# WHAT DOES STEERING CHANGE? DECODING WITH BASE-RELATIVE STEERING EVIDENCE

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Activation steering controls a frozen language model by adding a direction to its hidden states, but the outcome hinges on the steering strength. Weak steering may leave the output largely unchanged, while stronger steering can degrade content. We study this sensitivity by examining how steering changes next-token predictions relative to the unsteered model. At a fixed strength, steering-induced divergence varies widely within generations, and steering can promote alternative tokens before changing the most probable continuation. Crucially, base-relative gain predicts which alternatives become dominant at stronger interventions beyond what steered probability alone predicts. These observations motivate decoding with base-relative steering evidence. We instantiate this approach in Multi-Alpha Consensus Decoding (MAC), which evaluates several strengths, filters unreliable predictions, and scores tokens by base-relative gain anchored to base likelihood. In controlled comparisons, base-relative gain scoring can produce stronger style transfer than mixing steered probabilities. Across two authorship-transfer benchmarks and four model backbones, MAC achieves the highest target-style similarity among the compared methods, with a measurable cost in meaning preservation. Our results suggest that steering-strength sensitivity is partly a decoding problem that base-relative steering evidence helps address.

## 1 INTRODUCTION

Activation steering controls a frozen language model by adding a direction to its hidden states at inference time (Subramani et al., 2022; Li et al., 2023a; Turner et al., 2024; Rimsky et al., 2024; Andy Zou, 2023). Its strength, denoted by $\alpha$, scales this intervention. Choosing the direction is only part of the problem, since generation must express the target behavior while preserving content. Weak steering may leave the output unchanged, while stronger steering can distort meaning or induce repetition. The challenge is deciding how much steering should influence each next-token decision.

Recent work helps explain why choosing steering strength is difficult. Taimeskhanov et al. (2026) analyze difference-of-means steering in an idealized linear-softmax next-token prediction model and show that the effect of steering strength is not generally monotonic. Most individual token probabilities follow a *bump-shaped* response, rising with steering strength before reaching a peak and then declining. Their experiments on traits and topics such as joy and mathematics observe similar non-monotonic responses in practical language models. Importantly, the peak location depends on both the token and context, motivating adaptive steering rather than a globally fixed steering strength.

Existing adaptive methods address strength sensitivity across inputs or generation positions (Scalena et al., 2024; Kang & Kim, 2026; Wang et al., 2025; Cheng et al., 2025; Yu et al., 2025). Hybrid Adaptive Search (Tran et al., 2026) selects a coefficient for each input through sequence-level evaluation, while MoLaCE (Kim & Torr, 2025) mixes token probabilities from several strengths using input-level weights that remain fixed throughout generation. This leaves open how the relative contributions of different strengths should adapt at each generation step as the context evolves.

To understand what should guide this adaptation, we examine how a fixed steering strength changes next-token predictions over generation. Our analysis across four backbones shows that the divergence induced by a fixed strength varies greatly within generations. High token probability alone also does not distinguish a token promoted by steering from one that was already favored by the base model. We find that steering can raise the probability of an alternative token by at least tenfold while leaving



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.

the most probable token unchanged. Greedy decoding then still selects the same token as the base model, despite the large increase in the probability of an alternative. Larger increases also predict which alternatives become the most probable at stronger interventions, beyond their current steered probability. These observations motivate examining the probability changes introduced by steering during next-token selection.

We therefore introduce *base-relative steering evidence*, which compares each strength-conditioned prediction with the unsteered prediction. Base-relative gains measure the additional support introduced by steering for candidate tokens, making steering-induced changes explicit rather than relying on absolute steered probabilities alone. We instantiate this principle in **M**ulti-**A**lpha **C**onsensus Decoding (MAC), a training-free decoder that evaluates the unsteered model alongside a fixed strength grid on a shared prefix. It uses current and recent base-relative steering evidence to adapt strength contributions and combines their token-level base-relative gains with base likelihood and repetition control to select the next token. The model and direction remain fixed while the strength contributions adapt as the prefix evolves. We then ask whether using this evidence during decoding improves behavioral control while preserving content.

Answering this question requires tasks where behavioral control and content preservation can be assessed together throughout a response. Authorship style transfer directly tests this setting. Matching a target style may require widespread changes across a response or only localized changes to particular words or phrases, while other parts already align with the target style. Steering must remain available throughout generation, but applying it at every token is unnecessary, since excessive steering can strengthen target-style expression at the cost of meaning preservation. We evaluate this trade-off on two authorship transfer benchmarks (MUD (Khan et al., 2021) and LaMP Tweet Paraphrasing (Salemi et al., 2024)). IFEval (Zhou et al., 2023) provides a complementary setting with verifiable constraints. This lets us test the same trade-off when a response must satisfy the target constraint while still completing the requested task coherently and with good writing quality.

Our contributions are:

- We introduce base-relative steering evidence, which measures steering-induced support relative to the unsteered model rather than relying on absolute probability. We instantiate this principle in MAC, a training-free decoder that uses this evidence to adapt strength contributions and guide token selection.
- We show what steering changes at the token level. Steering promotes alternative tokens before they become the top prediction, and base-relative gain captures this emerging preference better than steered probability alone. These findings motivate base-relative decoding.
- Across two authorship-transfer benchmarks and four model backbones, MAC achieves the highest target-style scores, with a measurable cost in meaning preservation. On IFEval, it achieves the highest observed compliance in both instruction conditions, with near-greedy quality under explicit instructions and a quality cost without explicit instructions.

## 2 Background and Related Work

**Activation steering.** Steering directions can be extracted by optimizing hidden-state offsets to reconstruct target sentences (Subramani et al., 2022), contrasting activations from paired prompts (Turner et al., 2024), or averaging residual-stream differences between positive and negative behavioral examples (Rimsky et al., 2024). Given a prompt $P$ and a generated prefix $y_{<t}$, an additive intervention produces a strength-conditioned next-token distribution,

$$h_{t,\ell} \leftarrow h_{t,\ell} + \alpha v_\ell, \qquad p_{t,\alpha}(x) = p_\theta(x \mid P, y_{<t}; \alpha v). \tag{1}$$

Here $h_{t,\ell}$ is the hidden state at layer $\ell$, $v_\ell$ is a steering vector, and $x$ is a candidate token. The model parameters $\theta$ remain fixed, and $\alpha = 0$ gives the unsteered distribution $p_{t,0}$. Increasing positive $\alpha$ scales the intervention magnitude. Fixed-strength greedy decoding selects $y_t = \arg\max_x p_{t,\alpha}(x)$.

**Adaptive activation steering.** Dynamic Activation Composition (DAC) (Scalena et al., 2024) uses KL divergence between truncated unsteered and steered next-token distributions to adjust the steering strength at each position. DIRECTER (Kang & Kim, 2026) adapts steering through dynamic rejection of predictions that lack sufficient base-model support. Hybrid Adaptive Search (Tran et al.,

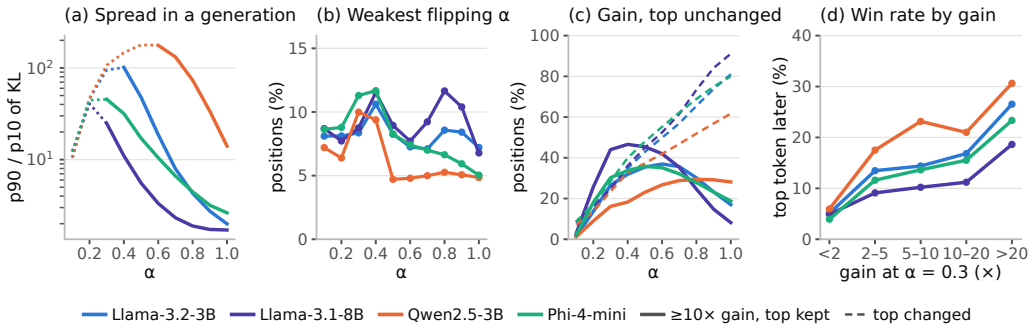



Figure 1: Steered versus unsteered next-token predictions on the same prefix. **(a)** Within-generation spread of $\mathrm{KL}(p_{t,\alpha} \,\|\, p_{t,0})$, shown as the median 90th/10th percentile ratio. **(b)** Weakest strength that changes the most probable token. **(c)** Positions with a tenfold-promoted token and an unchanged top token (solid) or a changed top token (dashed). **(d)** At $\alpha = 0.3$, the fraction of alternative tokens with $p_{t,\alpha} \geq 0.01$, excluding the unsteered top token, that become the top token at a higher strength on the same prefix, grouped by gain $p_{t,\alpha}/p_{t,0}$. Line styles refer to (c).

2026) instead selects a coefficient per input through sequence-level evaluation. Other approaches use calibrated controllers, trajectory signals, abstention, learned gates, or feedback (Wang et al., 2025; Cheng et al., 2025; Hedström et al., 2025; Yu et al., 2025; Ferrando et al., 2026; Hsu et al., 2026; Nguyen et al., 2026). Multi-attribute and memory-based methods also adapt intervention composition (Nguyen et al., 2025; Do et al., 2025). These approaches adapt which intervention is applied or how strongly it is applied. Our focus is on how changes by steering relative to the base model can guide the combination of predictions from multiple strengths for next-token selection.

MoLaCE (Kim & Torr, 2025) is the closest approach to ours. It retains an ensemble of strengths and forms an arithmetic mixture $\sum_{\alpha \in A} w(\alpha) p_{t,\alpha}(x)$, where $A$ is the strength grid and $w(\alpha)$ are the strength weights based on cosine similarity between the prompt representation and the steering direction. MAC also retains multiple strengths but differs in both its weighting signal, derived from current and recent base-relative steering evidence, and its token objective, which combines token-level base-relative gains with base likelihood. Even identical weights can yield different token choices (Appendix C).

**Likelihood-ratio guidance.** Classifier-free guidance for language models (Sanchez et al., 2024), contrastive decoding (Li et al., 2023b), DExperts (Liu et al., 2021), and Activation Steering Decoding (Su et al., 2025) guide generation with log-probability differences. MAC contributes an adaptive combination of positive gains across strengths rather than the ratio itself (Appendix C.1).

**Style transfer and instruction following.** DAC (Scalena et al., 2024) evaluates adaptive steering on behaviors including binary formality control. More broadly, style transfer aims to change stylistic properties while preserving meaning (Jin et al., 2022; Rao & Tetreault, 2018), with recent activation-steering work extending to exemplar-based and personalized generation (Horvitz et al., 2024; Zhang et al., 2025; Salemi et al., 2024; Tran et al., 2026). For instruction following, DIRECTER (Kang & Kim, 2026) studies adaptive steering, while Stolfo et al. (2025) show on IFEval (Zhou et al., 2023) that instruction-specific directions can improve constraint adherence with and without explicit instructions. We therefore ask whether decoding with base-relative steering evidence can further improve control on these tasks.

## 3 How Steering Changes Next-Token Predictions

We first examine how activation steering changes next-token predictions relative to the unsteered model. We construct an ablation set of 500 source–target authorship style-transfer pairs drawn from MUD (Khan et al., 2021) and generate prefixes with the unsteered model, yielding the base prediction $p_{t,0}$ at each position $t$. For the same prefix, we then evaluate the steered prediction $p_{t,\alpha}$ for $\alpha \in \{0.1, 0.2, \ldots, 1.0\}$. This yields 85,500 positions across four backbones. Section 5 describes

the models and experimental setup. Appendix A gives the full protocol and repeats the analysis on prefixes generated with fixed-strength steering.

**RQ1. Does a fixed strength have a consistent effect across positions?** We measure steering effect by $d_t(\alpha) = \mathrm{KL}(p_{t,\alpha} \,\|\, p_{t,0})$. At a fixed $\alpha$, divergence varies substantially across positions within the same generation (Figure 1a). The median within-generation 90th/10th percentile ratio is at least ten for $\alpha \in [0.1, 0.4]$ on every backbone, with peak ratios ranging from 41 to 178. Thus, fixed-strength steering has an inconsistent measured steering effect across generation positions.

**RQ2. Does the flip strength vary across positions?** Define the *flip strength* as the weakest $\alpha$ whose top token differs from the unsteered prediction. Each strength is the first flip for 5–12% of positions, while 9–37% never flip (Figure 1b). Counting no flip as a separate outcome, consecutive positions have different flip outcomes 72–84% of the time. Thus, the steering strength for changing the top prediction varies across positions, motivating adaptation throughout generation.

**RQ3. Can steering promote alternatives without changing the argmax?** We call a token tenfold-promoted if it ranks 2–10 under $p_{t,\alpha}$, has probability at least $0.01$, and gains at least $10\times$ over its unsteered probability. Figure 1c counts positions where such a token exists while the steered argmax remains unchanged compared to unsteered. At $\alpha = 0.3$, this occurs at 16–44% of positions, compared with 23–28% where the top token changes. Among positions with either event, 41–64% have an alternative token promoted while the top token remains unchanged. Thus, steering can increase the probability of alternative tokens without changing the unsteered top token.

**RQ4. Does gain predict later rank-1 transitions beyond current steered probability?** For each strength $\alpha \in \{0.1, \ldots, 0.9\}$, we consider tokens of steered rank 2–10 with $p_{t,\alpha} \geq 0.01$, excluding the unsteered top token. We group them by gain $g = p_{t,\alpha}/p_{t,0}$ and record whether they become the top token at any higher tested strength on the same prefix. On every backbone and starting strength, the highest-gain group reaches rank-1 more often than the group below twofold. Rates are 52–84% versus 8–10% at $\alpha = 0.1$, 19–31% versus 4–6% at $0.3$ (Figure 1d), and 3–8% versus 0–2% at $0.9$.

Because all comparisons follow the same steering trajectory, this later rank-1 transition could partly reflect the current steered probability. We therefore control for it. Within $p_{t,\alpha} \in [0.1, 0.3)$ at $\alpha = 0.3$, the transition rates are 50–56% for gains of at least twentyfold versus 12–15% below twofold. Adding log-gain to steered log-probability also raises the prediction AUC from 0.736–0.783 to 0.790–0.838 across backbones, and improves it in all 36 backbone–strength comparisons (Appendix A.1).

**What does steering change?** Steering can increase support for alternative tokens before changing the top prediction, and larger base-relative changes predict which alternatives later reach rank-1 under stronger interventions beyond what is indicated by current steered probability alone. Decoding based only on absolute steered probabilities can therefore miss steering-induced preferences. These findings motivate decoding with *base-relative steering evidence*, which we develop in Section 4.

## 4 Decoding with Base-Relative Steering Evidence

MAC uses this base-relative steering evidence for token selection and adaptive strength contributions, with base-model support and usability filtering to preserve content. Following Eq. 1, it evaluates in parallel the unsteered branch $p_{t,0}$ and a fixed grid $A = \{\alpha_1, \ldots, \alpha_m\}$ on a shared prefix. Figure 2 visualizes the decoder. The implementation details are in Appendix B.

### 4.1 Candidate Tokens and Steering Evidence

At position $t$, a strength is kept in the usable set $U_t \subseteq A$ when its distribution is not disrupted, its top token remains plausible under the base model, and that token is sufficiently confident. We measure distributional disruption by entropy, confidence by the top-token probability, and plausibility by its rank under the base model. Appendix H ablates these criteria and reports retention rates.

We form the candidate set $C_t$ from the top tokens of the base branch and each usable steered branch. For each candidate token $x$, we measure the positive steering-induced change in log probability

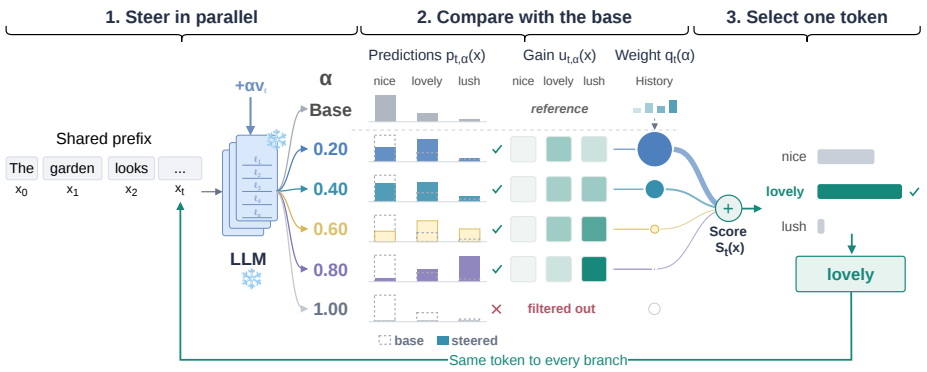



Figure 2: Multi-Alpha Consensus Decoding (MAC). (1) A frozen model evaluates the unsteered branch and multiple steering strengths on a shared prefix. (2) MAC filters unusable strengths and uses current and recent base-relative gains to weight the remaining strengths. (3) A score combining base likelihood, weighted gain, and repetition control guides token selection. The selected token updates every branch.

relative to the base model:

$$u_{t,\alpha}(x) = \begin{cases} \left[\log \dfrac{p_{t,\alpha}(x)}{p_{t,0}(x)}\right]_+, & \alpha \in U_t, \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

Here $[z]_+ = \max\{z, 0\}$, where $z = \log \frac{p_{t,\alpha}(x)}{p_{t,0}(x)}$. Every usable strength contributes a gain for every token in $C_t$. The log ratio measures multiplicative promotion relative to the base prediction, so the same probability ratio gives the same gain regardless of the initial probability. Positive clipping gives a steering bonus only to promoted tokens while a decrease contributes zero bonus. Because steered probabilities are normalized, promoting some tokens necessarily lowers others, so a decrease need not indicate evidence against a token. Clipping therefore lets steering reward the continuations it promotes without penalizing base-supported ones, which the base likelihood term already governs. In our ablation, clipping improves meaning preservation relative to the unclipped log ratio (Appendix H).

