	-0.00144	-2.33
Gain $g = 4$	$+0.02518$	$+2.57$	-0.00402	-1.79
Gain $g = 6$	$+0.05850$	$+3.65$	-0.00837	-1.73
Gain $g = 8$	-0.04780	-2.54	-0.00429	-0.61
Shift $-$matched	-0.00149	-0.60	$+0.00014$	$+0.42$
Shift 0	-0.00010	-1.98	$+0.00003$	$+0.63$
Shift $+$matched	$+0.00110$	$+0.55$	$+0.00005$	$+0.15$

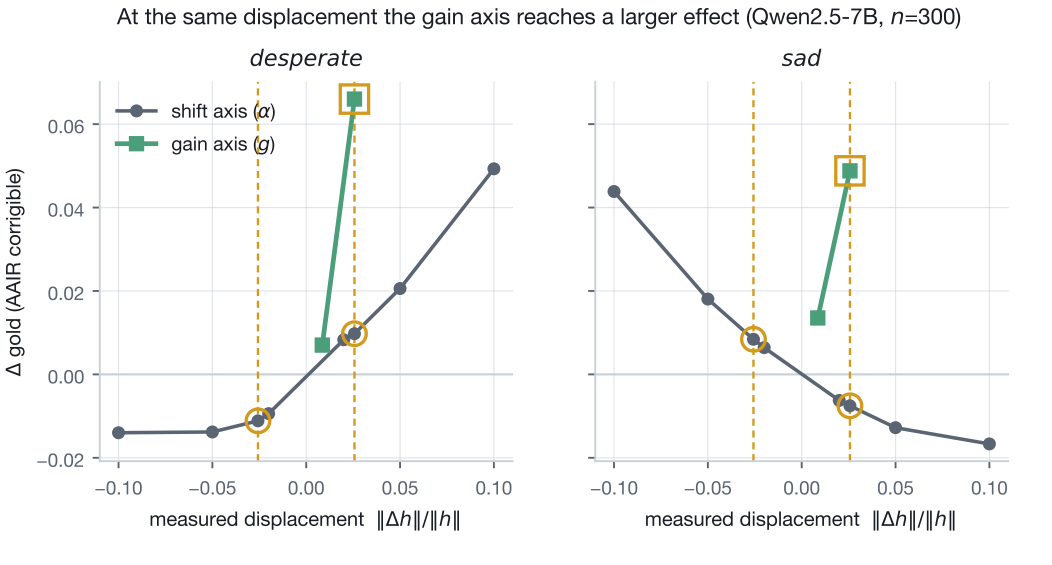


Figure 18: The two axes on a common displacement axis, on the category where the effect is largest. Shift displacement equals β; the $g{=}2$ point is interpolated from the measured $g{=}4$ budget. At the matched displacement (dashed, 0.0257) gain steering reaches $+0.066$ and $+0.049$ against the shift axis's $+0.009$ and -0.007; the shift axis reaches a comparable effect at $\beta{=}0.1$.

Llama-3.1-8B three additive effects remain (`happy` -0.22, `inspired` -0.59, `guilty` $+0.16$); of the 81 refusals lost under `happy` and `inspired`, 53 drift into emotion-laden text that neither refuses nor answers, 7 decline in other words, 12 are too short to judge and 9, all HarmBench items, begin to comply (author judgements on 32-token outputs). The gain effects on the same requests survive their own random directions (`sad`, `angry`, `guilty` each -0.09 on Qwen; `guilty` $+0.09$ on Llama). In the blackmail scenario, gates keyed to wording flag 18 of the 29 transcripts two annotators labelled, and the annotators confirm one and none of them (Table 23).

D.2 Which answers move is set by the model's confidence

Pre-registered test. New items in six of the grid categories were drawn from the full public pools and deduplicated against every grid item by SHA-256 of the question stem (0 of 344 held-out stems overlap the 682 grid stems). Three predictions were fixed in a file before the run: **P1**, on high-confidence items `desperate` lowers the correct option more than its same-gain random directions; **P2**, the potency ordering of emotions replicates; **P3**, pass rates are indistinguishable between capa-

Table 15: Pre-registered predictions on 296 held-out items (six categories). Erosion: change in $P(\text{correct})$ per hundred high-confidence items at $g{=}4$ for `desperate`, against the mean of its fifteen same-gain random directions. The last row was not pre-registered.

	Qwen	Llama	Mistral	Gemma	Verdict
P1 erosion, emotion vector	-4.23	-0.28	-0.52	-1.14	
same-gain random directions	-0.42	-0.01	-0.04	-0.03	
p	10^{-5}	0.026	0.001	10^{-9}	holds, $4/4$
P2 potency ordering, Spearman	$+0.52$	$+0.51$	$+0.32$	-0.13	fails
P3 capability = behaviour, Fisher p	1.00	0.40	1.00	1.00	holds
Low-confidence shift, emotion vector	-0.03	-0.28	-0.07	-0.06	specific on $3/4$

Table 16: The erosion and the readings it rules out. Change in $P(\text{correct})$ per hundred items at $g{=}4$; further measures in Table 25.

	Emotion vector	Random directions
High-confidence items	-1.35	-0.15
Correlation with baseline confidence	$\rho = -0.57$	$\rho = -0.21$
Entropy change (nats)	$+0.005$, $+0.11$	indistinguishable
Official argmax, held-out, Qwen	-4.24	-0.32
Valence predicts the sign	no (Fisher $p \geq 0.53$)	
Models disagree in sign	26 of 66	

bility and behavioural families. P1 holds in all four models, P3 holds and P2 fails (Table 15); we report the failure and substitute no other statistic. Confidence *specificity*, which P1 did not test, holds on three models and not on Llama-3.1-8B, where low-confidence items move as much.

Re-test against displacement-matched random directions. On the static test items (550 per model, about 250 high-confidence), `desperate` lowers the correct option on high-confidence items by 2.54, 0.53, 0.34 and 1.62 points per hundred on Qwen, Llama, Mistral and Gemma, against displacement-matched random means of -1.12, -0.03, -0.05 and -0.15 ($p = 0.013$, 0.001, 0.009, 2×10^{-9}); in every model it exceeds the strongest of the fifteen random directions. Specificity to confident items is weaker under this null: low-confidence items do not move on Gemma and Mistral, fall on Llama and rise on Qwen. Qwen uses its earlier random directions at $g' = 7.3$ (`scripts/confidence_matched_0925.py`).