### 4.2 Adaptive Strength Weights

To determine how much each usable strength contributes, we aggregate the gains of its candidate tokens and retain a smoothed evidence state. Usability failures are rare at weak strengths but rise starting from $\alpha = 0.3$ (Appendix A). We therefore favor weaker interventions and center the prior at $\alpha = 0.2$, where the degenerate steered prediction rate is 0%. The center is a fixed design choice. The strength grid is the only MAC decoder setting tuned on the development set (Appendix B.1).

For each strength $\alpha$, we obtain its current strength-level evidence by averaging its positive base-relative token gains over the candidate set:

$$g_t(\alpha) = \sum_{x \in C_t} \tilde{p}_{t,\alpha}(x) u_{t,\alpha}(x), \qquad \tilde{p}_{t,\alpha}(x) = \frac{p_{t,\alpha}(x)}{\sum_{z \in C_t} p_{t,\alpha}(z)}. \tag{3}$$

Here $\tilde{p}_{t,\alpha}$ renormalizes the steered probabilities over $C_t$. This expectation gives more weight to gains on candidates that the branch considers likely, rather than treating candidate gains equally.

To reduce sensitivity to a single scoring step, we combine the current and previous evidence states:

$$s_t(\alpha) = \frac{g_t(\alpha) + s_{\text{prev}}(\alpha)}{2}. \tag{4}$$

At the first evidence update, we initialize $s_t = g_t$. Here $s_{\text{prev}}$ denotes the smoothed evidence state from the most recent scoring step in the decoder instance. Direct consensus decisions and fallback decisions made when no strength is usable leave the state unchanged.

We combine this evidence with a Gaussian prior $\rho(\alpha)$:

$$q_t(\alpha) = \frac{\mathbf{1}[\alpha \in U_t]\rho(\alpha)\exp(s_t(\alpha))}{\sum_{a \in U_t} \rho(a)\exp(s_t(a))}. \tag{5}$$

A strength excluded at one position can contribute again if it later becomes usable. The prior specification and numerical settings are given in Appendix B.

### 4.3 Consensus Token Selection

MAC combines the expected base-relative gain across steering strengths with base likelihood and a local repetition cost:

$$S_t(x) = \log p_{t,0}(x) + \lambda \sum_{\alpha \in U_t} q_t(\alpha) u_{t,\alpha}(x) - \gamma c_t(x), \qquad y_t = \arg\max_{x \in C_t} S_t(x). \tag{6}$$

Here $c_t(x)$ counts recent occurrences of $x$ among the generated tokens (Appendix B). The base term prevents relative gain alone from determining the score, since a large ratio can arise for a token with low absolute probability. The repetition term discourages a promoted token from repeatedly winning the score and creating a local loop. The coefficients $\lambda$ and $\gamma$ are fixed across experiments. Appendix H compares token objectives and repetition control.

Before computing the weighted score, MAC checks whether all usable strengths agree on the same preferred continuation. This direct consensus reflects the hypothesis that tokens already aligned with the target behavior should remain stable across steering strengths. If the consensus token is also plausible under the base model and passes the repetition check, the decoder accepts it directly. Otherwise, the strength weights are updated from the current evidence and Eq. 6 resolves the disagreement. If no strength is usable, the decoder falls back to the base-model argmax. Every selected token is then appended to the shared prefix and used by all branches at the next position.

## 5 Experimental Setup

**Authorship style transfer.** We evaluate on MUD (Khan et al., 2021) and Tweet Paraphrasing from LaMP (Salemi et al., 2024) using Llama-3.2-3B and Llama-3.1-8B (Llama, 2024). For generalization, we additionally evaluate Qwen2.5-3B (Qwen, 2025) and Phi-4-mini (Microsoft, 2025) on MUD. We measure target-style similarity with StyleCAV (Wegmann et al., 2022) and meaning preservation against the source with SBERT (Reimers & Gurevych, 2019); higher values are better for both. We compare methods by their style–meaning trade-off. Additional style metrics, including StyleDistance (Patel et al., 2025), are reported in Appendix E. We also compare outputs using GPT-4.1 with the pairwise preference prompt of Tran et al. (2026). Appendix D gives the full evaluation protocol.

**Instruction following.** We evaluate the Format and Keywords categories of IFEval (Zhou et al., 2023) using Llama-3.2-3B. We report strict instruction-following accuracy with and without explicit instructions. We additionally assess task fidelity and text quality using the GPT-4.1 judge rubric of Kang & Kim (2026). Appendix G gives the full evaluation protocol.

**Baselines.** We compare MAC against unsteered decoding, fixed-strength steering, and three training-free methods that adapt strength at different levels, namely Hybrid Search (per-input coefficient search), DAC (per-token KL-rescaling of one strength), and MoLaCE (input-level mixture of several strengths, the closest to us). For authorship transfer, the unsteered baseline rewrites the source without target-style text. For IFEval, it uses the corresponding task prompt. Fixed-strength steering applies a single coefficient $\alpha$ throughout generation. We sweep $\alpha \in \{0.1, 0.2, \ldots, 1.0\}$ and report the strength with the best style–meaning trade-off. Appendix F reports the full sweep. DIRECTER is evaluated only on IFEval with explicit instructions because it requires an instruction span. Steering methods share prompts and direction construction within each task, following Tran et al. (2026) for authorship transfer and Stolfo et al. (2025) for IFEval. Appendices D and G give details.

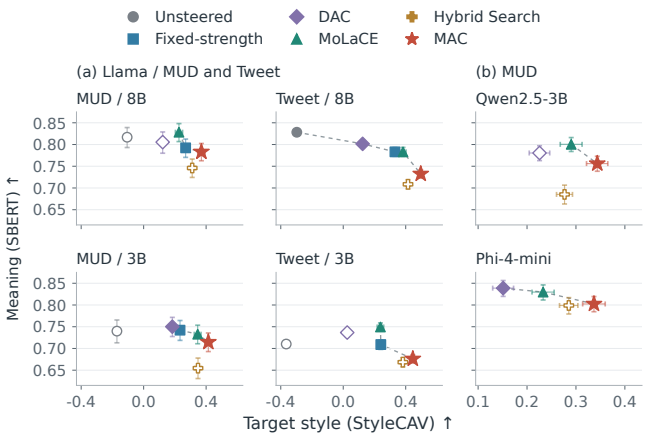



Figure 3: Target-style similarity and meaning preservation. **(a)** MUD and Tweet with Llama-3.1-8B (top) and Llama-3.2-3B (bottom). **(b)** MUD with Qwen2.5-3B (top) and Phi-4-mini (bottom). Error bars are 95% bootstrap intervals, and filled markers with dashed lines mark nondominated by means (Appendix E).

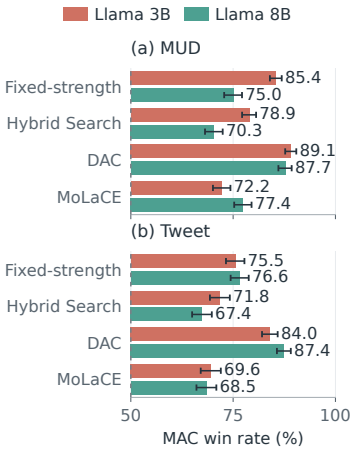



Figure 4: GPT-4.1 preferences. MAC win rates $W/(W+L)$ exclude ties and invalid verdicts. Error bars are 95% bootstrap intervals (Appendix D).

# 6 Results

## 6.1 Style-Transfer Results

**Automatic metrics.** Across both tasks and model sizes, MAC achieves the highest StyleCAV among the reported configurations (Figure 3(a); full metrics in Appendix E). On MUD, its StyleCAV is 0.370 on Llama-3.1-8B and 0.414 on Llama-3.2-3B, compared with 0.226 and 0.346 for MoLaCE. On Tweet, it reaches 0.496 and 0.446, compared with 0.383 and 0.238. The StyleCAV gains over every baseline are significant in all four settings (Holm-adjusted $p < 0.001$).

MAC lies on the style–meaning Pareto frontier in every setting. Relative to Hybrid Search, it improves both StyleCAV and SBERT across all four settings, with the SBERT gains significant except on Tweet with Llama-3.2-3B (Appendix E). MoLaCE retains higher SBERT, reflecting a different operating point with weaker target-style transfer but stronger source-meaning preservation.

**Pairwise style judgments.** GPT-4.1 prefers MAC in all baseline comparisons, with tie-excluded win rates of 70.3–89.1% on MUD and 67.4–87.4% on Tweet (Figure 4). All comparisons remain significant after Holm adjustment ($p < 0.001$), supporting the embedding-based style gains.

**Other model families.** We further compare MAC with DAC, MoLaCE, and Hybrid Search on MUD using Qwen2.5-3B and Phi-4-mini (Figure 3(b)). MAC again achieves the highest StyleCAV, at 0.344 on Qwen and 0.337 on Phi-4-mini versus 0.290 and 0.233 for MoLaCE. The StyleCAV differences from all baselines are significant after Holm adjustment ($p < 0.001$). Its target-style advantage therefore extends beyond the Llama family.

The style–meaning trade-off follows a similar pattern. Against Hybrid Search, MAC improves both style and SBERT on Qwen, with SBERT increasing from 0.685 to 0.756. On Phi-4-mini, stronger style transfer is accompanied by similar SBERT (0.802 vs. 0.799). MoLaCE retains higher SBERT on both backbones (0.800 vs. 0.756 on Qwen and 0.830 vs. 0.802 on Phi-4-mini).

## 6.2 Instruction-Following Results

**With explicit instructions.** MAC achieves the highest observed compliance (Figure 5, Appendix G), reaching 92.0% accuracy and improving by 3.6 percentage points over unsteered decoding. However, differences from several baselines are small, and no paired comparison remains significant

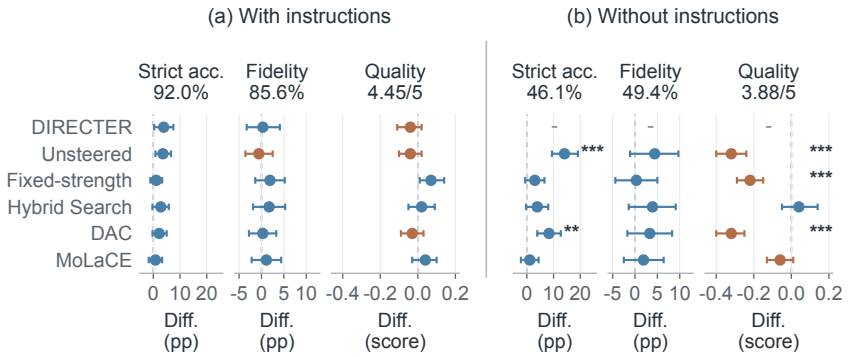



Figure 5: IFEval on Llama-3.2-3B. MAC minus each baseline (pp: percentage points) with unadjusted 95% paired bootstrap intervals. Stars mark Holm-adjusted tests ($^{**}p < 0.01$, $^{***}p < 0.001$).

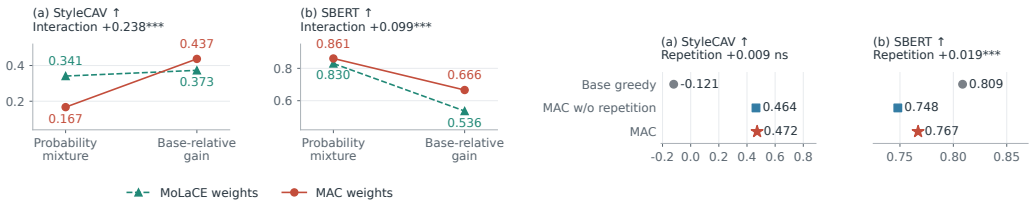



Figure 6: MUD ablations with Llama-3.2-3B. **Left:** Strength weighting and token objective. **Right:** Progression from greedy decoding to MAC. $^{***}p < 0.001$ after Holm adjustment and ns denotes no significant difference. Appendix H gives configurations and paired tests.

after Holm correction. Task fidelity and text quality are also similar to unsteered decoding, at 85.6% and 4.45 for MAC versus 86.2% and 4.49 for the unsteered baseline.

**Without explicit instructions.** MAC reaches 46.1% accuracy, exceeding unsteered decoding by 14.1 percentage points. The gains over unsteered decoding and DAC remain significant after Holm correction. MAC also reaches 49.4% task fidelity, but text quality falls to 3.88, significantly below unsteered decoding and DAC (both 4.20) and fixed-strength steering (4.10).

Together, these results suggest that decoding with base-relative steering evidence generalizes beyond style transfer. It can strengthen the steering signal and obtain stronger compliance, but at a measurable quality cost, highlighting the trade-off between behavioral control and generation quality.

## 7 Analysis

We test how base-relative steering evidence affects generation through strength weighting, token scoring, and decision rules on 500 MUD source-target samples with Llama-3.2-3B (Appendix H).

**Weighting and token selection interact.** Figure 6 (left) compares MoLaCE and MAC strength weights under probability mixing and base-relative gain scoring. Under gain scoring, replacing MoLaCE weights with MAC weights increases StyleCAV from 0.373 to 0.437 and SBERT from 0.536 to 0.666. Under probability mixing, the same replacement increases SBERT but reduces StyleCAV from 0.341 to 0.167. The weighting benefit therefore depends on the token objective. Relative to the MoLaCE-weighted probability mixture, the MAC-weighted gain configuration also has higher StyleCAV (0.437 vs. 0.341) and lower SBERT (0.666 vs. 0.830). Both interactions are significant after Holm adjustment ($p < 0.001$; Appendix H).

**Progression to the complete decoder.** Figure 6 (right) compares base greedy decoding with MAC before and after repetition control. The first transition introduces the decision procedure of Section 4

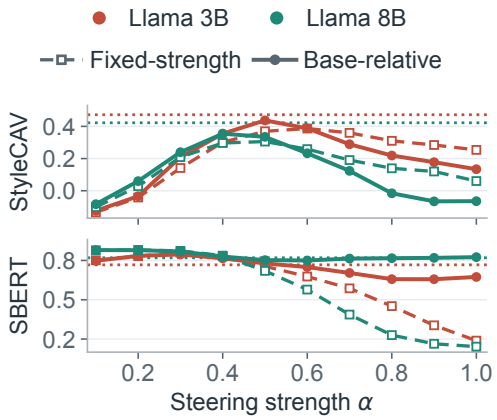



Figure 7: MUD single-strength sweep. Dotted lines mark MAC with its full grid.

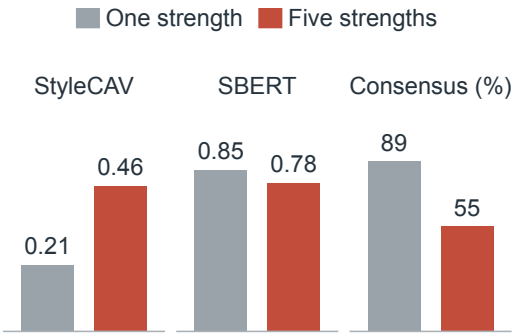



Figure 8: Scoring with $\alpha = 0.3$ when one or five strengths enter direct consensus (Llama-3.2-3B).

as a whole, and the second isolates repetition control. Relative to base greedy decoding, full MAC increases StyleCAV from $-0.121$ to 0.472 while reducing SBERT from 0.809 to 0.767. Adding repetition control recovers SBERT from 0.748 to 0.767 (Holm-adjusted $p < 0.001$).

**Single-strength decoding.** Base-relative decoding can improve the style–meaning balance even with one strength. We restrict MAC to a single $\alpha$ while retaining its score and decision rules, and sweep $\alpha \in 0.1, \ldots, 1.0$ against fixed-strength steering (Figure 7, Appendix I). At $\alpha = 0.5$, StyleCAV rises from 0.369 to 0.437 and SBERT from 0.755 to 0.778 on Llama-3.2-3B; on Llama-3.1-8B, the corresponding gains are 0.306 to 0.335 and 0.720 to 0.803. Using the full grid further increases StyleCAV by 0.035 and 0.068 at SBERT costs of 0.011 and 0.010, respectively (Table 29). Thus, base-relative decoding improves on fixed-strength steering, while multiple strengths shift the balance further toward target style.

**Where several strengths act.** Multiple strengths also affect when gain scoring is used. At $\alpha = 0.3$, using that strength for scoring while all five grid strengths participate in consensus yields StyleCAV 0.457 and SBERT 0.777, versus 0.209 and 0.846 with that strength alone (Figure 8). Consensus selects 89% of tokens with one strength but only 55% with the grid, making it more selective and leaving more positions to gain scoring. Disabling consensus in the five-strength setting lowers SBERT from 0.777 to 0.700 with little StyleCAV change (Appendix I). These results show that the grid contributes through the interaction between consensus and gain scoring.

## 8 Conclusion

Our analysis shows that steering-induced support often emerges before the argmax changes, revealing information about steering behavior that is not captured by the current prediction alone. Larger base-relative gains also predict which alternatives become the argmax at stronger steering strengths, beyond current steered probability. This observation led us to measure steering relative to the unsteered model, using log-probability gains to identify which tokens are actually promoted by steering. MAC uses these gains both to determine how much each strength contributes and to select the next token while retaining the base model likelihood and a repetition penalty.

On MUD and Tweet, MAC produces stronger target-style transfer, though the improvement comes with a measurable loss in meaning preservation. The IFEval results show a similar trade-off in a different setting. With explicit instructions, MAC keeps task fidelity and text quality close to unsteered decoding while achieving a small increase in compliance, although this difference is not statistically significant. Without explicit instructions, steering produces larger compliance gains at the cost of lower text quality. Overall, changes relative to the base model provide useful decoding information that absolute probability alone does not, but using this information does not eliminate the trade-off between stronger control and content preservation.

## 9 LIMITATIONS

**Scope of the analysis.** Our analysis identifies recurring patterns in how activation steering changes next-token distributions, including the promotion of alternative tokens before they become the most probable continuation. These patterns appear consistently across the evaluated backbones and settings, but we study only a subset of steering methods and tasks. Future work should examine whether similar dynamics arise for other forms of activation steering, different behavioral targets, and a broader range of language models.

**Scope of the decoder.** MAC is one instantiation of decoding with base-relative steering evidence. It combines usability filtering, a strength prior, gain clipping, consensus decisions, and repetition control. Although the ablations indicate that several of these components contribute to the final style–meaning trade-off, they do not establish that this formulation of gain, filtering, weighting, and token selection is optimal. Base-relative gain also measures steering-induced promotion rather than target-behavior success, so it does not verify that content is preserved. Alternative definitions of steering evidence or mechanisms for combining strengths may yield similar or better results.

**Computational cost.** MAC evaluates the unsteered branch and several steered branches at each decoding step. With five strengths batched in one model call, it takes 35.0 ms per generated token with Llama-3.2-3B on an RTX A6000 ($1.42\times$ greedy decoding) and uses 0.28 GB more peak allocated memory (Appendix K). It adapts only within a selected strength grid and a fixed direction, so strength-grid selection remains necessary.

### REPRODUCIBILITY STATEMENT

Section 4 defines the decoder, and Appendix B specifies its fixed constants and strength-grid search. Appendix A defines the token-level measurements. Appendices D–G document the steering-direction construction, intervention layers, dataset construction, full results, judge prompts, and statistical protocols. Appendices H–K report ablation configurations, single-strength comparisons, strength-grid sensitivity, and the runtime protocol. Code will be released upon acceptance.

### AI USE STATEMENT

We used generative AI tools to review and provide feedback on research methodology and experiments, assist with drafting and polishing the manuscript, search and summarize related literature, create visualizations, and assist with mathematical notation and implementation details based on ideas, hypotheses, and assumptions developed by the authors. The core experimental design and broader code implementation were developed by the authors. All AI-assisted materials were reviewed, verified, edited, and revised by the authors, who take full responsibility for the final content.

## REFERENCES

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

Zifeng Cheng, Jinwei Gan, Zhiwei Jiang, Cong Wang, Yafeng Yin, Xiang Luo, Yuchen Fu, and Qing Gu. Steering when necessary: Flexible steering large language models with backtracking. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=l75RyRcevf.

Van Dai Do, Quan Hung Tran, Svetha Venkatesh, and Hung Le. Dynamic steering with episodic memory for large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 13731–13749, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.706. URL https://aclanthology.org/2025.findings-acl.706/.

Guodong DU, Junlin Lee, Jing Li, Runhua Jiang, Yifei Guo, Shuyang Yu, Hanting Liu, Sim Kuan Goh, Ho-Kin Tang, Daojing He, and Min Zhang. Parameter competition balancing for model merging. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=l5SbrtvSRS.

Alex Ferrando, Xavier Suau, Jordi Gonzàlez, and Pau Rodriguez. Dynamically scaled activation steering, 2026. URL https://arxiv.org/abs/2512.03661.

Anna Hedström, Salim I. Amoukou, Tom Bewley, Saumitra Mishra, and Manuela Veloso. To steer or not to steer? mechanistic error reduction with abstention for language models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=fUCPq5RvmH.

Zachary Horvitz, Ajay Patel, Kanishk Singh, Chris Callison-Burch, Kathleen McKeown, and Zhou Yu. TinyStyler: Efficient few-shot text style transfer with authorship embeddings. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 13376–13390, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.781. URL https://aclanthology.org/2024.findings-emnlp.781/.

Brandon Hsu, Daniel Beaglehole, Adityanarayanan Radhakrishnan, and Mikhail Belkin. Contextual linear activation steering of language models, 2026. URL https://arxiv.org/abs/2604.24693.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205, March 2022. doi: 10.1162/coli_a_00426. URL https://aclanthology.org/2022.cl-1.6/.

Minjae Kang and Jaehyung Kim. Enhancing instruction following of LLMs via activation steering with dynamic rejection. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=OpuPBNcQwe.

Aleem Khan, Elizabeth Fleming, Noah Schofield, Marcus Bishop, and Nicholas Andrews. A deep metric learning approach to account linking. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5275–5287, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.415. URL https://aclanthology.org/2021.naacl-main.415/.

Hazel Kim and Philip Torr. Single llm debate, molace: Mixture of latent concept experts against confirmation bias, 2025. URL https://arxiv.org/abs/2512.23518.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023a. URL https://openreview.net/forum?id=aLLuYpn83y.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12286–12312. Association for Computational Linguistics, 2023b.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. DExperts: Decoding-time controlled text generation with experts and anti-experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6691–6706. Association for Computational Linguistics, 2021.

Team Llama. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Team Microsoft. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras, 2025. URL https://arxiv.org/abs/2503.01743.

Dung Viet Nguyen, Yen Nhi Pham, Hieu M. Vu, Lei Zhang, and Tan Minh Nguyen. Activation steering with a feedback controller. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=vzkEX2SwFD.

Duy Nguyen, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. Multi-attribute steering of language models via targeted intervention. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 20619–20634, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1007. URL https://aclanthology.org/2025.acl-long.1007/.

Ajay Patel, Jiacheng Zhu, Justin Qiu, Zachary Horvitz, Marianna Apidianaki, Kathleen McKeown, and Chris Callison-Burch. StyleDistance: Stronger content-independent style embeddings with synthetic parallel examples. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8662–8685, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.436. URL https://aclanthology.org/2025.naacl-long.436/.

Team Qwen. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Sudha Rao and Joel Tetreault. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 129–140, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1012. URL https://aclanthology.org/N18-1012/.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. LaMP: When large language models meet personalization. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7370–7392, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.399. URL https://aclanthology.org/2024.acl-long.399/.

Guillaume Sanchez, Alexander Spangher, Honglu Fan, Elad Levi, Pawan Sasanka Ammanamanchi, and Stella Biderman. Stay on topic with classifier-free guidance, 2024. URL https://openreview.net/forum?id=RmRA7Q0lwQ.

Daniel Scalena, Gabriele Sarti, and Malvina Nissim. Multi-property steering of large language models with dynamic activation composition. In Yonatan Belinkov, Najoung Kim, Jaap Jumelet, Hosein Mohebbi, Aaron Mueller, and Hanjie Chen (eds.), *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 577–603, Miami, Florida, US, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.blackboxnlp-1.34. URL https://aclanthology.org/2024.blackboxnlp-1.34/.

Alessandro Stolfo, Vidhisha Balachandran, Safoora Yousefi, Eric Horvitz, and Besmira Nushi. Improving instruction-following in language models through activation steering. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=wozhdnRCtw.

Jingran Su, Jingfan Chen, Hongxin Li, Yuntao Chen, Li Qing, and Zhaoxiang Zhang. Activation steering decoding: Mitigating hallucination in large vision-language models through bidirectional hidden state intervention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12964–12974. Association for Computational Linguistics, 2025.

Nishant Subramani, Nivedita Suresh, and Matthew Peters. Extracting latent steering vectors from pretrained language models. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 566–581, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.48. URL https://aclanthology.org/2022.findings-acl.48/.

Magamed Taimeskhanov, Samuel Vaiter, and Damien Garreau. Towards understanding steering strength. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=A52LNgpGKy.

Hieu Tran, Calvin Bao, and Marine Carpuat. Can activation steering capture multidimensional authorship style?, 2026. URL https://arxiv.org/abs/2609.04792.

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

Tianlong Wang, Xianfeng Jiao, Yinghao Zhu, Zhongzhi Chen, Yifan He, Xu Chu, Junyi Gao, Yasha Wang, and Liantao Ma. Adaptive activation steering: A tuning-free LLM truthfulness improvement method for diverse hallucinations categories. In *THE WEB CONFERENCE 2025*, 2025. URL https://openreview.net/forum?id=NBHOdQJ1VE.

Anna Wegmann, Marijn Schraagen, and Dong Nguyen. Same author or just same topic? towards content-independent style representations. In Spandana Gella, He He, Bodhisattwa Prasad Majumder, Burcu Can, Eleonora Giunchiglia, Samuel Cahyawijaya, Sewon Min, Maximilian Mozes, Xiang Lorraine Li, Isabelle Augenstein, Anna Rogers, Kyunghyun Cho, Edward Grefenstette, Laura Rimell, and Chris Dyer (eds.), *Proceedings of the 7th Workshop on Representation Learning for NLP*, pp. 249–268, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.repl4nlp-1.26. URL https://aclanthology.org/2022.repl4nlp-1.26/.

Manjiang Yu, Hongji Li, Priyanka Singh, Xue Li, Di Wang, and Lijie Hu. Pixel: Adaptive steering via position-wise injection with exact estimated levels under subspace calibration, 2025. URL https://arxiv.org/abs/2510.10205.

Jinghao Zhang, Yuting Liu, Wenjie Wang, Qiang Liu, Shu Wu, Liang Wang, and Tat-Seng Chua. Personalized text generation with contrastive activation steering. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7128–7141, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.353. URL https://aclanthology.org/2025.acl-long.353/.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023. URL https://arxiv.org/abs/2311.07911.

# A TOKEN-LEVEL ANALYSIS OF STEERING

The analysis uses the MUD ablation set of Appendix H, covering 34 source texts and 15 target authors. Each backbone uses the steering directions, intervention layers, and prompt of the main experiments. At every position, we evaluate the same prefix without steering ($\alpha = 0$) and at each $\alpha \in \{0.1, 0.2, \ldots, 1.0\}$, applying steering to the prompt and prefix as in Eq. 1. We compare the resulting full-vocabulary distributions $p_{t,0}$ and $p_{t,\alpha}$.

We consider two prefix policies. *Unsteered prefixes* extend the sequence with the unsteered argmax, so the target directions associated with a source are evaluated on the same generated text. *Fixed-strength (plain-steering) prefixes* extend it with the argmax at one fixed steering strength, corresponding to fixed-strength greedy decoding. Table 1 summarizes these generations. Inputs shorter than six words are copied unchanged and contribute no positions. A generation denotes one source–target pair. Confidence intervals resample source texts.

The divergence is $d_t(\alpha) = \mathrm{KL}(p_{t,\alpha} \,\|\, p_{t,0})$, measured in nats over the full vocabulary. Before taking logarithms or ratios, we floor values at $10^{-2}$ nats to limit ratios driven by near-zero divergence. For token-level statistics, we record the union of the top ten tokens from the unsteered branch and every steered branch, together with each token's log-probability and full-vocabulary rank under every branch. A token is *tenfold-promoted* at strength $\alpha$ if it ranks 2–10 under $p_{t,\alpha}$, has $p_{t,\alpha} \geq 0.01$, and satisfies $p_{t,\alpha}/p_{t,0} \geq 10$. A steered prediction fails the usable-set conditions if its entropy increase exceeds 6 nats, its peak probability is below $0.05$, or its most probable token has base rank above $1000$ (Eq. 8). Figures and tables label this event *degenerate*. It is a distributional diagnostic, not a judgment of generated-text quality.

**Per-strength statistics.** Table 2 reports divergence pooled over positions, the median within-generation 90th/10th percentile ratio, and the within-generation share of log-divergence variance. For $z_{i,t} = \log \max\{d_{i,t}, 10^{-2}\}$, this share is $\mathbb{E}_i[\mathrm{Var}_t z_{i,t}]/(\mathbb{E}_i[\mathrm{Var}_t z_{i,t}] + \mathrm{Var}_i[\mathbb{E}_t z_{i,t}])$, with equal weight per generation. It ranges from 77 to 90% at $\alpha \leq 0.2$ and from 60 to 81% at $\alpha = 0.3$, but falls to 25% for Llama-3.1-8B at $\alpha = 1.0$ (Figure 9, top). This measures variation at a fixed strength. Token-event frequencies pool positions. Unlike Figure 1c, the tenfold-promotion rows in Tables 2 and 5 include positions whose most probable token changes. Figure 11 shows the pooled divergence distribution and these broader promotion frequencies.

**Position-level statistics.** Table 3 reports, for both prefix policies and for the full ablation set and its 101-pair subset, flip outcomes, promoted-token peak strengths, and takeover at stronger strengths. Adjacent-position flip comparisons include no flip as a separate outcome. Peak comparisons use transitions where both positions have a selected promoted token. Takeover requires a stronger tested strength and therefore excludes tokens first promoted at the largest grid value. Controls share the rank range 2–10 but are not matched to promoted tokens by rank or probability. The bottom row of Figure 9 shows the flip- and peak-strength distributions, and Table 4 gives three representative positions.