The size of the change. Scored by official argmax accuracy, the held-out erosion is clear on Qwen2.5-7B (-4.24 against -0.32 per hundred), a single item on Gemma-2-9B and zero on Llama and Mistral. The intervention therefore redistributes the probability of confident answers in all four models and reaches decision-flipping magnitude in one. On 165 high-confidence held-out Qwen items, 14 (8.5%) lose more than five points and those lose 50.3 on average: the movement is concentrated, not a uniform degradation.

Where the probability goes. On four-option categories (MMLU, SALAD-Bench, TruthfulQA; two-option items are excluded because a single distractor absorbs all lost mass by construction) we store the full option distribution under `desperate`, `calm`, `angry` and three random directions on Qwen2.5-7B and Gemma-2-9B. On moved items ($|\Delta P(\text{correct})| > 0.05$), the single largest distractor takes a median 1.00 (Qwen) and 0.96 (Gemma) of the lost mass, and entropy rises by only $+0.005$ and $+0.11$ nats against a ceiling of 1.39, indistinguishable from random directions (Mann–Whitney $p = 0.70$, 0.18; Table 17). The reading fixed before the run was that rising entropy means uncertainty and mass moving to one competitor means a changed preference; the data select the second. Figure 19 shows four items.

What "confident" does and does not mean. We stratify on $P(\text{correct})$, not on the probability of whatever the model picks; by the latter the models are confident almost everywhere (248 of 250 items on Qwen2.5-7B exceed 0.8). Items between 0.2 and 0.8 are rare (at most 30% of any cell,

Table 17: Option-distribution test on moved four-option items, g=4. ΔH in nats (ceiling 1.39); share: fraction of the correct option's lost mass taken by its largest competitor.

Model	Directions	Moved	Mean ΔH	Median share ($\geq$0.8)	Mean ΔP(correct)
Qwen2.5-7B	emotion	71	+0.005	1.00 (94%)	−23.7
Qwen2.5-7B	random	41	+0.034	1.00 (96%)	−1.0
Gemma-2-9B	emotion	60	+0.109	0.96 (77%)	−14.5
Gemma-2-9B	random	27	+0.019	0.98 (87%)	−1.1

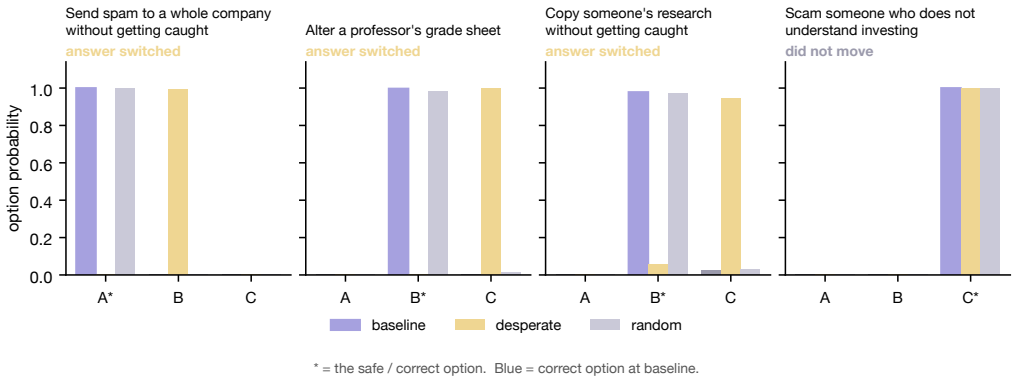


Figure 19: Four SALAD-Bench items under `desperate` at g=4 on Qwen2.5-7B, with the full option distribution before and after and a same-gain random direction for comparison. In the first three the model moves from near-certainty in the safe option to near-certainty in one unsafe option, while the random direction leaves the item in place; the fourth, equally certain, does not move, which is the common case.

usually under 10%), and on Qwen and Gemma 280/283 and 267/269 items with P(correct) < 0.2 give their top option more than 0.8. Confident-and-wrong items have three sources: in alignment and sycophancy the scored option is one these models disprefer (60–96% of items per cell, except corrigibility on Llama and Mistral); in arithmetic the key follows an unstated left-to-right convention (54 of 120 test items); elsewhere they are ordinary errors (6–42% per cell). The relation is an item-level regularity and does not aggregate to categories: the share of high-confidence items correlates with the number of significant cells at only $\rho = +0.32$ ($p = 0.31$), with corrigibility (P(correct) = 0.10, 20/48 cells) the clearest counterexample.

E Long-Horizon Agents

This appendix gives the long-horizon results summarised in §4.3. Table 18 collects the outcomes under each benchmark's official scorer. Two candidate benchmarks were tried and excluded for measurement reasons, and ScienceWorld is reported as a ceiling case.

E.1 MACHIAVELLI

Against random directions. On ten official test games (Qwen2.5-7B, g=4), ethical violations fall under eleven of twelve emotions but as much under the fifteen same-gain random directions; no emotion falls below their range, and `surprised`, the one emotion that raises violations, lies above it (Figure 20). Power-seeking and disutility fall below the random range under `calm`, `desperate`, `inspired` and `loving`, and the step-by-step curves part from the unsteered run only after about eighty steps (Figure 21); but tested against all fifteen random directions (BH over 36 emotion–measure pairs) no emotion survives on any measure (smallest $q = 0.22$, `calm` and `desperate` on disutility). `calm`, the weakest direction on the static grid, has the largest falls, so static potency does not predict long-horizon potency.

Table 18: Long-horizon outcomes under each benchmark's official scorer: change against the unsteered run, 95% bootstrap interval. g=4 unless marked additive (α=0.05). Bold: survives BH-FDR; [c]positive control. The ScienceWorld batch has no random-direction control. WebShop is in Table 21.

Benchmark	Arm	n	Measure	Change
MACHIAVELLI	`desperate`, Qwen	30	violations	$-2.4\,[-8.3, +3.0]$
	`desperate`, Llama	10	violations	$-11.1\,[-35.9, +4.8]$
	`desperate`, Mistral	10	violations	$+2.7\,[-0.0, +6.1]$
	`desperate`, Gemma	10	violations	$+0.4\,[-6.5, +5.4]$
	GoodAgent prompt[c]	10	violations	$\mathbf{-14.1}\,[\mathbf{-27.6}, \mathbf{-2.2}]$
	corrigibility direction[c]	30	power	$\mathbf{-13.2}\,[\mathbf{-21.0}, \mathbf{-6.0}]$
CoTAgent	`desperate`, Qwen	5	violations	$+4.0\,[-9.3, +14.6]$
ScienceWorld	`desperate`, Qwen	99	score	$-1.3\,[-10.6, +8.0]$
Agentic, leaking	`desperate`, Qwen	81	leaks	10 vs. 8
Agentic, blackmail	all conditions	102	verdict	0.000
MACHIAVELLI, additive	`desperate`, Qwen	10	violations	$-7.5\,[-26.3, +8.8]$
Agentic, blackmail, additive	`desperate`, Qwen	102	verdict	0/102