Table 1: The text on which the analysis runs. Unsteered prefixes are greedy generations of the unsteered model, and plain-steering prefixes are greedy generations of the model steered at one fixed strength.

| Backbone | Prefix | generations | positions | mean tokens | StyleCAV | SBERT |
|---|---|---|---|---|---|---|
| Llama-3.2-3B | unsteered | 500 | 23,295 | 49 | -0.202 | 0.788 |
| | plain steering ($\alpha = 0.4$) | 500 | 20,469 | 43 | 0.257 | 0.809 |
| Llama-3.1-8B | unsteered | 500 | 22,620 | 48 | -0.177 | 0.865 |
| | plain steering ($\alpha = 0.4$) | 500 | 23,107 | 49 | 0.286 | 0.824 |
| Qwen2.5-3B | unsteered | 500 | 20,406 | 43 | -0.123 | 0.900 |
| | plain steering ($\alpha = 0.4$) | 500 | 28,217 | 60 | 0.229 | 0.859 |
| Phi-4-mini | unsteered | 500 | 19,179 | 41 | -0.149 | 0.871 |
| | plain steering ($\alpha = 0.3$) | 500 | 30,719 | 65 | 0.251 | 0.873 |

Table 2: Per-strength statistics on unsteered prefixes (MUD ablation set). Divergence is $\mathrm{KL}(p_{t,\alpha} \,\|\, p_{t,0})$, and values below $10^{-2}$ nats are set to $10^{-2}$. A steered prediction is counted as degenerate if its entropy exceeds the unsteered entropy by more than 6 nats, its peak probability is below 0.05, or its top token has unsteered rank above 1000.

| $\alpha$ | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1.0 |
|---|---|---|---|---|---|---|---|---|---|---|
| *Median divergence (nats)* | | | | | | | | | | |
| Llama-3.2-3B | 0.02 | 0.07 | 0.24 | 0.61 | 1.35 | 2.64 | 4.89 | 7.91 | 10.46 | 12.03 |
| Llama-3.1-8B | 0.02 | 0.14 | 0.52 | 1.33 | 2.95 | 5.82 | 9.74 | 13.17 | 15.13 | 16.27 |
| Qwen2.5-3B | 0.01 | 0.03 | 0.09 | 0.26 | 0.60 | 1.31 | 2.55 | 4.43 | 7.18 | 10.13 |
| Phi-4-mini | 0.02 | 0.10 | 0.38 | 1.05 | 2.34 | 4.33 | 6.92 | 9.60 | 11.93 | 13.83 |
| *p90 / p10 of the divergence inside one generation* | | | | | | | | | | |
| Llama-3.2-3B | 10 | 46 | 95 | 101 | 48 | 19 | 7.9 | 4.3 | 2.7 | 2.0 |
| Llama-3.1-8B | 12 | 41 | 25 | 11 | 5.6 | 3.3 | 2.3 | 1.9 | 1.7 | 1.7 |
| Qwen2.5-3B | 10 | 46 | 107 | 147 | 178 | 176 | 132 | 74 | 34 | 14 |
| Phi-4-mini | 12 | 44 | 45 | 32 | 17 | 10 | 6.6 | 4.5 | 3.2 | 2.6 |
| *Variance of the log-divergence inside generations (%)* | | | | | | | | | | |
| Llama-3.2-3B | 90 | 90 | 81 | 67 | 61 | 58 | 58 | 59 | 57 | 56 |
| Llama-3.1-8B | 87 | 81 | 64 | 51 | 47 | 43 | 36 | 29 | 26 | 25 |
| Qwen2.5-3B | 82 | 85 | 67 | 55 | 54 | 54 | 55 | 56 | 56 | 57 |
| Phi-4-mini | 83 | 77 | 60 | 50 | 50 | 51 | 53 | 55 | 56 | 54 |
| *Top token changed (% of positions)* | | | | | | | | | | |
| Llama-3.2-3B | 8 | 16 | 24 | 35 | 43 | 50 | 57 | 65 | 74 | 81 |
| Llama-3.1-8B | 9 | 16 | 25 | 36 | 45 | 52 | 62 | 73 | 84 | 91 |
| Qwen2.5-3B | 7 | 14 | 23 | 33 | 37 | 42 | 47 | 52 | 57 | 62 |
| Phi-4-mini | 9 | 17 | 28 | 40 | 48 | 55 | 62 | 69 | 75 | 80 |
| *Another token gained* $\geq 10\times$ *(% of positions)* | | | | | | | | | | |
| Llama-3.2-3B | 2 | 19 | 42 | 60 | 74 | 83 | 89 | 92 | 93 | 91 |
| Llama-3.1-8B | 3 | 34 | 64 | 80 | 88 | 89 | 91 | 90 | 85 | 94 |
| Qwen2.5-3B | 1 | 12 | 28 | 39 | 51 | 61 | 69 | 77 | 82 | 87 |
| Phi-4-mini | 2 | 26 | 51 | 68 | 79 | 88 | 93 | 96 | 97 | 98 |
| *Degenerate steered prediction (% of positions)* | | | | | | | | | | |
| Llama-3.2-3B | 0 | 0 | 0 | 2 | 5 | 14 | 21 | 31 | 45 | 61 |
| Llama-3.1-8B | 0 | 0 | 0 | 4 | 14 | 24 | 45 | 66 | 81 | 87 |
| Qwen2.5-3B | 0 | 0 | 2 | 5 | 6 | 8 | 10 | 12 | 14 | 17 |
| Phi-4-mini | 0 | 0 | 3 | 9 | 11 | 14 | 17 | 23 | 30 | 37 |

Table 3: Position-level statistics in percent. *None* counts positions where no strength of the grid changes the top token. *Flip differs* counts consecutive positions whose flip strengths differ. *Interior peak* counts promoted tokens whose probability peaks strictly inside the grid, and *peak differs* counts consecutive positions whose peak strengths differ. *Becomes top* counts promoted tokens that are the top token at some stronger strength, and *control* gives the same frequency for tokens of the same rank range that gained less than twofold.

| Backbone | positions | none | flip differs | interior peak | peak differs | becomes top | control |
|---|---|---|---|---|---|---|---|
| *Ablation set, unsteered prefix* | | | | | | | |
| Llama-3.2-3B | 23,295 | 18 | 84 | 91 | 76 | 30.4 | 8.5 |
| Llama-3.1-8B | 22,620 | 9 | 82 | 95 | 71 | 23.7 | 9.8 |
| Qwen2.5-3B | 20,406 | 37 | 72 | 67 | 75 | 32.3 | 6.4 |
| Phi-4-mini | 19,179 | 19 | 82 | 83 | 79 | 32.1 | 6.9 |
| *Ablation set, plain-steering prefix* | | | | | | | |
| Llama-3.2-3B | 20,469 | 24 | 78 | 88 | 75 | 25.6 | 4.9 |
| Llama-3.1-8B | 23,107 | 8 | 74 | 94 | 68 | 18.2 | 5.9 |
| Qwen2.5-3B | 28,217 | 53 | 57 | 54 | 64 | 21.5 | 5.6 |
| Phi-4-mini | 30,719 | 26 | 75 | 74 | 74 | 30.4 | 3.5 |
| *101-pair subset, unsteered prefix* | | | | | | | |
| Llama-3.2-3B | 5,814 | 17 | 84 | 93 | 77 | 30.2 | 4.4 |
| Llama-3.1-8B | 5,543 | 8 | 81 | 95 | 72 | 23.4 | 4.7 |
| Qwen2.5-3B | 5,315 | 38 | 72 | 67 | 76 | 31.1 | 3.8 |
| Phi-4-mini | 4,878 | 21 | 83 | 82 | 81 | 31.1 | 3.3 |
| *101-pair subset, plain-steering prefix* | | | | | | | |
| Llama-3.2-3B | 5,496 | 26 | 79 | 87 | 77 | 23.7 | 4.7 |
| Llama-3.1-8B | 6,184 | 7 | 80 | 95 | 74 | 18.6 | 5.7 |
| Qwen2.5-3B | 7,342 | 53 | 58 | 57 | 69 | 21.7 | 6.1 |
| Phi-4-mini | 6,434 | 27 | 78 | 75 | 77 | 28.7 | 3.3 |

Table 4: Three positions from unsteered Llama-3.2-3B generations on MUD. The first row of each pair gives the unsteered model's top token, which is still the top token at $\alpha = 0.3$. The second row gives a token with negligible unsteered probability that steering promotes. It becomes the top token at a stronger strength and declines again.

| Prefix | Token | $p_{t,0}$ | $p_{t,0.1}$ | $p_{t,0.2}$ | $p_{t,0.3}$ | $p_{t,0.4}$ | $p_{t,0.5}$ | $p_{t,0.7}$ | $p_{t,1.0}$ |
|---|---|---|---|---|---|---|---|---|---|
| `...Verratti` | `␣is` | 0.638 | 0.749 | 0.876 | 0.642 | 0.085 | 0.048 | 0.007 | 0.0003 |
| | `␣IS` | $<10^{-4}$ | $<10^{-4}$ | $<10^{-4}$ | 0.257 | 0.715 | 0.435 | 0.077 | 0.001 |
| `... that Mourinho believes he can` | `␣still` | 0.887 | 0.887 | 0.845 | 0.435 | 0.013 | 0.001 | $<10^{-4}$ | $<10^{-4}$ |
| | `␣PLAY` | $<10^{-4}$ | $<10^{-4}$ | $<10^{-4}$ | 0.150 | 0.457 | 0.333 | 0.018 | 0.0002 |
| `...d by the interest of top teams` | `.` | 0.973 | 0.972 | 0.971 | 0.816 | 0.274 | 0.052 | 0.004 | $<10^{-4}$ |
| | `!` | $<10^{-4}$ | $<10^{-4}$ | 0.0005 | 0.125 | 0.400 | 0.254 | 0.058 | 0.006 |

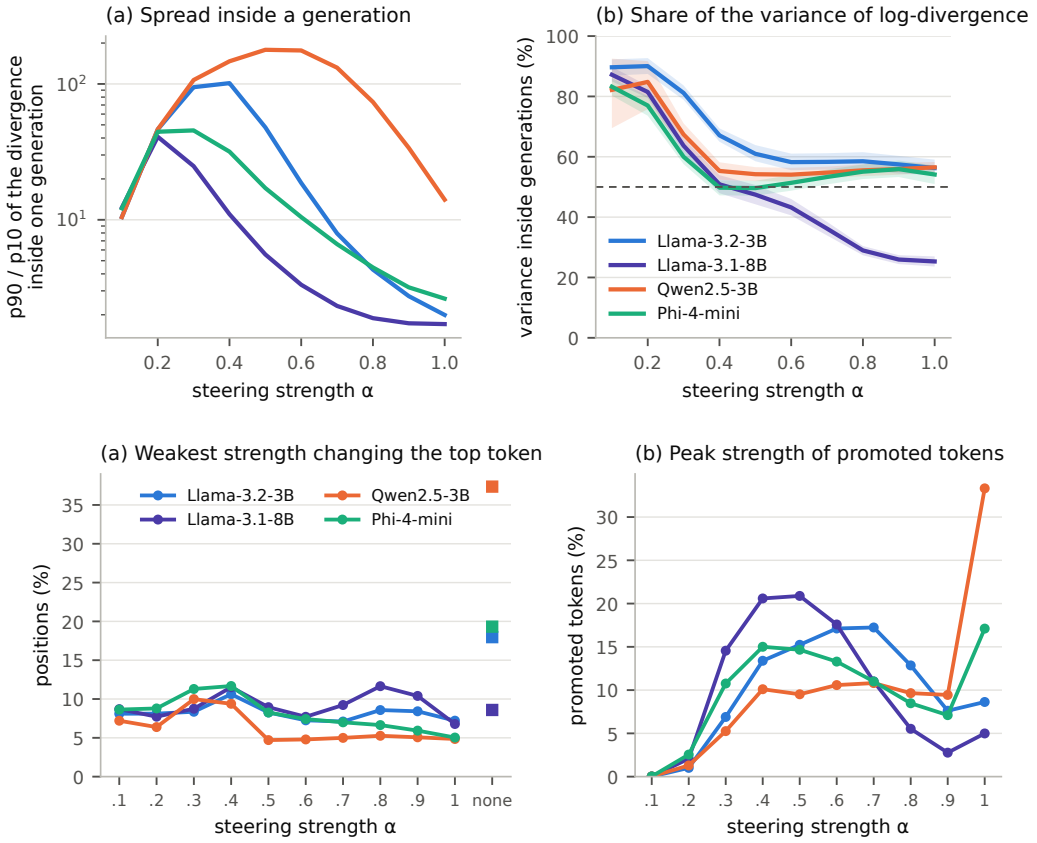



Figure 9: Position-level strength sensitivity on unsteered prefixes. **Top left:** ratio of the 90th to the 10th percentile of divergence within a generation, median over generations. **Top right:** share of log-divergence variance within generations, with intervals that resample source texts. **Bottom left:** weakest strength that changes the most probable token; squares denote positions unchanged across the grid. **Bottom right:** strength at which a promoted token reaches its maximum probability.

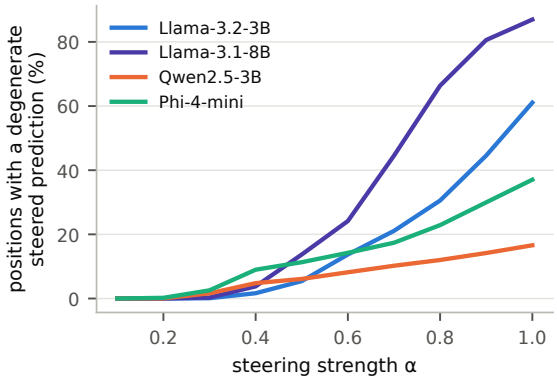



Figure 10: Positions with a degenerate prediction at each strength on unsteered prefixes, using the usable-set conditions in Eq. 8.

**Usability across strengths and positions.** Predictions fail the usable-set conditions at no more than 3% of positions up to $\alpha = 0.3$, rising to 17–87% at $\alpha = 1.0$ (Figure 10). At $\alpha = 0.7$, a failure follows another failure with probability 49–79%, compared with 6–16% after a prediction that passes the checks. Thus, usability depends on the position as well as the strength.

**Fixed-strength prefixes.** We repeat the analysis on prefixes generated at $\alpha = 0.4$, or 0.3 for Phi-4-mini (Table 1). Divergence again varies widely within generations at weak and moderate strengths. Tenfold promotion, including positions with a changed top token, outnumbers top-token changes from $\alpha = 0.2$ upward (Table 5 and Figure 11, bottom). These prefixes contain different generated text from the unsteered runs, so the comparison tests whether the pattern recurs under another prefix policy.

### A.1 Predictive Value of Base-Relative Gain

We extend the MUD ablation analysis with a token-level prediction task. A unit is a token at a fixed prefix and strength $\alpha$, with rank 2–10 under $p_{t,\alpha}$, excluding the unsteered top token. Its outcome is whether it attains rank 1 at any tested strength greater than $\alpha$ on the same prefix. All such tokens are present in the recorded candidate union. Across four backbones, there are 749,629 units at $\alpha = 0.3$, of which 246,556 satisfy $p_{t,\alpha} \geq 0.01$.

**Gain magnitude and probability strata.** Figure 12 groups tokens by $g = p_{t,\alpha}/p_{t,0}$ using $g < 2$, $[2, 5)$, $[5, 10)$, $[10, 20)$, and $g \geq 20$. At $\alpha = 0.3$, tokens in the highest bin take over more frequently than those in the lowest bin on every backbone, both with and without the probability floor (Table 6). We additionally stratify by $p_{t,\alpha}$ using boundaries 0.01, 0.03, 0.1, and 0.3. The highest-gain bin exceeds the lowest in all 20 backbone–probability strata, each with at least 30 observations in both bins. These comparisons control current probability coarsely, rather than matching it exactly. Fixing the starting strength also gives every token the same set of stronger strengths to be evaluated.

**Prediction beyond current probability.** We compare AUCs from the raw steered log-probability and log-gain with those from a logistic model using both. The joint model uses five folds grouped by source text, with feature standardization fitted on each training fold and predictions evaluated on held-out sources. Models are fitted separately for each backbone and starting strength. At $\alpha = 0.3$, adding gain improves AUC by 0.054–0.073 (Table 6). The joint model also has higher AUC in all 36 backbone–strength comparisons over $\alpha \in \{0.1, \ldots, 0.9\}$. Models that additionally include the log-probability margin to the current top token or unsteered entropy reach AUCs of 0.795–0.840 and 0.792–0.841, respectively, at $\alpha = 0.3$.