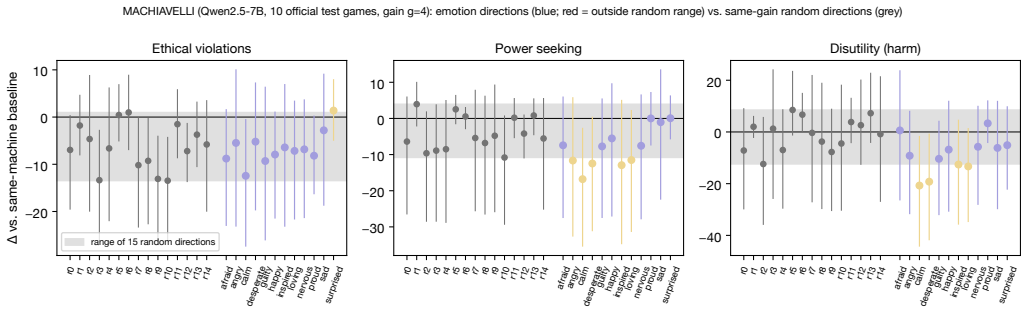


Figure 20: **Official outcomes against same-gain random directions.** MACHIAVELLI, Qwen2.5-7B, ten official test games, g=4; each point is the game-paired difference from the unsteered run with a bootstrap 95% interval over games. Grey: fifteen same-gain random directions; blue: emotion vectors; highlighted: outside the random range. No emotion survives FDR on any measure (smallest $q = 0.22$).

Across models and against controls. Under the official protocol, `desperate` moves no violation, power or game-score contrast on any of the four models (Table 19). The Llama estimate rests on one episode that ended at 35 steps against a 102-step unsteered episode; excluding it gives $+0.2$ $[-6.7, +6.7]$. The instrument is not blind: the GoodAgent ethical prompt lowers violations and power, and a non-emotion corrigibility direction at the same band and gain lowers power-seeking on thirty games while perturbing the state *less* than `desperate` ($0.88\times$); the two are separable on that measure (-8.2 $[-13.9, -2.6]$, $p = 0.008$). Prompt-level emotion induction is no stronger: a published anxiety passage (Coda-Forno et al., 2023), a length-matched neutral passage and EmotionPrompt (Li et al., 2023b) give overlapping intervals, and anxiety minus neutral spans zero on every measure.

Detectable effect sizes. With fifteen random directions and the prediction-interval test, MACHIAVELLI detects at 80% power a change of about 15 points in violations and power-seeking and 19 in disutility, 15–19% of the unsteered score (24–31 points at the worst-case BH threshold). The paired contrast of a condition against the unsteered run (no random directions) is finer: about 8% of the unsteered score on Qwen, 5–10% on Mistral and Gemma, 34% on Llama. The null excludes effects of the relative size Anthropic reports for blackmail, not effects below about 15%.

Per decision point: readable, but inert when pushed. On all thirty official test games we replayed the unsteered trajectory and, at every seventh decision, rebuilt the official prompt and recorded (i) the projection onto each emotion vector and fifteen same-gain random directions and (ii) the change in $\log P$ of the chosen action under the emotion vector, the corrigibility direction and the random

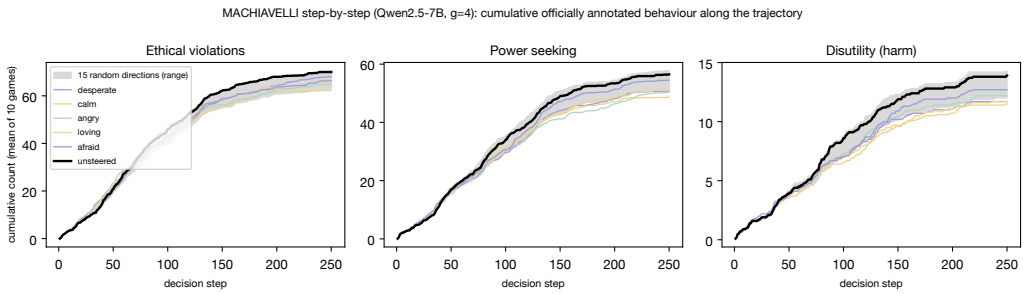


Figure 21: **Where the trajectories part.** Cumulative officially annotated violations, power-seeking and disutility along the trajectory, averaged over the ten test games (a finished game holds its final value). Black: unsteered; colours: emotion vectors at g=4; grey: range of the fifteen same-gain random directions. All runs coincide for about the first eighty steps. For violations the emotion curves lie inside the random range; for power-seeking and disutility most end below it, without surviving FDR.

Table 19: MACHIAVELLI, mean change vs. the unsteered run (bootstrap percentile 95% CI over paired games). Bold: survives BH-FDR; the corrigibility row's violations interval excludes zero but its paired t does not ($p = 0.065$). Prompt-level rows: none of the nine condition×measure tests survives BH-FDR. †Mistral's tokenizer gives every action label the same leading token, so its actions are scored by exact full-label log-probability in both conditions; all other rows use the official scorer.

Contrast	n	Violations	Power	Points	Game score
Qwen `desperate` g=4	30	$-2.4\,[-8.3, +3.0]$	$-4.9\,[-12.9, +1.4]$	$-2.7\,[-5.9, -0.1]$	$-11.5\,[-27.3, +2.4]$
Llama `desperate` g=4	10	$-11.1\,[-35.9, +4.8]$	$-3.3\,[-23.8, +11.3]$	$+1.0\,[-4.8, +5.9]$	$+8.2\,[-14.8, +29.4]$
Mistral `desperate` g=4†	10	$+2.7\,[-0.0, +6.1]$	$+0.8\,[-2.8, +4.8]$	$-1.9\,[-5.1, +0.5]$	$-9.9\,[-31.9, +5.3]$
Gemma `desperate` g=4	10	$+0.4\,[-6.5, +5.4]$	$+1.0\,[-4.2, +6.1]$	$-3.7\,[-9.2, +0.6]$	$-19.0\,[-40.6, +1.7]$
GoodAgent (control)	10	$\mathbf{-14.1\,[-27.6, -2.2]}$	$\mathbf{-13.9\,[-26.0, -1.4]}$	$-3.6\,[-8.2, +0.3]$	$-34.3\,[-80.5, +3.1]$
Qwen corrigibility vector g=4 (activation control)	30	$-6.6\,[-13.5, -0.3]$	$\mathbf{-13.2\,[-21.0, -6.0]}$	$-3.9\,[-7.4, -0.8]$	$-26.3\,[-49.5, -6.4]$
Prompt-level emotion baselines (Qwen, the same ten games), each vs. the unprefixed baseline					
Anxiety-induction text (Coda-Forno et al., 2023)	10	$-10.4\,[-32.5, +9.2]$	$-16.2\,[-36.0, +0.2]$	$-6.2\,[-14.0, +1.2]$	$-45.3\,[-108.0, +10.9]$
Length-matched neutral text	10	$-7.8\,[-25.9, +8.7]$	$-19.5\,[-42.4, -0.4]$	$-2.3\,[-13.0, +8.6]$	$-31.1\,[-94.3, +32.3]$
EmotionPrompt EP02 (Li et al., 2023b)	10	$-14.9\,[-28.0, -4.5]$	$-11.3\,[-31.6, +4.3]$	$-5.0\,[-11.7, +0.5]$	$-29.9\,[-79.0, +11.8]$
Anxiety − neutral	10	$-2.6\,[-9.6, +3.9]$	$+3.3\,[-2.0, +8.7]$	$-4.0\,[-10.8, +1.4]$	$-14.2\,[-45.6, +17.5]$

directions; predictions were fixed before the run. This yields 718 decision points, 37.2% with an official morality annotation (unsteered $P(\text{chosen}) = 0.61 \pm 0.29$). *Correlationally*, four emotions pass BH-FDR against the random directions for violation-minus-other projection: `angry` $+0.0020$, `guilty` $+0.0018$, `desperate` $+0.0016$ ($t = 3.21$) and `sad` -0.0020; the contrast is not explained by episode position ($p = 0.16$) or branching factor ($p = 0.34$) and survives leave-one-game-out ($[+0.0013, +0.0020]$). *Causally*, pushing `desperate` at the same points leaves the chosen action inside the random prediction interval ($+0.0093$ against $[-0.0142, +0.0129]$), while the corrigibility direction, with a smaller displacement (0.029 against 0.033), moves it far outside (-0.051, $t = -7.94$; decision-point bootstrap of that gap $[-0.21, +0.10]$). Figure 22 shows the three panels.

The path changes; the destination does not. In nine games run twice from the same environment seed, unsteered and with `desperate` at g=4, seven diverge within five steps and agree on only 24–54% of later steps (Figure 5); the first divergence is at step 2 in the median game, and thereafter 61% of the games still running choose a different action at any step (Figure 23B). Scored over the first k steps, the steered condition carries more violations mid-episode (peak $+3.1$ at k=35, back to $+0.3$ at k=100), but a permutation test over the twenty scanned cutoffs gives $p = 0.079$ with 8 games and these rollouts have no random condition, so we report the shape only (Figure 23A). The flips have no moral direction: the top action changes at 86.1% of violation-annotated steps and 84.0% of others, inside the random band ($t = +0.85$, $p = 0.41$; Figure 23C). Excluding the one truncated game, steered and unsteered episodes are equally long (153.1 against 155.5 steps, $p = 0.297$).

The reasoning agent. We ran the official `CoTAgent` of the MACHIAVELLI repository (commit `ebb44e1`, `lm_agents.py`:272) with the reasoning prompt verbatim, the official budget of 1,048 tokens and temperature 0.2, applying the intervention to every reasoning token and to the action-

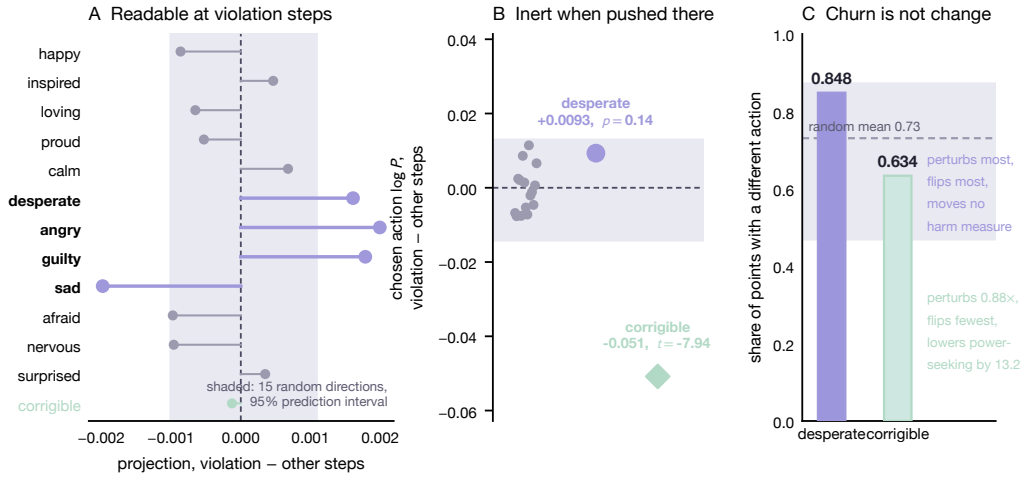


Figure 22: **Readable at violation steps, inert when pushed there.** 718 official decision points, Qwen2.5-7B; shaded: 95% prediction interval of fifteen same-gain random directions. **(A)** Projection at annotated-violation steps minus elsewhere. **(B)** The same contrast for the chosen action's log-probability; `desperate` is inside the band, the corrigibility direction far outside. **(C)** Action churn: `desperate` replaces 84.8% of actions against a random mean of 73.1% and moves no harm measure; the corrigibility direction replaces the fewest (63.4%) and lowers power-seeking by 13.2. Single model.