Restricting the population to $p_{t,\alpha} \geq 0.01$ retains the improvement, with AUCs of 0.78–0.82 for the joint predictor versus 0.71–0.74 for probability alone. Restricting the outcome to takeover at strengths in $(0.3, 0.6]$ also preserves the high-gain versus low-gain ordering, at 15–27% versus 2–3%. Thus,

Table 5: Per-strength statistics on plain-steering prefixes (MUD ablation set). Quantities and thresholds match Table 2.

| $\alpha$ | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1.0 |
|---|---|---|---|---|---|---|---|---|---|---|
| *Median divergence (nats)* | | | | | | | | | | |
| Llama-3.2-3B | 0.02 | 0.07 | 0.18 | 0.42 | 0.92 | 1.94 | 3.71 | 6.05 | 8.04 | 9.52 |
| Llama-3.1-8B | 0.03 | 0.17 | 0.54 | 1.34 | 2.86 | 5.05 | 7.32 | 9.83 | 11.65 | 12.62 |
| Qwen2.5-3B | 0.01 | 0.03 | 0.07 | 0.18 | 0.40 | 0.83 | 1.53 | 2.55 | 3.96 | 5.60 |
| Phi-4-mini | 0.01 | 0.05 | 0.14 | 0.42 | 1.07 | 2.32 | 4.25 | 6.47 | 8.73 | 10.76 |
| *p90 / p10 of the divergence inside one generation* | | | | | | | | | | |
| Llama-3.2-3B | 11 | 40 | 74 | 85 | 41 | 17 | 8.0 | 4.5 | 2.8 | 2.2 |
| Llama-3.1-8B | 13 | 27 | 16 | 8.6 | 4.9 | 3.1 | 2.3 | 1.9 | 1.7 | 1.7 |
| Qwen2.5-3B | 11 | 45 | 92 | 142 | 180 | 182 | 131 | 73 | 31 | 15 |
| Phi-4-mini | 12 | 42 | 52 | 43 | 26 | 15 | 10 | 6.0 | 4.3 | 3.1 |
| *Variance of the log-divergence inside generations (%)* | | | | | | | | | | |
| Llama-3.2-3B | 80 | 79 | 75 | 71 | 66 | 63 | 64 | 64 | 61 | 58 |
| Llama-3.1-8B | 75 | 70 | 62 | 54 | 50 | 45 | 38 | 31 | 27 | 25 |
| Qwen2.5-3B | 83 | 84 | 81 | 78 | 74 | 71 | 69 | 68 | 66 | 66 |
| Phi-4-mini | 78 | 75 | 70 | 66 | 64 | 63 | 62 | 63 | 64 | 65 |
| *Top token changed (% of positions)* | | | | | | | | | | |
| Llama-3.2-3B | 8 | 15 | 21 | 26 | 32 | 38 | 46 | 56 | 67 | 76 |
| Llama-3.1-8B | 8 | 14 | 20 | 25 | 32 | 42 | 54 | 71 | 82 | 92 |
| Qwen2.5-3B | 7 | 12 | 16 | 20 | 23 | 27 | 31 | 35 | 40 | 46 |
| Phi-4-mini | 7 | 13 | 19 | 25 | 32 | 41 | 50 | 59 | 66 | 73 |
| *Another token gained* $\geq 10\times$ *(% of positions)* | | | | | | | | | | |
| Llama-3.2-3B | 2 | 18 | 37 | 54 | 69 | 80 | 88 | 92 | 94 | 94 |
| Llama-3.1-8B | 5 | 31 | 60 | 77 | 85 | 88 | 92 | 90 | 83 | 93 |
| Qwen2.5-3B | 3 | 14 | 27 | 40 | 52 | 63 | 73 | 81 | 86 | 90 |
| Phi-4-mini | 3 | 19 | 38 | 55 | 69 | 80 | 88 | 94 | 97 | 98 |
| *Degenerate steered prediction (% of positions)* | | | | | | | | | | |
| Llama-3.2-3B | 0 | 0 | 0 | 1 | 3 | 9 | 16 | 25 | 37 | 53 |
| Llama-3.1-8B | 0 | 0 | 1 | 6 | 17 | 31 | 51 | 67 | 79 | 83 |
| Qwen2.5-3B | 0 | 0 | 0 | 0 | 1 | 1 | 2 | 3 | 5 | 7 |
| Phi-4-mini | 0 | 0 | 0 | 1 | 2 | 4 | 7 | 12 | 18 | 26 |

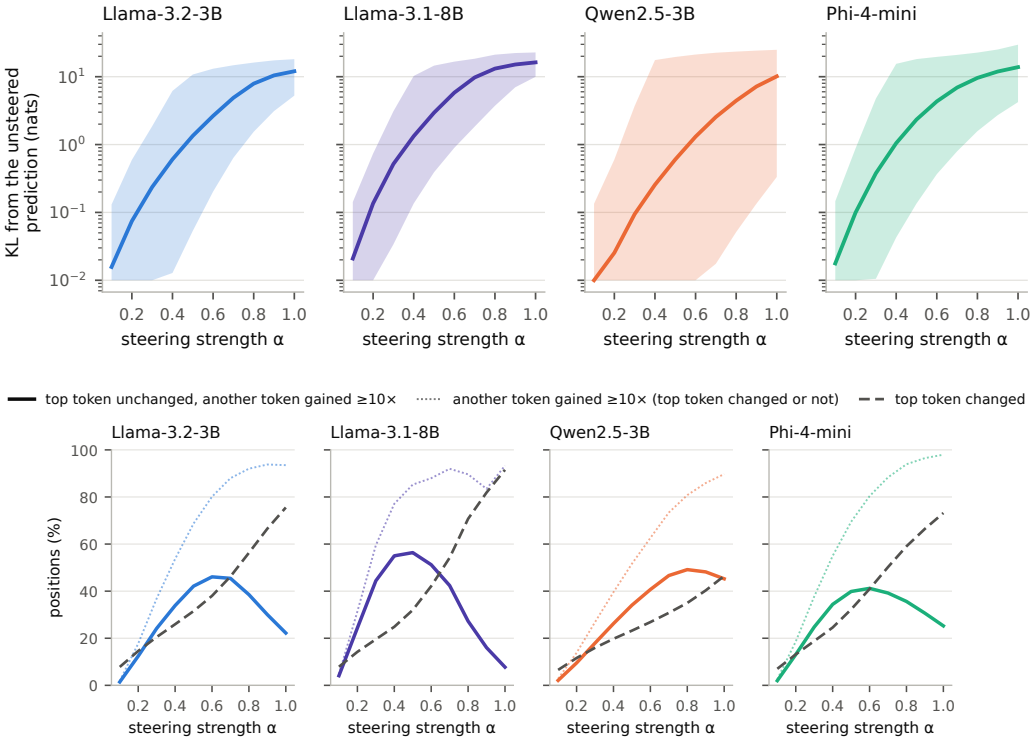



Figure 11: Per-strength comparison across four backbones. **Top row:** divergence $\mathrm{KL}(p_{t,\alpha} \,\|\, p_{t,0})$ pooled over positions of unsteered prefixes; lines give medians and bands span the 10th to 90th percentiles. **Bottom row:** on fixed-strength prefixes, positions with a tenfold-promoted non-modal token (solid) and positions whose most probable token changed (dashed).

Table 6: Full-set predictive analysis at $\alpha = 0.3$. Takeover percentages apply the $p_{t,\alpha} \geq 0.01$ floor. AUCs use all eligible rank-2–10 tokens, with $n$ token–position units. $P$ and $G$ are raw log-probability and log-gain predictors. $P + G$ is the source-grouped cross-validated logistic model.

| | | Takeover rate | | AUC | | |
|---|---|---|---|---|---|---|
| Backbone | $n$ | $g < 2$ | $g \geq 20$ | $P$ | $G$ | $P + G$ |
| Llama-3.2-3B | 204,286 | 5% | 27% | 0.756 | 0.707 | 0.829 |
| Llama-3.1-8B | 198,314 | 5% | 19% | 0.739 | 0.631 | 0.801 |
| Qwen2.5-3B | 179,199 | 6% | 31% | 0.783 | 0.752 | 0.838 |
| Phi-4-mini | 167,830 | 4% | 23% | 0.736 | 0.719 | 0.790 |

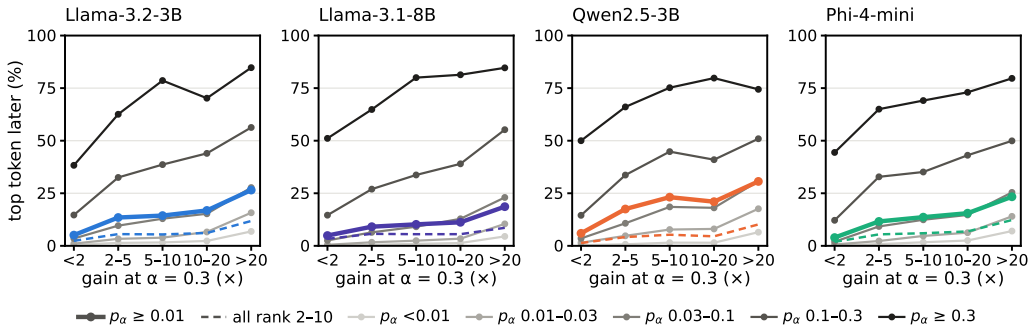



Figure 12: Gain at $\alpha = 0.3$ and takeover at a stronger tested strength on the same unsteered prefix, on the full MUD ablation set. Units exclude the unsteered top token and have steered rank 2–10. Colored solid lines apply the probability floor $p_{t,\alpha} \geq 0.01$, and colored dashed lines include all units. Gray lines, from light to dark, show probability strata $[0, 0.01)$, $[0.01, 0.03)$, $[0.03, 0.1)$, $[0.1, 0.3)$, and $[0.3, 1]$. Cells with fewer than 30 observations are omitted. The final gain bin includes gains equal to 20.

the association is present without requiring the strongest tested interventions. These results establish predictive value along the intervention trajectory, motivating gain as an additional decoding signal alongside likelihood.

# B Implementation Details

## B.1 Decoder Design and Strength Grid

**Fixed decoder design.** MAC uses the same decoder constants across MUD, Tweet, IFEval, and all evaluated backbones. Each nominating branch contributes its top five candidates, the prior is centered at the reference strength $0.2$, and the token score uses gain and repetition coefficients of $2$ and $4$, respectively. Repetition counts cover the last 24 generated tokens. These choices and the admissibility conditions below define a shared decoding rule and remain fixed throughout the experiments.

**Strength grid.** The only searched MAC decoder setting is the strength grid. We consider five-value contiguous windows of $\{0.1, 0.2, \ldots, 1.0\}$,

$$A_j = \{0.1j, 0.1(j+1), \ldots, 0.1(j+4)\}, \qquad j \in \{1, \ldots, 6\}. \tag{7}$$

For example, $A_2 = \{0.2, 0.3, 0.4, 0.5, 0.6\}$. The selected grid is fixed throughout a run. For each Llama backbone, we reuse the same grid across MUD and Tweet. IFEval and other backbones use their corresponding grid configurations. Appendix E describes the development sets used for authorship steering-strength selection. Appendix J compares MAC and MoLaCE across all six grids on the MUD development set.

**Usable set.** Writing $x^*_{t,\alpha} = \arg\max_x p_{t,\alpha}(x)$, the usable set is

$$U_t = \left\{ \alpha \in A : \begin{array}{l} H(p_{t,\alpha}) - H(p_{t,0}) \le 6, \\ \mathrm{rank}_{p_{t,0}}(x^*_{t,\alpha}) \le 1000, \\ p_{t,\alpha}(x^*_{t,\alpha}) \ge 0.05 \end{array} \right\}. \tag{8}$$

Entropy uses natural logarithms. Ranks are one-based: a token's rank is one plus the number of logits strictly greater than its own. Tied logits therefore receive equal rank.

**Decision order.** If $U_t = \varnothing$, emit the base argmax. Otherwise, consider the base model's top three tokens. A token is eligible for direct consensus when $\mathrm{rank}_{p_{t,\alpha}}(x) = 1$ for every $\alpha \in U_t$ and $c_t(x) < 2$, where $c_t$ counts the last 24 generated tokens. Accept an eligible token directly. If none is eligible, form the candidate set, compute gains, update the evidence and weights, and maximize Eq. 6. Consensus and base fallback leave the evidence state unchanged. The same checks apply when only one strength is usable.

**Candidate nomination.** The base branch, every usable steered branch, and the reference branch nominate their top five tokens for the shared candidate set. The reference branch is the grid strength nearest the prior center $0.2$. It nominates candidates even when it is outside $U_t$, but only usable strengths contribute steering evidence. Every usable strength is then evaluated on all tokens in the candidate set. The positive gain in Eq. 2 enters both the strength-level evidence and the token score, with $\tilde{p}_{t,\alpha}$ normalized over the shared candidate set.

**Prior and evidence.** The Gaussian prior in Eq. 5 is

$$\rho(\alpha) = \exp\left[-\frac{(\alpha - 0.2)^2}{2\sigma^2}\right], \qquad \sigma = \max\left\{\frac{\max A - \min A}{4}, 10^{-6}\right\}. \tag{9}$$

Every five-value grid in Eq. 7 spans $0.4$, giving the same prior width $\sigma = 0.1$ throughout these experiments. If the prior mass on usable strengths is at most $10^{-12}$, the decoder uses a uniform prior on that set. Current evidence is the candidate-normalized, probability-weighted mean of positive gains in Eq. 3, with the evidence state updated according to Eq. 4.

**Repetition.** The local count $c_t(x)$ in Eq. 6 counts occurrences of token $x$ among the last 24 generated tokens. The window and its coefficient remain fixed across settings.

**Numerical computation.** Base and steered log probabilities are computed directly with `log_softmax`. The weight update normalizes the masked prior, adds $10^{-12}$ before taking its logarithm, and subtracts the maximum log weight before exponentiation. Normalization denominators also include $10^{-12}$; the scorer normalizes the weights once more before computing expected gain. Masked strengths can retain negligible numerical mass, but their steering gain is zero.

# C Properties of the Decoder

We analyze Eq. 6 with normalized strength weights, before the numerical guards in Appendix B.

**Base-relative behavior.** When all candidate gains are zero, the score reduces to $S_t(x) = \log p_{t,0}(x) - 4c_t(x)$. Among candidates with equal repetition counts, the score therefore follows the base model's ranking. More generally, for equal base probabilities and repetition counts, the ranking depends only on the expected steering gain. With two equally weighted strengths, gains $(1, 1)$ and $(2, 0)$ receive equal scores. Unanimous top-ranked support is handled by the separate consensus rule in Section 4.

**Comparison with probability mixing.** Consider two usable strengths with equal weights and two candidates $x$ and $z$ with equal base probability $b > 0$ and repetition count. Suppose their log gains are $u(x) = (1.1, 1.1)$ and $u(z) = (2, 0)$. Their expected gains are 1.1 and 1, so the gain score prefers $x$:

$$S_t(x) - S_t(z) = 2(1.1 - 1) = 0.2. \tag{10}$$

An arithmetic probability mixture with the same weights instead assigns $be^{1.1}$ to $x$ and $b(e^2 + 1)/2$ to $z$. Since $e^{1.1} \approx 3.004 < (e^2 + 1)/2 \approx 4.195$, it prefers $z$. For example, $b = 0.06$ allows these probabilities to coexist in normalized distributions with both candidates among the leading predictions. This comparison holds the weights fixed to isolate the token objective; MAC and MoLaCE can also assign different weights to the same strengths.

**Rank invariance.** The base rank of the steered modal token is unchanged by strictly increasing transformations of the base and steered logits, because these preserve the base ordering and steered argmax. This property applies to the rank condition; entropy differences and peak probabilities still depend on logit scale.

## C.1 Relation to Likelihood-Ratio Guidance

Fix nonnegative weights $q_a$ with $\sum_a q_a = 1$, remove rectification, usability filtering, and repetition cost, and bypass direct consensus. The scoring rule becomes

$$\begin{aligned} S(x) &= \log p_0(x) + \lambda \sum_a q_a \big(\log p_a(x) - \log p_0(x)\big) \\ &= (1 - \lambda) \log p_0(x) + \lambda \sum_a q_a \log p_a(x). \end{aligned} \tag{11}$$

Thus, over the scoring domain, maximizing $S$ maximizes $p_0^{1-\lambda} \prod_a p_a^{\lambda q_a}$. For $\lambda > 1$, the base factor has a negative exponent and acts as a reference penalty. A full-vocabulary decoder additionally removes MAC's candidate restriction.

With one strength and $\lambda = 2$, this gives $2 \log p_a - \log p_0$, the classifier-free guidance form with the steered and unsteered distributions replacing conditional and unconditional branches (Sanchez et al., 2024). It is related to, but not identical to, the original contrastive-decoding objective $\log p_{\text{expert}} - \log p_{\text{amateur}}$ with its expert plausibility constraint (Li et al., 2023b). DExperts adds an expert–anti-expert log-ratio to a base-model score (Liu et al., 2021). ASD uses $(1 + \beta)z_{+\alpha} - \beta z_{-\alpha}$, contrasting opposite interventions rather than using an unsteered reference (Su et al., 2025).

MAC determines which steering evidence contributes at each position. Rectification gives demoted tokens zero steering bonus rather than a negative contribution. The usable set excludes branches with excessive entropy increase, low peak probability, or an implausible modal token. Multiple strengths provide alternative usable predictions, and the prior and evidence state determine their

relative contributions. Consensus and repetition control further distinguish the complete decoder from the simplified score. On the MUD ablation set, base fallback occurs at 0.4% and 0.1% of generated tokens for MAC on Llama-3.2-3B and Llama-3.1-8B, respectively, compared with 6.2% and 5.3% for the selected rectified single-strength controls.

**Style improvement with bounded meaning loss.** We compare four single-strength guidance controls on MUD ablation set. The rectified control retains MAC's positive gain, candidate set, usability filter, consensus, and repetition rules. The signed control removes only rectification. The CFG-form control uses $\log p_0 + \lambda(\log p_\alpha - \log p_0)$ over the full vocabulary without these guards. The ASD-form control uses opposite interventions with the same authorship direction, adapting its decoding formula rather than reproducing its visual-hallucination setup.