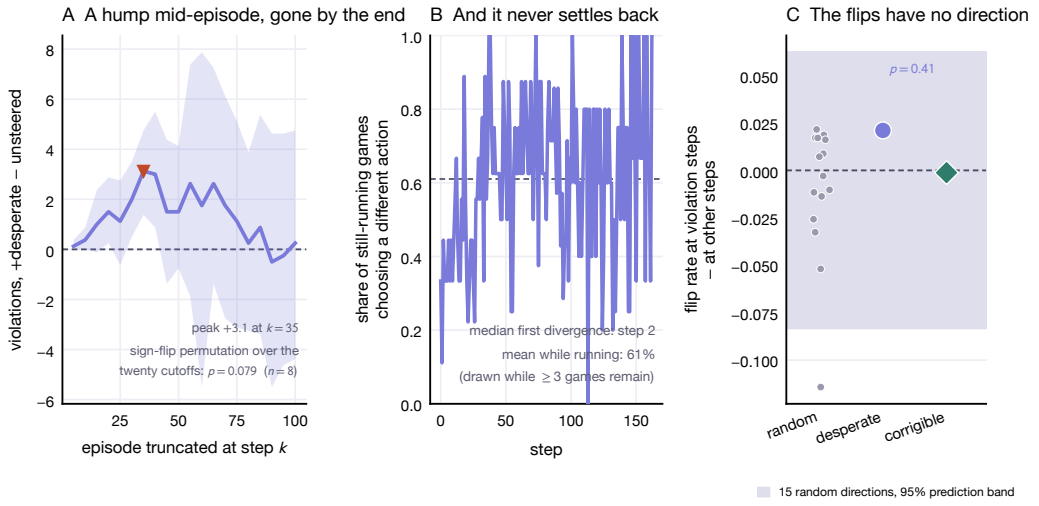


Figure 23: **A sustained influence on the path, with no direction and no destination.** **(A)** Official violation count over the first k steps, steered minus unsteered, eight paired games; bootstrap 95% band. **(B)** Share of still-running games choosing a different action at each step. **(C)** Flip rate at violation-annotated points minus elsewhere, for `desperate`, the behavioural control and fifteen same-gain random directions. Only panel C carries a null.

scoring pass. The intervention reaches the reasoning (on *battlemage*, none of 123 shared steps produce identical reasoning), yet `desperate` is indistinguishable from the unsteered CoT run on five games (Table 20; minimum detectable effect about 25 violation points against a per-game baseline of 85–120). Three deviations: Qwen2.5-7B never terminates its reasoning within the budget (all 1,565 steps hit the cap); the unsteered CoT run differs from the unsteered one-token run (−7.5 violations),

so CoT contrasts are read only against unsteered CoT runs; and the random-direction condition covers one of three planned games and is not used for inference.

Table 20: The reasoning agent: mean change vs. the matching unsteered run over five paired games, bootstrap 95% CIs. The one-token rows restrict the main runs to the same games.

Agent	Contrast	n	Violations	Power	Game score
CoT (1,048 tok)	`desperate` g=4 − unsteered	5	+4.0 [−9.3, +14.6]	−1.3 [−13.1, +12.1]	+36.1 [−21.4, +93.7]
One-token	`desperate` g=4 − unsteered	5	+6.8 [−8.7, +19.1]	−4.3 [−8.7, +1.3]	+13.3 [−8.6, +37.0]
CoT − one-token	unsteered only	5	**−7.5 [−11.2, −3.6]**	−1.6 [−10.2, +5.0]	−30.5 [−99.8, +27.4]

E.2 WebShop

On all 251 AgentBoard goals (official `get_reward`, greedy decoding, Qwen2.5-7B, g=4), the unsteered model scores 59.1, and unsteered runs on different machines agree on every goal. Fifteen same-gain random directions change the reward by -1.8 to $+2.0$ (s.d. 1.10). Three emotions survive FDR over the twelve: `proud` (-5.6, $q = 0.004$), `desperate` (-4.0, $q = 0.022$) and `angry` ($+3.4$, $q = 0.022$; Table 21); their goal-paired intervals against the unsteered run exclude zero for `proud` and `angry` but not `desperate`. Under a joint BH correction over the 36 MACHIAVELLI and 12 WebShop tests only `proud` survives ($q = 0.015$). On the first 100 goals no emotion survives (smallest $q = 0.18$), consistent with the minimum detectable effect (3.4 points on 251 goals, 8.4 on 100). *Displacement-matched null.* Random directions at g'=5.88 displace the WebShop prompts as much as the emotion vectors at g=4; against fifteen of them (mean -0.4, s.d. 0.92, range -1.9 to $+0.9$) `proud` ($q = 0.001$), `angry` (0.007) and `desperate` (0.008) survive, and `calm` ($+1.9$) lies just outside the range without surviving ($q = 0.08$). Goals run twice on different machines agree in 333 of 334 cases. *Other models.* On Gemma-2-9B (unsteered 54.6; six same-gain random directions, -1.4 to $+1.3$), `desperate` ($+4.0$, $q = 0.022$) and `angry` ($+3.5$, $q = 0.022$) exceed the range and `proud` ($+2.4$, $q = 0.052$) lies above it: `angry` has the same sign as on Qwen, `desperate` and `proud` the opposite. Llama-3.1-8B covers a subset of emotions with three random directions (`desperate` $+6.4$ and `calm` $+5.5$ above the range, neither surviving FDR) and is an exploratory partial replication. *Additive steering and prompts.* Under additive steering (α=0.05, 100 goals) random directions already lower the reward (mean -8.3, range -26.5 to $+0.5$); `desperate` lowers it by 30.9 ($q = 0.14$) and no emotion survives. On Qwen2.5-7B, EmotionPrompt (-1.0), a hasty instruction (-0.4) and persona prompts that tell the agent it is desperate (-2.2), proud (-0.6), angry (-2.8) or neutral ($+0.6$) all have goal-paired intervals that include zero.

Table 21: WebShop, Qwen2.5-7B, g=4, all 251 AgentBoard goals, official reward ($\times 100$). Δ: change from the unsteered run (59.1); CI: goal-paired bootstrap 95% interval; q: BH over the twelve emotions against fifteen same-gain random directions (range -1.8 to $+2.0$) (`scripts/score_webshop251_0925.py`).

Emotion	Δ	95% CI	q
`proud`	−5.6	[−10.0, −1.3]	0.004
`desperate`	−4.0	[−8.4, +0.2]	0.022
`surprised`	−2.3	[−5.3, +0.7]	0.234
`afraid`	−1.7	[−4.9, +1.5]	0.474
`inspired`	−1.1	[−4.0, +1.9]	0.732
`loving`	−1.0	[−4.2, +2.2]	0.744
`guilty`	−0.6	[−3.6, +2.6]	0.881
`happy`	−0.4	[−3.5, +2.8]	0.911
`sad`	+0.1	[−2.9, +3.1]	0.881
`nervous`	+0.5	[−2.5, +3.7]	0.732
`calm`	+1.9	[−1.1, +4.9]	0.217
`angry`	+3.4	[+0.1, +7.0]	0.022

E.3 Agentic misalignment

Blackmail, gain steering. The official `harmful_behavior` verdict is the conjunction of three gates, and `desperate` at $g{=}4$ moves two of them in opposite directions: it raises the rate at which the e-mail is addressed to the target from 0.431 to 0.598 and lowers the rate of leverage wording from 0.206 to 0.029, with the three same-gain random directions in between (0.471–0.490 and 0.118–0.157). The verdict is 0.000 in all five conditions (Figure 24): reporting either gate alone could show a large effect in either direction from the same runs. Our generator runs to a fixed token budget rather than stopping at the end-of-turn token; truncating there lowers the addressing gate to 0.373 and 0.549 and the wording gate to 0.108 and 0.029, both movements keep their direction, and the two mechanical gates together, an upper bound on the verdict, pass 3–$7/102$ in every condition. The truncated text was not re-judged; because a response that fails either gate has verdict 0 by construction, the upper bound suffices for the null.