For each family, we select the highest development StyleCAV among settings whose development SBERT is at least MAC's development SBERT minus $\delta$. We evaluate $\delta = 0.03$ and $0.05$, absolute SBERT differences. Rectified and signed controls sweep $\alpha \in \{0.2, 0.3, 0.4, 0.5, 0.6\}$, with $\lambda \in \{0.5, 1, 2, 3, 4\}$ and $\{0.5, 1, 2, 4\}$, respectively. CFG-form controls use the same strengths and $\lambda \in \{1.5, 2, 3\}$, excluding $(3, 0.6)$ on 8B. ASD-form controls sweep $\alpha \in \{0.3, 0.4, 0.5\}$ and $\beta \in \{0.5, 1, 2\}$. No ASD setting meets the development meaning bound, so we report its most meaning-preserving setting. MAC uses its default grid and $\lambda = 2$ in the same run batch.

Table 7: Held-out comparisons on the MUD ablation set. Controls are selected on the MUD development pairs using a 0.03 absolute SBERT margin. Differences are MAC minus control, with source-bootstrap 95% confidence intervals. MAC's held-out StyleCAV/SBERT are 0.456/0.758 (3B) and 0.404/0.817 (8B). ASD reports the most meaning-preserving development setting because none meets the selection margin.

| Control | $(\lambda, \alpha)$ or $(\beta, \alpha)$ | $\Delta$StyleCAV [95% CI] | $\Delta$SBERT [95% CI] |
|---|---|---|---|
| **Llama-3.2-3B** | | | |
| Rectified | $(2, 0.5)$ | $+0.043$ $[0.016, 0.071]$ | $-0.005$ $[-0.024, 0.015]$ |
| Signed | $(2, 0.5)$ | $+0.045$ $[0.017, 0.072]$ | $-0.002$ $[-0.024, 0.021]$ |
| CFG form | $(1.5, 0.3)$ | $+0.074$ $[0.042, 0.108]$ | $-0.001$ $[-0.023, 0.021]$ |
| ASD form | $(0.5, 0.3)$ | $+0.036$ $[0.009, 0.062]$ | $+0.104$ $[0.081, 0.127]$ |
| **Llama-3.1-8B** | | | |
| Rectified | $(2, 0.4)$ | $+0.065$ $[0.032, 0.102]$ | $-0.007$ $[-0.027, 0.011]$ |
| Signed | $(2, 0.4)$ | $+0.071$ $[0.037, 0.107]$ | $-0.007$ $[-0.026, 0.008]$ |
| CFG form | $(1.5, 0.2)$ | $+0.072$ $[0.025, 0.121]$ | $-0.021$ $[-0.039, -0.004]$ |
| ASD form | $(0.5, 0.3)$ | $+0.140$ $[0.080, 0.200]$ | $+0.442$ $[0.378, 0.502]$ |

We test the joint criterion that the lower 95% confidence bound of $\Delta\mathrm{StyleCAV}$ exceeds zero and that of $\Delta\mathrm{SBERT}$ exceeds $-\delta$, with differences defined as MAC minus control. Intervals use paired bootstrap resampling over source texts. MAC satisfies both conditions at $\delta = 0.03$ against the rectified and signed controls on both backbones. Against the CFG-form control, the criterion holds at $0.03$ on 3B and $0.05$ on 8B. Against the tested ASD-form control, both style and meaning improve. These results support stronger style expression with bounded meaning loss at the reported operating points.

**How often is gain scoring used?** The rectified single-strength controls retain the same consensus checks as MAC. Their comparisons therefore evaluate the complete decoding procedures, including consensus, rather than gain scoring at every position. Table 8 reports the share of all generated tokens using each decision path on the MUD ablation set. Consensus occurs more often for the selected single-strength controls, while MAC uses weighted scoring more frequently.

Table 8: Decision paths as percentages of generated tokens on the MUD ablation set. The single-strength configurations are $\alpha = 0.5$ on 3B and $0.4$ on 8B, both with $\lambda = 2$.

| Backbone | Decoder | Consensus | Scoring | Base fallback |
|---|---|---|---|---|
| Llama-3.2-3B | Single strength | 71.4 | 21.2 | 6.2 |
| | MAC | 54.5 | 44.0 | 0.4 |
| Llama-3.1-8B | Single strength | 78.8 | 14.7 | 5.3 |
| | MAC | 62.7 | 36.5 | 0.1 |

# D Experimental Details for Authorship Style Transfer

## D.1 Authorship Steering Directions

For authorship transfer, let $X$ be the source text and $E$ a target-author exemplar. The decoder receives a prompt containing $X$ and $E$ and a fixed layerwise style direction. We construct this direction following Tran et al. (2026).

**Four-aspect contrastive directions.** For every authorship-transfer task, we follow the A3S construction (Tran et al., 2026) with four aspects: $\mathcal{D} = \{\textit{Figurative Language}, \textit{Tone}, \textit{Perspective}, \textit{Structure}\}$. Let $\bar{h}_\ell(r)$ denote the residual-stream representation of text $r$, mean-pooled over its tokens at layer $\ell$. The target exemplar provides the positive text $E_d^+ = E$. For each aspect $d$, we generate $K = 4$ contrastive rewrites $E_{d,k}^-$ that change that aspect while preserving content, and form

$$v_{\ell,d} = \frac{1}{K} \sum_{k=1}^{K} \left[ \bar{h}_\ell(E_d^+) - \bar{h}_\ell(E_{d,k}^-) \right]. \tag{12}$$

Each direction captures an exemplar-grounded contrast along one stylistic aspect. The same frozen backbone constructs the contrastive texts and supplies their activations.

**PCB aggregation.** We merge the four directions independently at each intervention layer using the activation-space adaptation of Parameter Competition Balancing (PCB) (DU et al., 2024; Tran et al., 2026):

$$v_\ell = \mathrm{PCB}_{r=0.5}\left(\{v_{\ell,d} : d \in \mathcal{D}\}\right). \tag{13}$$

PCB weights components using their importance and agreement across directions; the A3S implementation retains the top half of the importance scores before min–max normalization. This construction is shared across the authorship steering baselines and MAC. It supplies one merged direction per layer, held fixed during decoding for that target profile. The aspect decomposition and PCB aggregation are inherited from A3S; the contribution here is how the resulting strength-conditioned predictions determine the next token.

Authorship steering methods share the A3S generation prompt and direction construction, with a generation budget of 256 tokens. We use the A3S contrastive prompts and apply steering during both prefill and decoding. Table 9 gives the intervention layers. Relative indices count backward from the final transformer block ($-1$); ranges include both endpoints. For every backbone, the layer range is identical on MUD and Tweet.

Table 9: Authorship-transfer intervention layers. Absolute block indices are zero-based; negative indices are relative to the final transformer block.

| Backbone | # Blocks | Relative range | Absolute range |
|---|---|---|---|
| Llama-3.2-3B | 28 | $-19$ to $-10$ | 9–18 |
| Llama-3.1-8B | 32 | $-19$ to $-10$ | 13–22 |
| Qwen2.5-3B | 36 | $-24$ to $-5$ | 12–31 |
| Phi-4-mini | 32 | $-24$ to $-5$ | 8–27 |

### D.2 Pairwise LLM Evaluation of Authorship Style

We use GPT-4.1 to compare MAC against fixed-strength, Hybrid Search, DAC, and MoLaCE on MUD and Tweet with Llama-3.2-3B and Llama-3.1-8B. Each baseline comparison covers all 3,600 MUD pairs or all 1,496 Tweet pairs. Dataset construction is given in Appendix E.

**Judge prompt.** We reuse the pairwise style-preference prompt from Tran et al. (2026) (Figure 13), whose agreement with human judgments was studied in A3S. The target-author exemplar fills `reference_text`, and the two generated outputs fill `post_one` and `post_two`. The judge selects which output is more likely to share the exemplar's author based on writing style, provides a brief stylistic justification, and may return a tie. The original source text is not provided, so the judge evaluates target-style similarity rather than content preservation. Each pair is judged once with randomized output order.

**Paired inference.** The decisive win rate is wins divided by wins plus losses, excluding ties and invalid verdicts. Its 95% percentile intervals use 10,000 bootstrap resamples, jointly resampling wins and decisive totals. Paired sign-flip tests use 10,000 draws and test the win–loss margin against zero. Both use seed 0 and the sampling units of Appendix E, with one sign per MUD source cluster or Tweet pair. Holm adjustment covers four baselines per dataset–backbone setting. All adjusted $p$-values are below 0.001.

Table 10: GPT-4.1 style preferences for Llama-3.2-3B. $W$, $U$, and $L$ are MAC wins, non-decisions (ties or invalid verdicts), and losses. Win rates exclude ties and invalid verdicts, $W/(W+L)$. Net preference uses all pairs, $(W-L)/n$. Both are percentages with 95% bootstrap intervals. All paired sign-flip comparisons have Holm-adjusted $p < 0.001$.

| Baseline | $W$ | $U$ | $L$ | Win rate | Net preference |
|---|---|---|---|---|---|
| *MUD ($n = 3,600$ per baseline)* | | | | | |
| Fixed-strength | 2511 | 661 | 428 | 85.4 [84.0, 86.9] | 57.9 [54.0, 61.7] |
| Hybrid Search | 2332 | 646 | 622 | 78.9 [77.2, 80.6] | 47.5 [43.8, 51.2] |
| DAC | 2629 | 650 | 321 | 89.1 [87.7, 90.4] | 64.1 [60.1, 68.1] |
| MoLaCE | 2063 | 744 | 793 | 72.2 [70.1, 74.3] | 35.3 [31.4, 39.2] |
| *Tweet ($n = 1,496$ per baseline)* | | | | | |
| Fixed-strength | 1032 | 130 | 334 | 75.5 [73.2, 77.8] | 46.7 [42.4, 50.8] |
| Hybrid Search | 991 | 115 | 390 | 71.8 [69.3, 74.2] | 40.2 [35.6, 44.7] |
| DAC | 1174 | 98 | 224 | 84.0 [82.1, 85.9] | 63.5 [59.8, 67.1] |
| MoLaCE | 949 | 132 | 415 | 69.6 [67.1, 72.0] | 35.7 [31.1, 40.2] |

Table 11: GPT-4.1 style preferences for Llama-3.1-8B. $W$, $U$, and $L$ are MAC wins, non-decisions (ties or invalid verdicts), and losses. Win rates exclude ties and invalid verdicts, $W/(W+L)$. Net preference uses all pairs, $(W-L)/n$. Both are percentages with 95% bootstrap intervals. All paired sign-flip comparisons have Holm-adjusted $p < 0.001$.

| Baseline | $W$ | $U$ | $L$ | Win rate | Net preference |
|---|---|---|---|---|---|
| *MUD ($n = 3,600$ per baseline)* | | | | | |
| Fixed-strength | 2074 | 836 | 690 | 75.0 [72.8, 77.2] | 38.4 [34.6, 42.3] |
| Hybrid Search | 2009 | 742 | 849 | 70.3 [68.1, 72.5] | 32.2 [28.4, 36.0] |
| DAC | 2489 | 763 | 348 | 87.7 [86.1, 89.3] | 59.5 [55.3, 63.5] |
| MoLaCE | 2131 | 846 | 623 | 77.4 [75.3, 79.5] | 41.9 [38.0, 46.0] |
| *Tweet ($n = 1,496$ per baseline)* | | | | | |
| Fixed-strength | 1076 | 91 | 329 | 76.6 [74.4, 78.8] | 49.9 [45.7, 54.2] |
| Hybrid Search | 965 | 64 | 467 | 67.4 [65.0, 69.8] | 33.3 [28.7, 37.9] |
| DAC | 1252 | 64 | 180 | 87.4 [85.7, 89.1] | 71.7 [68.2, 74.9] |
| MoLaCE | 960 | 94 | 442 | 68.5 [66.0, 70.9] | 34.6 [30.0, 39.2] |

```
You are an experienced linguist specializing in authorship and writing
style analysis.
Given a REFERENCE TEXT and two TARGET TEXTS, determine which TARGET TEXT
is significantly more likely to have been written by the same author as
the REFERENCE TEXT.

Base your judgment solely on writing style, including (but not limited
to):
- Linguistic and stylistic patterns
- Sentence structure and organization
- Ordering and presentation of information
- Tone, formality, and rhetorical habits
Do not rely on topic similarity or factual content except where it
reflects stylistic choices.

For your response, follow these instructions:
1.  Output TEXT ONE if it is significantly more likely to be written by
the same author as the REFERENCE TEXT.
2.  Output TEXT TWO if it is significantly more likely to be written by
the same author as the REFERENCE TEXT.
3.  Output BOTH if either text could plausibly have been written by the
same author, or if neither stands out as more likely.

Here are the texts:
REFERENCE TEXT:
{reference_text}
TEXT ONE:
{post_one}
TEXT TWO:
{post_two}

First, provide a brief justification for your judgment based on
stylistic evidence.
Then output exactly one of the following on a new line:  TEXT ONE, TEXT
TWO, or BOTH.
```

Figure 13: Pairwise preference prompt adapted from Tran et al. (2026).

# E FULL STYLE-TRANSFER RESULTS

**Dataset construction.** We follow the authorship-transfer setup of A3S (Tran et al., 2026). For MUD (Khan et al., 2021), we sample 15 source and 15 target authors from the test split. Each of the 16 texts per source author is paired with every target author, represented by a concatenated exemplar, giving $16 \times 15 \times 15 = 3{,}600$ evaluation pairs. For LaMP Tweet Paraphrasing (Salemi et al., 2024), each source Tweet is paired with a distinct target user. Nine historical posts are randomly selected and concatenated to form the target exemplar, yielding 1,496 examples. For each task, we additionally construct a development set of 100 samples for hyperparameter search, kept separate from the final evaluation set.

**Metrics.** Table 12 reports the three evaluation scores for the methods compared in Figure 3. StyleCAV (Wegmann et al., 2022) and StyleDistance (Patel et al., 2025) measure towards-target similarity. SBERT (Reimers & Gurevych, 2019) measures similarity to the source. Higher values are better for all three metrics. StyleCAV remains on its original $[-1, 1]$ scale.

Table 12: Full style-transfer results. Higher is better for every metric. **Bold** marks the best value within each task and backbone.

| Method | StyleCAV ↑ | StyleDistance ↑ | SBERT ↑ |
|---|---|---|---|
| *MUD — Llama-3.1-8B* | | | |
| Unsteered Decoding | -0.106 | 0.635 | 0.817 |
| Fixed-strength | 0.269 | 0.734 | 0.792 |
| DAC | 0.122 | 0.698 | 0.806 |
| MoLaCE | 0.226 | 0.732 | **0.828** |
| Hybrid Search | 0.311 | 0.747 | 0.746 |
| **MAC (ours)** | **0.370** | **0.766** | 0.783 |
| *MUD — Llama-3.2-3B* | | | |
| Unsteered Decoding | -0.171 | 0.623 | 0.740 |
| Fixed-strength | 0.233 | 0.726 | 0.742 |
| DAC | 0.184 | 0.714 | **0.750** |
| MoLaCE | 0.346 | 0.743 | 0.733 |
| Hybrid Search | 0.348 | 0.741 | 0.655 |
| **MAC (ours)** | **0.414** | **0.757** | 0.715 |
| *Tweet — Llama-3.1-8B* | | | |
| Unsteered Decoding | -0.297 | 0.530 | **0.828** |
| Fixed-strength | 0.331 | 0.700 | 0.783 |
| DAC | 0.124 | 0.648 | 0.802 |
| MoLaCE | 0.383 | 0.725 | 0.783 |
| Hybrid Search | 0.415 | 0.719 | 0.709 |
| **MAC (ours)** | **0.496** | **0.764** | 0.732 |
| *Tweet — Llama-3.2-3B* | | | |
| Unsteered Decoding | -0.365 | 0.525 | 0.710 |
| Fixed-strength | 0.241 | 0.681 | 0.709 |
| DAC | 0.025 | 0.627 | 0.737 |
| MoLaCE | 0.238 | 0.687 | **0.749** |
| Hybrid Search | 0.382 | 0.707 | 0.669 |
| **MAC (ours)** | **0.446** | **0.745** | 0.676 |

**Qualitative examples.** Tables 14 and 15 compare MAC and MoLaCE on three illustrative examples.

**Statistical Comparisons** Authorship-transfer comparisons pair outputs on the same source–target examples. Resampling uses source-text clusters for MUD and individual pairs for Tweet. The full evaluation sets contain 240 MUD sources with 15 targets each (3,600 pairs) and 1,496 Tweet pairs. Figure 3 shows marginal 95% bootstrap intervals for metric means. Significance annotations use paired tests with Holm adjustment across baselines within each setting and metric.