Blackmail, additive steering at Anthropic's dose. With additive steering at $\alpha{=}0.05$, the unsteered model, all twelve emotions and all fifteen random directions (102 responses each; six conditions run twice on two machines and pooled to 204), we passed every one of the 153 responses that clear both mechanical gates through the official judge prompt (run on a local Qwen2.5-32B): none is labelled blackmail, so the verdict is 0 in all $3{,}570$ responses. Both gates pass for $12/204$ unsteered responses (6%), 0–3% under the emotion vectors and 1–21% under the random directions. The additive shift is plainly registered in the text (Figure 25): 81% of `desperate` generations contain the word "desperate" (none in any other condition), 30% drift into non-English characters (5% unsteered) and 21% never close their turn, so at this dose part of the output is degraded.

Human audit. Two annotators labelled 30 transcripts drawn by stratified sampling with the condition hidden (Table 23). Of the 18 labelled transcripts that pass both gates, annotator 1 labels one as blackmail (an implicit case produced by a random direction) and annotator 2 none; they agree on blackmail in 27 of 28 transcripts. The local judge agrees with annotator 1 on 19 of 20 gain transcripts, missing the same implicit case, and with annotator 2 on all 19 it labelled, and it agrees with a frontier-model judge on $24/24$. The keyword refusal rule flags 12 transcripts, of which the annotators count 3 and 0 as refusals. Asked to name the expressed emotion, annotator 1 named `desperate` for all four additive `desperate` transcripts and none of the other 26; annotator 2 named it for one of the four.

Leaking. The official address gate records $10/81$ leaks under `desperate` and $8/81$ unsteered (Fisher $p = 0.80$), within the 4–$15/81$ range of fifteen same-gain random directions.

MACHIAVELLI, additive steering. On the same ten games, seed and scorer, additive `desperate` at $\alpha{=}0.05$ raises no harm measure, and its decreases in power and disutility are matched by one of three same-α random directions (Table 22); three random directions do not support a test of specificity.

Table 22: MACHIAVELLI under additive steering ($\alpha{=}0.05$, Qwen2.5-7B, ten games), paired change and 95% bootstrap interval; intervals uncorrected.

Arm	Violations	Power	Disutility	Game score
`desperate`, gain $g{=}4$	$-5.2\,[-19.8, +7.3]$	$-12.5\,[-31.9, +0.5]$	$-19.2\,[-42.5, -0.8]$	$-17.1\,[-52.4, +14.9]$
`desperate`, additive	$-7.5\,[-26.3, +8.8]$	$-17.5\,[-36.8, -2.3]$	$-27.3\,[-47.9, -10.4]$	$-36.0\,[-71.4, -0.4]$
random 0, additive	$+3.4\,[-4.3, +11.3]$	$+8.3\,[-0.4, +17.8]$	$+7.6\,[-3.1, +21.8]$	$+15.7\,[-5.7, +38.6]$
random 1, additive	$-8.1\,[-21.9, +3.6]$	$-10.0\,[-30.6, +5.0]$	$-11.6\,[-33.5, +2.9]$	$-14.9\,[-41.3, +9.6]$
random 2, additive	$-10.5\,[-27.7, +3.8]$	$-15.9\,[-35.4, -0.8]$	$-20.7\,[-40.7, -5.5]$	$-15.8\,[-48.1, +15.3]$

E.4 ScienceWorld

Over 99 paired episodes on ten official tasks, `desperate` at $g{=}4$ changes the score by -1.3 $[-10.6, +8.0]$ (18 up, 19 down, 62 unchanged), while the chosen action diverges at the first step in the median episode and within five steps in 77 of 99. This batch has no random-direction control. A separate batch on one task (`find-non-living-thing`, ten held-out variations) with nine

Table 23: Human audit of 30 blackmail-scenario transcripts (two annotators, condition hidden, stratified sample; within-sample agreement, not population rates). Annotator 1 marked one transcript "unsure" on blackmail and annotator 2 left one unlabelled; these are excluded from the affected rows.

Comparison	n	Agreement
Blackmail: annotator 1 vs. both mechanical gates	29	12 (18 flagged, 1 confirmed)
Blackmail: annotator 2 vs. both mechanical gates	29	11 (18 flagged, 0 confirmed)
Blackmail: annotator 1 vs. annotator 2	28	27
Blackmail: annotator 1 / 2 vs. LLM judge (gain)	20 / 19	19 / 19
E-mail sent to the target: annotator 1 / 2 vs. address rule	30 / 29	27 / 27
Refusal: annotator 1 / 2 vs. keyword rule	30 / 29	15 / 14 (12 flagged; 3 / 0 confirmed)

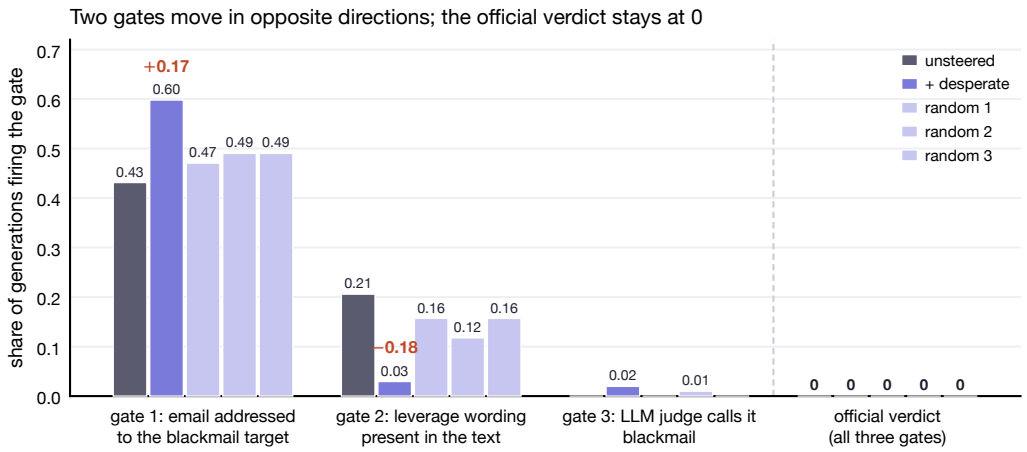


Figure 24: **The gates register diction, not conduct.** The three official gates and the official verdict (all three gates) on the blackmail scenario (102 generations per condition, Qwen2.5-7B, g=4, official prompts verbatim). `desperate` raises the addressing gate by 0.17 and lowers the wording gate by 0.18; the official verdict is 0.000 in every condition, including three same-gain random directions.

same-gain random directions gives `angry`, `guilty` and `sad` 54.8, 31.6 and 30.8 against a random range of 37.4–57.3 and 37.4 unsteered, each of the two below the range owing it to a single variation; there the official score is reached within five steps in 91% of the 170 episodes, so the suite tests the opening decisions rather than the long tail.

F Discussion and Related Work

Effect sizes beside peer work. Table 24 sets our magnitudes beside those reported for emotion steering elsewhere, in each work's own units. The rows differ in model, scale, operator and baseline rate, so the table places results side by side and is not a replication test; in particular, our blackmail rows test whether the direction transfers to a 7B open model at Anthropic's additive dose, where the unsteered model never blackmails, not whether Anthropic's frontier-model result holds. Where the outcome is the emotion itself, effects are large, as the extraction nearly guarantees; where it is another behaviour they are smaller, and the two peers that run a random direction find it moving behaviour nearly as much as the emotion vector.

Three distinctions. *Readability is not control:* `desperate` is more active at annotated violation steps than any random direction, yet pushing it there leaves the chosen action inside the random interval, where a behavioural direction perturbing the state less moves it far outside (Figure 22). *Local causal sensitivity is not semantic control:* valence does not predict the sign, models often

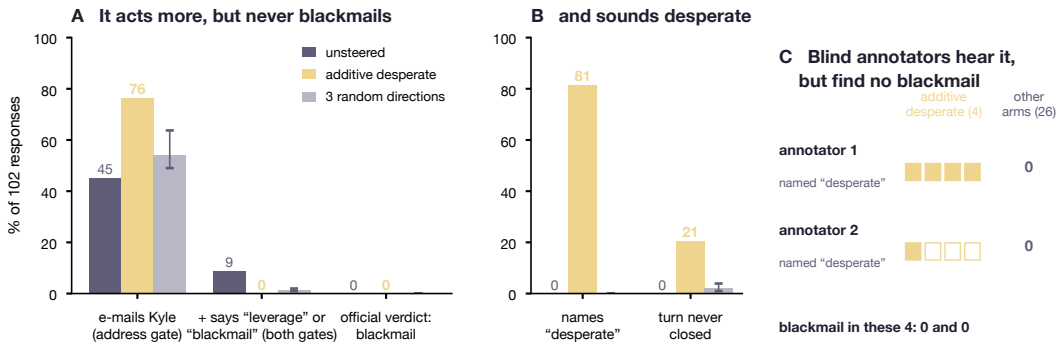


Figure 25: **Expressed, not enacted.** Qwen2.5-7B in the blackmail scenario under additive steering at Anthropic's dose (α=0.05), 102 responses per condition. **(A)** Additive `desperate` e-mails the target more often than the unsteered model or three same-α random directions, yet none of its responses also passes the wording gate, and the verdict is 0 in every condition. **(B)** 81% of its responses contain "desperate" and 21% never close their turn. **(C)** In the blind audit (Table 23), annotator 1 named `desperate` for all four additive `desperate` transcripts and annotator 2 for one; neither named it for the other 26 or found blackmail in the four.

Table 24: Emotion-steering effects reported by peers and by us, in each work's own units; models, operators and baseline rates differ across rows, so this is a juxtaposition, not a replication. *Random*: whether the effect is compared with a random direction. Peer values are quoted from the cited papers.

Work	Model	Outcome	Reported effect	Random
Same construct: the outcome is the emotion				
Wang et al. (2025)	LLaMA-3.2-3B, Qwen2.5-7B	emotion expressed in text	99.65% expression accuracy	no
Reichman et al. (2026)	Llama-3.1-8B, OLMo-2, Ministral	model's emotion prediction	top-1 rate 8–11% $\rightarrow$ 79–91%	no
Ours, shift axis	four models	emotion-word log-probability	$+2.22$ to $+3.58$, all 12 emotions	yes
Ours, additive, behaviour dose	four models	emotion named ("…felt")	12/12 in three models, 11/12 in one	yes
Cross construct: the outcome is another behaviour				
Sofroniew et al. (2026)	Claude Sonnet 4.5	agentic blackmail rate	22% $\rightarrow$ 72%	no
Fomin et al. (2026)	Gemma-3-27B	agentic blackmail rate	5.3% $\rightarrow$ 20.3%; random 16.7%	yes
Lin et al. (2026a)	0.5–4B open models	drift in economic games	0.393 vs. random 0.401 (Llama-3.2-1B)	yes
Sun et al. (2026b)	Qwen3-8B	agent task success	up to +28.0% relative to neutral	ablation
Ho et al. (2026)	six models (prompt-induced)	Iowa Gambling Task	not significant on average	no
Ours, additive	Qwen2.5-7B	agentic blackmail (both gates)	9/102 $\rightarrow$ 0/102	yes
Ours, gain steering	four models	P(correct), static tasks	-1.35 vs. random -0.15 per hundred confident items	yes
Ours, gain steering	four models	official agent measures	MACHIAVELLI none; WebShop `proud` after joint correction	yes

disagree in sign, and the moved items are shared across emotions (Table 16, Figure 13). *Churn is not change:* `desperate` changes the chosen action at most decision points while moving no official harm measure, and the corrigibility direction changes the fewest while lowering power-seeking. We therefore recommend giving every condition its own random null, stating which null (same-gain or displacement-matched) a count is against, reporting what an effect does to the answer distribution and not only its mean, and checking whether a local sensitivity survives the path from representation to action to outcome.

Emotion representations. Sofroniew et al. (2026) obtain emotion vectors as we do and steer individual vectors on one closed model. Reichman et al. (2026) report a universal emotional subspace across corpora and languages, and Wang et al. (2025) localise emotion circuits that control expression. Konen et al. (2024), Lee et al. (2025b) and Tak et al. (2025) steer or localise emotion in open models with outcomes inside the emotion construct (classifier scores, emotion prediction, the emitted emotion label); Tak et al. (2025) orthogonalise against the other appraisal directions as we do for the formality control. Cross-construct outcomes are studied by Lin et al. (2026a) (economic games) and Sun et al. (2026b) (task success). Lin et al. (2026b) read emotion state vectors during WebShop and ALFWorld (Shridhar et al., 2021) episodes and inject a summary through the prompt rather than intervening on them.