Table 13: Style transfer across model families on MUD. StyleCAV and StyleDistance measure target-style similarity; SBERT measures source-meaning preservation. Higher is better for all columns; **bold** marks the best value within each model.

| Method | StyleCAV ↑ | StyleDistance ↑ | SBERT ↑ |
|---|---|---|---|
| *Qwen2.5-3B* | | | |
| DAC | 0.226 | 0.719 | 0.781 |
| MoLaCE | 0.290 | 0.736 | **0.800** |
| Hybrid Search | 0.277 | 0.736 | 0.685 |
| **MAC (ours)** | **0.344** | **0.744** | 0.756 |
| *Phi-4-mini* | | | |
| DAC | 0.151 | 0.700 | **0.839** |
| MoLaCE | 0.233 | 0.724 | 0.830 |
| Hybrid Search | 0.285 | 0.737 | 0.799 |
| **MAC (ours)** | **0.337** | **0.747** | 0.802 |

Table 14: Two authorship-transfer examples with the same target-author exemplar. Excerpts retain its capitalization, spelling, and informal phrasing. MAC makes stronger surface-style changes, with content errors discussed in the text. Colored spans highlight selected output differences.

| | |
|---|---|
| **Target** | `toasties da best friend :))))  [...]  how FUCKING GOOD IS TYHIS GAME AND HOW IT NEVER ACTGUALLY GETS BETTER [...] someTimes I jUST get FURIOUS and REALLY fucking mad beCASUE they're JUST dumb and FUCK this GAME [...]  Edit:  cuz u sound like a megadweeb getting so mad ))` |
| *Example 1: recalibrating visual estimates* | |
| **Source** | `Just a heads up, remember to recalibrate your eyes (weigh and measure) periodically to make sure that your eyeballing method is still accurate.` |
| **MAC** | `Heads up:  Make Sure To RECALIBRATE YOUR EYES REGULARLY REMEMBER TO PERIODECICALLY CHECK YOUR EYE MEASUREMENTS (WEIGHT AND MEASURE) TO ENSURE YOUR EYEBALLING METHOD IS STILL ACCURATE` |
| **MoLaCE** | `Just a heads up, remember to check your eye calibration (calibrate your eyes) periodically to make sure your eyeballing method is still accurate.` |
| *Example 2: continued enjoyment* | |
| **Source** | `It's incredible.  Using it for months and haven't gotten sick of it like some other non standard flavors.` |
| **MAC** | `I'm using IT for like 3 MONTHS now and I dont GET SICK OF IT its flavor,its the only non standard FLAVOR IVE BEENUSING` |
| **MoLaCE** | `I'm using it for months now and I'm still not sick of it, which is crazy for a non standard flavor` |

The tables report MAC minus each baseline. Stars denote adjusted significance ($^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$), and ns indicates that the test does not establish a difference. Neither overlapping intervals nor nonsignificant tests establish equivalence. Pareto membership describes observed means, and dashed segments between evaluated methods do not represent measured intermediate settings.

## F Fixed-Strength Sensitivity

The fixed-strength sweeps show how steering strength changes target-style similarity and meaning preservation. StyleCAV and StyleDistance measure towards-target similarity, while SBERT measures similarity to the source. Figure 3 reports the strength with the best style–meaning trade-off.

Table 15: An authorship-transfer example with informal spelling, sparse punctuation, and laughter in the target style. MAC adopts more of these features, while both outputs retain the broad baseball-related message. Colored spans highlight selected output differences.

| | |
|---|---|
| **Target** | This analogy doesnt really work for jim, per se [...] for chrissakes hes saying BRAVO again! tss tell us how you really feel [...] well lets just do it for another year or two, just to be safe. hahahahahaha....i doubt it. was that a joke? [...] its been 2 years and were still here. i still havent |
| *Example 3: concern about Strasburg* | |
| **Source** | Sucks for Strasburg. He's had a hell of a season after some issues last year and looked like he had it all sorted out. Hopefully it's not TJx2 for him...We haven't had a Harvey/Strasburg match up when both are firing on all cylinders in a long long time... |
| **MAC** | Bad lucks strasburg hes had a insane season after gettin thru some issues last year looked like he good to go hopefully aint tjjx2 for him<br>we havent had a harveystraburg match up when both are goin full throttle in a woooow long time lolol. |
| **MoLaCE** | Tough break for Strasburg. He's had a great season after dealing with some issues last year and looked like he was back on track. Fingers crossed its not a repeat of the TJ injury for him...We havent had a harvey/strausberg matchup when both are goinng good in a long tim |

Table 16: Paired differences (MAC minus baseline) for Llama-3.1-8B.

| Baseline | $\Delta StyleCAV$ | $\Delta SBERT$ |
|---|---|---|
| *MUD — Llama-3.1-8B* | | |
| Unsteered Decoding | $+0.476^{***}$ | $-0.034^{***}$ |
| Fixed-strength | $+0.101^{***}$ | $-0.009^{ns}$ |
| DAC | $+0.248^{***}$ | $-0.022^{**}$ |
| MoLaCE | $+0.144^{***}$ | $-0.045^{***}$ |
| Hybrid Search | $+0.059^{***}$ | $+0.037^{***}$ |
| *Tweet — Llama-3.1-8B* | | |
| Unsteered Decoding | $+0.794^{***}$ | $-0.096^{***}$ |
| Fixed-strength | $+0.165^{***}$ | $-0.051^{***}$ |
| DAC | $+0.372^{***}$ | $-0.069^{***}$ |
| MoLaCE | $+0.114^{***}$ | $-0.051^{***}$ |
| Hybrid Search | $+0.081^{***}$ | $+0.024^{***}$ |

Table 17: Paired differences (MAC minus baseline) for Llama-3.2-3B.

| Baseline | $\Delta StyleCAV$ | $\Delta SBERT$ |
|---|---|---|
| *MUD — Llama-3.2-3B* | | |
| Unsteered Decoding | $+0.585^{***}$ | $-0.025^{**}$ |
| Fixed-strength | $+0.181^{***}$ | $-0.028^{***}$ |
| DAC | $+0.231^{***}$ | $-0.036^{***}$ |
| MoLaCE | $+0.068^{***}$ | $-0.018^{***}$ |
| Hybrid Search | $+0.066^{***}$ | $+0.060^{***}$ |
| *Tweet — Llama-3.2-3B* | | |
| Unsteered Decoding | $+0.811^{***}$ | $-0.034^{***}$ |
| Fixed-strength | $+0.205^{***}$ | $-0.033^{***}$ |
| DAC | $+0.420^{***}$ | $-0.061^{***}$ |
| MoLaCE | $+0.207^{***}$ | $-0.073^{***}$ |
| Hybrid Search | $+0.064^{***}$ | $+0.008^{ns}$ |

Table 18: Paired differences (MAC minus baseline) for Qwen2.5-3B, Phi-4-mini.

| Baseline | $\Delta StyleCAV$ | $\Delta StyleDistance$ | $\Delta SBERT$ |
|---|---|---|---|
| *MUD — Qwen2.5-3B* | | | |
| DAC | $+0.118^{***}$ | $+0.025^{***}$ | $-0.024^{***}$ |
| MoLaCE | $+0.053^{***}$ | $+0.008^{***}$ | $-0.044^{***}$ |
| Hybrid Search | $+0.067^{***}$ | $+0.008^{***}$ | $+0.071^{***}$ |
| *MUD — Phi-4-mini* | | | |
| DAC | $+0.186^{***}$ | $+0.047^{***}$ | $-0.036^{***}$ |
| MoLaCE | $+0.104^{***}$ | $+0.023^{***}$ | $-0.027^{***}$ |
| Hybrid Search | $+0.051^{***}$ | $+0.010^{***}$ | $+0.004^{\mathrm{ns}}$ |

Table 19: Fixed-strength steering sweeps. Higher is better for every metric.

| | Llama-3.1-8B | | | Llama-3.2-3B | | |
|---|---|---|---|---|---|---|
| $\alpha$ | StyleCAV | StyleDistance | SBERT | StyleCAV | StyleDistance | SBERT |
| *MUD* | | | | | | |
| 0.0 | -0.106 | 0.635 | 0.817 | -0.171 | 0.623 | 0.740 |
| 0.1 | -0.064 | 0.651 | 0.833 | -0.128 | 0.639 | 0.752 |
| 0.2 | 0.038 | 0.680 | 0.843 | -0.069 | 0.658 | 0.768 |
| 0.3 | 0.183 | 0.719 | 0.845 | 0.092 | 0.699 | 0.776 |
| 0.4 | 0.269 | 0.734 | 0.792 | 0.233 | 0.726 | 0.742 |
| 0.5 | 0.276 | 0.730 | 0.699 | 0.296 | 0.732 | 0.685 |
| 0.6 | 0.232 | 0.714 | 0.575 | 0.300 | 0.721 | 0.618 |
| 0.7 | 0.170 | 0.697 | 0.418 | 0.283 | 0.704 | 0.545 |
| 0.8 | 0.142 | 0.682 | 0.298 | 0.242 | 0.694 | 0.443 |
| *Tweet Paraphrasing* | | | | | | |
| 0.0 | -0.297 | 0.530 | 0.828 | -0.365 | 0.525 | 0.710 |
| 0.1 | -0.224 | 0.550 | 0.850 | -0.317 | 0.538 | 0.735 |
| 0.2 | 0.092 | 0.637 | 0.842 | -0.119 | 0.591 | 0.763 |
| 0.3 | 0.331 | 0.700 | 0.783 | 0.241 | 0.681 | 0.709 |
| 0.4 | 0.405 | 0.720 | 0.621 | 0.319 | 0.680 | 0.466 |
| 0.5 | 0.370 | 0.701 | 0.314 | 0.208 | 0.627 | 0.175 |
| 0.6 | 0.196 | 0.639 | 0.137 | 0.143 | 0.602 | 0.090 |
| 0.7 | 0.088 | 0.598 | 0.094 | 0.128 | 0.593 | 0.077 |
| 0.8 | 0.047 | 0.580 | 0.083 | 0.113 | 0.586 | 0.074 |

# G Experimental Details for IFEval Evaluation

IFEval (Zhou et al., 2023) evaluates adherence to instructions that can be checked programmatically. We evaluate 362 single-instruction prompts from the Format and Keywords groups, following the task decomposition of Stolfo et al. (2025), using Llama-3.2-3B. These prompts impose verifiable constraints on the response, such as required formatting or the inclusion of specified words.

All methods are evaluated on the same prompts within each input condition. This evaluation provides a complementary test beyond authorship transfer, examining whether the methods can satisfy explicit, verifiable constraints while preserving task fidelity and writing quality. Table 20 illustrates the types of constraints considered in the Format and Keywords groups; the examples describe the instruction types rather than reproduce individual evaluation queries.

Table 20: Representative instruction types in the Format and Keywords groups. Examples are illustrative paraphrases. "Format" follows their broad grouping, which also includes response-language constraints.

| Group | Instruction types | Illustrative requirement |
|---|---|---|
| Format | JSON; quotation | Return a valid JSON response; enclose the response in quotation marks. |
| | Letter case | Use lowercase or uppercase throughout, or include a specified number of uppercase words. |
| | Punctuation | Omit commas. |
| | Sections; bullet points | Produce $n$ labeled sections or exactly $n$ bullets. |
| | Highlighting; title | Highlight specified text spans or delimit the title. |
| | Ending; restricted reply | End with a specified phrase or choose from a fixed set of replies. |
| | Response language | Write in a specified language. |
| Keywords | Inclusion; exclusion | Include a specified word, or avoid it throughout the response. |

**Base queries and input conditions.** Following Stolfo et al. (2025), we distinguish the base query, which specifies the requested content, from the instruction that constrains its presentation or word choice. For example, a request to explain photosynthesis can be paired with a requirement to produce JSON. With explicit instructions, the model receives both parts. Without explicit instructions, it receives the base query alone, while the steering direction represents the target constraint. The response is checked against that same constraint in both conditions. This separates reinforcement of a stated instruction from control supplied through steering alone. Appendix G describes the direction construction.

**Constraint compliance.** We use the IFEval checker to compute strict and loose prompt-level accuracy from the generated responses (Zhou et al., 2023). Strict accuracy applies the constraint checker directly to the response. Loose accuracy also permits the benchmark's prescribed transformations, including removing Markdown asterisks or an opening or closing line, to reduce superficial false negatives. It does not use an LLM to decide compliance. Since each evaluated prompt has one target instruction, prompt-level and instruction-level correctness coincide. We report strict and loose scores separately.

**Instruction-Following Steering Directions** For IFEval, we follow the direction construction of Stolfo et al. (2025). For a target instruction $I$, let $P_i$ be a base query and $P_i^I$ the same query with $I$ included. Let $h_\ell^{\text{last}}(P)$ denote the residual-stream state at the final input token of $P$ at layer $\ell$. We average the paired differences and normalize the resulting direction,

$$d_{\ell,I} = \frac{1}{N_I} \sum_{i=1}^{N_I} \left[ h_\ell^{\text{last}}(P_i^I) - h_\ell^{\text{last}}(P_i) \right], \qquad v_{\ell,I} = \frac{d_{\ell,I}}{\|d_{\ell,I}\|_2}, \tag{14}$$

where $N_I$ is the number of query pairs used to construct the direction. MAC uses $v_{\ell,I}$ in Eq. 1 and combines the resulting strength-conditioned predictions during decoding.

```
You are an impartial evaluator of instruction-following tasks.
You will receive a "query" and a "response".  In this dataset, each
query contains:
(1) the task to be completed, and
(2) additional formatting instructions.

For this evaluation, IGNORE the formatting instructions entirely because
they are evaluated elsewhere.  Focus only on whether the response
completed the TASK content correctly and whether the writing quality is
adequate.

Scoring rules:
- task_fidelity:  1 if (and only if) the response fulfills the task
asked in the query; else 0.
- text_quality:  an integer from 1 to 5 reflecting clarity, coherence,
fluency, and appropriateness of tone for the task (1=very poor,
5=excellent).

Provide also a brief justification (1-3 concise sentences) that explains
your decision, referring only to task fulfillment and writing quality.
Return STRICT JSON with the keys:  task_fidelity, text_quality,
justification.

Query:
{query}

Response:
{response}
```

Figure 14: Task fidelity and text quality LLM-as-a-Judge prompt adapted from Kang & Kim (2026).

**LLM judging protocol.** Constraint satisfaction alone does not establish that a response answers the query well. We therefore assess task fidelity and text quality using GPT-4.1 with the judge template from Kang & Kim (2026). The full prompt is shown in Figure 14. In a blinded study on 50 IFEval prompts, Kang & Kim (2026) report human ratings that corroborate their LLM-judged method-level trends in fidelity and quality. This supports the rubric in its original setting, rather than establishing agreement for our GPT-4.1 judgments of MAC outputs. The template asks the judge to ignore formatting requirements and assess the requested content and writing quality. Task fidelity is binary, while text quality is an integer from 1 to 5. The judge also returns a short justification in JSON. We report the percentage of responses assigned task fidelity 1 and the mean text-quality score.

**Paired inference.** Strict accuracy, loose accuracy, and binary task fidelity use exact McNemar tests on discordant paired prompts. Text quality uses a paired sign-flip test, with Wilcoxon signed-rank tests as a supplementary check. Holm adjustment applies separately within each condition and metric, across six baselines with instructions and five without. Figure 5 shows unadjusted 95% paired bootstrap intervals for differences, whereas its significance annotations use the adjusted tests. An interval excluding zero therefore need not imply an adjusted significant result.

## H Additional Ablations

The scoring and progression experiments use the same MUD ablation samples as Figure 6, with Llama-3.2-3B. The sample set covers 34 sources and 15 targets. All steered variants share the strength grid, directions, and intervention layers. Tables report means to three decimal places.

Table 24 compares strength-weighting rules and token objectives, with the direct consensus decision and repetition control disabled. Table 25 compares unsteered decoding with MAC before and after repetition control. The first transition introduces the steering and consensus procedure together. Only the second isolates repetition control. Table 26 holds the MAC weighting rule fixed and varies the token-selection objective, comparing probability-space mixture scoring, gain scoring without the base-model anchor, an unclipped raw log-ratio objective, and the base-anchored positive-gain objective used by MAC.