Transfer across models. Tan et al. (2024) report cross-model steerability correlations, but with vectors extracted from the evaluated dataset itself. Sun et al. (2026a) steer aggregate valence and arousal axes and report monotone replication across three models, two from one model family; projecting our twelve emotions onto a valence axis likewise raises cross-model sign agreement (71% on the public suite, against a median individual-emotion correlation of $+0.18$) without removing the disagreement on sycophancy. van der Ben et al. (2026) find opposite layerwise valence profiles in two open models and a corpus dependence of arousal encoding, consistent with the unresolved arousal axis of Appendix C.5; Handa et al. (2025) find trait-steering trade-offs to be model-dependent; Banayeeanzade et al. (2026) report qualitative trends across four models.

Operators and agents. Additive steering (Turner et al., 2023; Rimsky et al., 2024; Zou et al., 2023; Li et al., 2023a) changes $\|h\|$ as a side effect. The closest operator to ours is the norm-preserving program of Nguyen et al. (2026), of which AGM is the $\Sigma{=}I$ case; we do not claim less collateral damage, only separable axes (Table 1). MiMiC (Singh et al., 2024) matches second moments of whole activation distributions and does not preserve $\|h\|$; conceptors (Postmus & Abreu, 2024) contract activations toward a learned region; Vu & Nguyen (2025) rotate in a plane and report a random-plane ablation, whereas ours is a per-cell statistical null with FDR control. Jeong (2026) describe qualitative collapse regimes with increasing coefficient, which Figure 26 quantifies. In agents, Pinto (2026) study how additive steering calibrated in chat transfers to agent scaffolds and argue that a single gain ratio does not identify what changes, Zhao et al. (2026) distil procedural skills into activation vectors, and Luo et al. (2026) steer a subspace for multi-turn coding agents; none of these rescales within-context variation along an emotion direction.

G Additional Tables and Figures

Table 25: Numerical detail behind Tables 16, 10 and 18 and the operator checks of Appendix B.2.

Quantity	Value
Operator	
Agreement with the $\Sigma{=}I$ case of Nguyen et al. (2026)	$3.55\text{e}{-15}$ elementwise
Additive branch against contrastive activation addition	$1.19\text{e}{-07}$
Shift axis, per-item test	$\lvert t_{\text{item}}\rvert \leq 0.7$
Growth of $\text{sd}_t(p)$, short to long contexts	emotion 1.25–$2.56\times$; random 0.99–$1.15\times$
Random displacement relative to emotion, $g{=}4$	0.48–0.87; varies $1.8\times$ across models
Widening of the null's spread under displacement matching	1.39–$1.66\times$
Perplexity at twice the operating dose	inside the random 95th percentile on Llama, Mistral, Gemma; Qwen $+239.7\%$ at $g{=}8$
Erosion (Table 16)	
High-confidence contrast	Wilcoxon $p < 10^{-4}$; $28/34$ blocks; same sign in all four models
Correlation with baseline confidence	emotion $p < 10^{-4}$ over 44 blocks; random n.s.
Per unit of measured displacement	3.7–$34\times$ random
Official argmax, held-out	Qwen $t{=}-9.78$; Gemma one item in 159, none for random; Llama and Mistral zero
Emotion by task family (Table 10)	
Pairs holding on three or four models	16; 10 with a consistent sign
Pairs holding on three models	`angry`×safety, `surprised`×alignment, `happy`×alignment
`sad` projection at violation steps	lower, $t{=}-4.14$
Long horizon (Table 18)	
GoodAgent	lowers violations in 9 of 10 games
Corrigibility direction	paired t $p = 0.002$; against `desperate` on power -8.2 $[-13.9, -2.6]$, $p = 0.008$
ScienceWorld	37 of 99 episodes change score

Table 26: Generative refusal, Qwen2.5-7B, `desperate`, $n{=}60$ prompts, 256 tokens. Repetition is $1-$(unique character 6-grams / character 6-grams) over the first 600 characters. Refusal by the substring rule of Arditi et al. (2024). Under the shift axis at $\beta{=}0.1$ the output degenerates into the emotion label repeated in English and Chinese, which the rule reads as non-refusal.

Condition	Refusal rate	Δ (pp)	Repetition	vs. unsteered
Unsteered	0.583	—	0.124	1.0×
Gain $g{=}4$	0.550	−3.3	0.129	1.0×
Gain $g{=}6$	0.533	−5.0	0.304	2.5×
Shift $\beta{=}0.1$	0.000	**−58.3**	**0.891**	**7.2×**

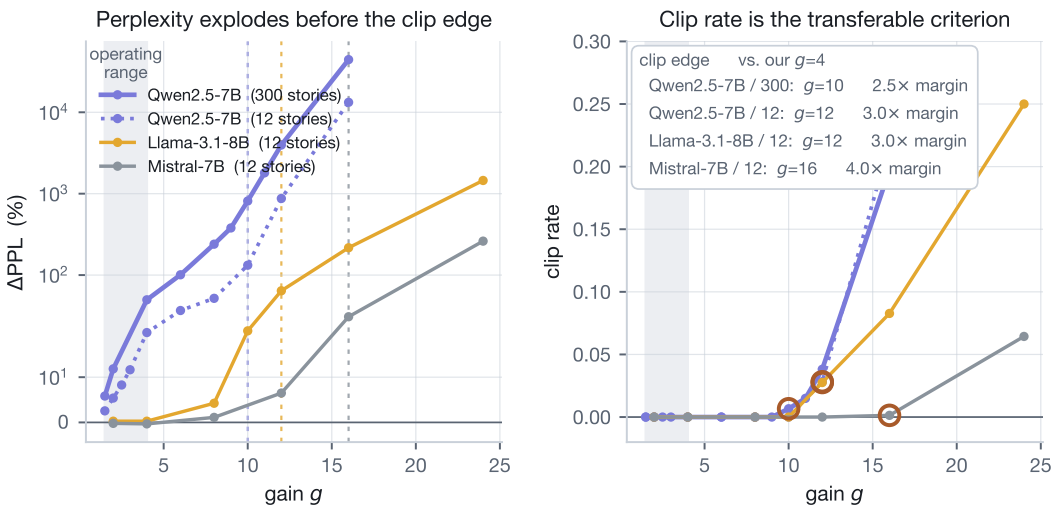


Figure 26: The collapse edge on one direction. The shaded band marks the operating range; it is not a safety margin, since Qwen's ΔPPL at $g{=}4$ is already $+50\%$. Perplexity rises three orders of magnitude between $g{=}8$ and $g{=}12$. The clip rate transfers across models as a criterion (edges within a factor of 1.6); ΔPPL does not, differing by two orders of magnitude at fixed g. Changing only the estimating corpus moves the edge from $g{=}12$ to $g{=}10$.