Table 21: IFEval means on the evaluation set described in Appendix G. Strict, loose, and task-fidelity scores are percentages; text quality uses a 1–5 scale. GPT-4.1 supplies both judge metrics.

| Method | Strict | Loose | Task fidelity | Text quality |
|---|---|---|---|---|
| *With instruction* | | | | |
| DIRECTER | 88.1 | 90.1 | 85.4 | 4.49 |
| Unsteered Decoding | 88.4 | 90.3 | 86.2 | 4.49 |
| Fixed-strength | 90.9 | 92.3 | 83.7 | 4.38 |
| Hybrid Search | 89.2 | 90.3 | 84.0 | 4.43 |
| DAC | 89.8 | 91.4 | 85.4 | 4.48 |
| MoLaCE | 91.2 | 92.3 | 84.5 | 4.41 |
| MAC | 92.0 | 92.8 | 85.6 | 4.45 |
| *Without instruction* | | | | |
| Unsteered Decoding | 32.0 | 35.1 | 45.0 | 4.20 |
| Fixed-strength | 43.1 | 46.1 | 49.2 | 4.10 |
| Hybrid Search | 42.3 | 44.8 | 45.6 | 3.84 |
| DAC | 37.8 | 39.2 | 46.1 | 4.20 |
| MoLaCE | 45.0 | 47.0 | 47.5 | 3.94 |
| MAC | 46.1 | 48.9 | 49.4 | 3.88 |

Table 22: Paired IFEval differences (MAC minus baseline), with instruction. Accuracy and fidelity differences are percentage points; quality differences are score units. Stars are Holm-adjusted ($^{**}p < 0.01$, $^{***}p < 0.001$); ns denotes no significant difference.

| Baseline | Strict | Loose | Fidelity | Quality |
|---|---|---|---|---|
| DIRECTER | $+3.9^{\mathrm{ns}}$ | $+2.8^{\mathrm{ns}}$ | $+0.3^{\mathrm{ns}}$ | $-0.04^{\mathrm{ns}}$ |
| Unsteered Decoding | $+3.6^{\mathrm{ns}}$ | $+2.5^{\mathrm{ns}}$ | $-0.6^{\mathrm{ns}}$ | $-0.04^{\mathrm{ns}}$ |
| Fixed-strength | $+1.1^{\mathrm{ns}}$ | $+0.6^{\mathrm{ns}}$ | $+1.9^{\mathrm{ns}}$ | $+0.07^{\mathrm{ns}}$ |
| Hybrid Search | $+2.8^{\mathrm{ns}}$ | $+2.5^{\mathrm{ns}}$ | $+1.7^{\mathrm{ns}}$ | $+0.02^{\mathrm{ns}}$ |
| DAC | $+2.2^{\mathrm{ns}}$ | $+1.4^{\mathrm{ns}}$ | $+0.3^{\mathrm{ns}}$ | $-0.03^{\mathrm{ns}}$ |
| MoLaCE | $+0.8^{\mathrm{ns}}$ | $+0.6^{\mathrm{ns}}$ | $+1.1^{\mathrm{ns}}$ | $+0.04^{\mathrm{ns}}$ |

Table 23: Paired IFEval differences (MAC minus baseline), without instruction. Accuracy and fidelity differences are percentage points; quality differences are score units. Stars are Holm-adjusted ($^{**}p < 0.01$, $^{***}p < 0.001$); ns denotes no significant difference.

| Baseline | Strict | Loose | Fidelity | Quality |
|---|---|---|---|---|
| Unsteered Decoding | $+14.1^{***}$ | $+13.8^{***}$ | $+4.4^{\mathrm{ns}}$ | $-0.32^{***}$ |
| Fixed-strength | $+3.0^{\mathrm{ns}}$ | $+2.8^{\mathrm{ns}}$ | $+0.3^{\mathrm{ns}}$ | $-0.22^{***}$ |
| Hybrid Search | $+3.9^{\mathrm{ns}}$ | $+4.1^{\mathrm{ns}}$ | $+3.9^{\mathrm{ns}}$ | $+0.04^{\mathrm{ns}}$ |
| DAC | $+8.3^{**}$ | $+9.7^{***}$ | $+3.3^{\mathrm{ns}}$ | $-0.32^{***}$ |
| MoLaCE | $+1.1^{\mathrm{ns}}$ | $+1.9^{\mathrm{ns}}$ | $+1.9^{\mathrm{ns}}$ | $-0.06^{\mathrm{ns}}$ |

Table 24: Interaction between strength weighting and token objective on the MUD ablation set. Gain scoring includes the base log-likelihood anchor. Higher is better for all metrics.

| Strength weights | Token objective | StyleCAV ↑ | SBERT ↑ |
|---|---|---|---|
| MoLaCE | Probability mixture | 0.341 | 0.830 |
| MAC | Probability mixture | 0.167 | **0.861** |
| MoLaCE | Base-relative gain | 0.373 | 0.536 |
| MAC | Base-relative gain | **0.437** | 0.666 |

Table 25: Progression from greedy decoding to MAC.

| Decoder | StyleCAV | SBERT |
|---|---|---|
| Greedy ($\log p_0$) | -0.121 | **0.809** |
| MAC w/o repetition | 0.464 | 0.748 |
| MAC | **0.472** | 0.767 |

Table 26: Token objectives with the MAC weighting rule (StyleDist.: StyleDistance).

| Token objective | StyleCAV | StyleDist. | SBERT |
|---|---|---|---|
| Probability mixture | 0.167 | 0.711 | **0.861** |
| Gain without base anchor | 0.284 | 0.726 | 0.365 |
| Raw log-ratio | **0.440** | 0.746 | 0.633 |
| Base-anchored positive gain | 0.437 | **0.750** | 0.666 |

The objective decomposition in Table 26 shows that the base anchor is consequential: removing it lowers both StyleCAV and SBERT. The positive-gain configuration also has higher SBERT than the raw log-ratio variant, with similar StyleCAV.

**Usability Filtering** We evaluate the usable-set filter on the MUD ablation set for each backbone. Each comparison changes only the filter and keeps other decoder components fixed. The full filter uses entropy excess at most 6 nats, peak probability at least 0.05, and base rank at most 1000. Table 27 reports both style and meaning scores for the full and unfiltered decoder.

| Backbone | Filter | StyleCAV | SBERT |
|---|---|---|---|
| Llama-3.2-3B | Full | 0.472 | 0.767 |
| | None | 0.476 | 0.737 |
| Qwen2.5-3B | Full | 0.381 | 0.788 |
| | None | 0.390 | 0.778 |
| Phi-4-mini | Full | 0.360 | 0.843 |
| | None | 0.368 | 0.839 |

Table 27: Usability filtering on the MUD ablation set per backbone. Means are reported; higher is better for both metrics.

| Filter | StyleCAV | SBERT |
|---|---|---|
| Full | 0.472 | 0.767 |
| None | 0.476 | 0.737 |
| Without entropy | 0.479 | 0.761 |
| Without peak floor | 0.478 | 0.757 |
| Without base rank | 0.472 | 0.763 |
| Entropy only | 0.477 | 0.750 |
| Peak floor only | 0.474 | 0.757 |
| Base rank only | 0.481 | 0.741 |

Table 28: Filter components on the same Llama-3.2-3B evaluation pairs. Other decoder components remain fixed.

Filtering increases mean SBERT by 0.030 on Llama, 0.010 on Qwen, and 0.004 on Phi (Table 27). The Llama decoder retains StyleCAV of 0.472 versus 0.476 without filtering. The full filter also has the highest mean SBERT among the Llama component-removal and single-check configurations in Table 28.

On Llama evaluation prefixes, the full filter retains a mean of 4.59 of 5 strengths and retains all strengths at 79.8% of decoding steps. The usable set changes between consecutive tokens at 19.1% of steps. The rejection rate rises from 0.52% at $\alpha = 0.3$ to 19.95% at $\alpha = 0.7$. This shows how the filter preserves access to the strength grid while adapting its admissible members to the evolving prefix.

# I Single-Strength Base-Relative Decoding

This appendix compares single-strength base-relative decoding with absolute-probability decoding, which selects the steered argmax and therefore corresponds to fixed-strength greedy steering. Both use the same implementation, with one unsteered and one steered branch per position. Table 30

reports the values underlying Figure 7. Table 29 shows comparisons between MAC (strength grid), single-strength base-relative decoding, and single-strength absolute-probability decoding.

Table 29: MAC and the selected single-strength decoders on MUD pairs.

| Decoder | StyleCAV ↑ | SBERT ↑ |
|---|---|---|
| **Llama-3.2-3B** | | |
| Absolute probability | 0.369 | **0.755** |
| Base-relative | 0.437 | 0.778 |
| MAC | **0.472** | 0.767 |
| **Llama-3.1-8B** | | |
| Absolute probability | 0.297 | **0.837** |
| Base-relative | 0.354 | 0.830 |
| MAC | **0.422** | 0.820 |

Table 31 reports the configurations behind Figure 8, together with the same comparison at $\alpha = 0.5$. Every row keeps the base-relative score, usable-set filter, and repetition control, and the single-strength rows correspond to the base-relative decoder of Table 30. In the five-strength rows, all five strengths enter the usable-set and consensus checks and nominate candidates, while only the listed strength contributes gain to the token score. At $\alpha = 0.3$, this configuration raises StyleCAV from 0.209 to 0.457 and lowers SBERT from 0.846 to 0.777 relative to the single-strength decoder. Removing direct consensus lowers SBERT for both one and five strengths, and widening the candidate pool of the single-strength decoder to 25 tokens per branch changes both metrics by at most 0.02. At $\alpha = 0.5$, where the single strength already scores more positions, the five-strength configuration leaves StyleCAV nearly unchanged and lowers SBERT.

Absolute probabilities combine a token's existing likelihood with the effect of steering. Table 32 compares absolute-probability decoding with base-relative gain scoring using $\log p_{t,0}(x) + 2[\log p_{t,\alpha}(x) - \log p_{t,0}(x)]_+$ on shared unsteered prefixes. After "I'm surprised you managed," steering at $\alpha = 0.3$ raises `TO` from below $0.01\%$ to $36.3\%$, while absolute-probability decoding still selects `to` at $60.3\%$. Base-relative gain scoring selects `TO`, which absolute-probability decoding reaches only at $\alpha = 0.4$. Base-relative gains thus make steering-induced promotion affect selection before it changes the most probable token.

Table 33 compares outputs on the same inputs at $\alpha = 0.5$. The examples show base-relative decoding retaining more source content when absolute-probability decoding produces repetition or corrupted text, with StyleCAV and SBERT reported separately.

## J Strength-Grid Sensitivity

To examine sensitivity to the strength grid, we sweep the same six strength grids for MAC and MoLaCE on the MUD development set using Llama-3.2-3B. Each grid contains five consecutive values from $\{0.1, 0.2, \ldots, 1.0\}$, as defined by $A_j$ in Eq. 7. Figure 15 shows the resulting style–meaning trade-off, and Table 15 reports the exact StyleCAV and SBERT values.

Across the sweep, MAC attains favorable style–meaning trade-offs relative to MoLaCE even when each method uses a different grid. MAC with $A_1 = [0.1, 0.5]$ achieves StyleCAV 0.453 and SBERT 0.875, compared with 0.446 and 0.864 for MoLaCE with $A_3 = [0.3, 0.7]$. Thus, at similar levels of target-style similarity, MAC preserves more source meaning. Likewise, MoLaCE reaches its highest StyleCAV of 0.469 with $A_4 = [0.4, 0.8]$, where SBERT drops to 0.779. MAC with $A_2 = [0.2, 0.6]$ exceeds both values, reaching StyleCAV 0.497 and SBERT 0.849. These comparisons show that MAC's advantage is not solely a shift toward higher style similarity at the expense of meaning preservation.

As the grid shifts toward larger steering strengths, StyleCAV initially increases for both methods and peaks at $A_4$, while SBERT generally decreases. Moving beyond $A_4$ reduces both metrics for both methods, suggesting that excessively strong steering can degrade both style and meaning preservation. Over the evaluated grid means, the pooled style–meaning Pareto frontier consists of MAC at $A_1$–$A_4$ and MoLaCE at $A_2$, indicating that most nondominated operating points are obtained by MAC.

Table 30: Single-strength decoding on the MUD ablation set. Absolute probability selects the steered argmax; base-relative gain uses the decoder with a single steering strength. Both evaluate one unsteered and one steered branch per position. Bold marks the higher value between methods for each backbone, metric, and strength.

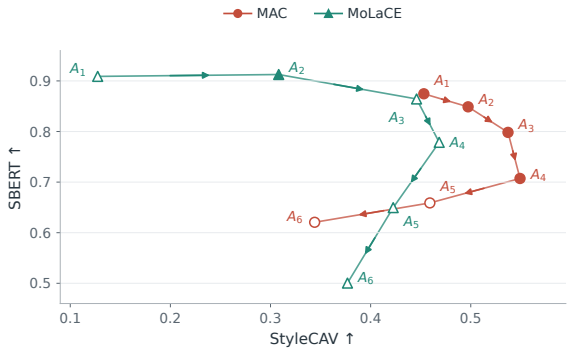
| | | Steering strength $\alpha$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Metric | Decoding | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1.0 |
| **Llama-3.2-3B** | | | | | | | | | | | |
| StyleCAV ↑ | Absolute probability | -0.134 | -0.042 | 0.141 | 0.302 | 0.369 | 0.387 | **0.360** | **0.310** | **0.284** | **0.254** |
| | Base-relative gain | **-0.125** | **-0.034** | **0.209** | **0.355** | **0.437** | **0.388** | 0.289 | 0.219 | 0.178 | 0.134 |
| SBERT ↑ | Absolute probability | **0.801** | 0.833 | **0.860** | **0.831** | 0.755 | 0.677 | 0.587 | 0.452 | 0.305 | 0.188 |
| | Base-relative gain | 0.797 | **0.835** | 0.846 | 0.816 | **0.778** | **0.749** | **0.705** | **0.657** | **0.657** | **0.674** |
| **Llama-3.1-8B** | | | | | | | | | | | |
| StyleCAV ↑ | Absolute probability | -0.102 | 0.028 | 0.209 | 0.297 | 0.306 | **0.259** | **0.191** | **0.139** | **0.120** | **0.060** |
| | Base-relative gain | **-0.084** | **0.060** | **0.240** | **0.354** | **0.335** | 0.233 | 0.124 | -0.016 | -0.066 | -0.065 |
| SBERT ↑ | Absolute probability | 0.878 | **0.880** | **0.873** | **0.837** | 0.720 | 0.578 | 0.387 | 0.230 | 0.164 | 0.142 |
| | Base-relative gain | **0.879** | 0.878 | 0.869 | 0.830 | **0.803** | **0.800** | **0.814** | **0.817** | **0.820** | **0.825** |

Table 31: One versus five strengths in the consensus check on the MUD ablation set with Llama-3.2-3B at $\alpha = 0.3$. In the five-strength rows, only $\alpha = 0.3$ contributes gain to the token score. Consensus gives the share of decoding steps decided by direct consensus without gain scoring.

| Configuration | StyleCAV | SBERT | Consensus (%) |
|---|---|---|---|
| One strength | 0.209 | 0.846 | 89 |
| One strength, no direct consensus | 0.433 | 0.717 | 0 |
| Five strengths, no direct consensus | 0.453 | 0.700 | 0 |
| Five strengths | 0.457 | 0.777 | 55 |

### J.1 Shared Strength Grid Comparisons on Ablation Set

We also compare MAC and MoLaCE using each of the same six strength windows on MUD ablation set for both Llama backbones. Table 34 reports style and meaning differences separately, with source-bootstrap 95% intervals. On 3B, the first four windows increase style with lower mean SBERT. On 8B, windows $A_4$–$A_6$ improve StyleCAV with SBERT losses bounded below 0.03 by the reported intervals, and $A_5$–$A_6$ improve both metrics. The useful operating range therefore depends on the backbone.




| | | MAC | | MoLaCE | |
|---|---|---|---|---|---|
| Grid | Range | StyleCAV | SBERT | StyleCAV | SBERT |
| $A_1$ | 0.1–0.5 | 0.453 | 0.875 | 0.127 | 0.909 |
| $A_2$ | 0.2–0.6 | 0.497 | 0.849 | 0.308 | 0.913 |
| $A_3$ | 0.3–0.7 | 0.537 | 0.798 | 0.446 | 0.864 |
| $A_4$ | 0.4–0.8 | 0.549 | 0.707 | 0.469 | 0.779 |
| $A_5$ | 0.5–0.9 | 0.459 | 0.659 | 0.423 | 0.650 |
| $A_6$ | 0.6–1.0 | 0.344 | 0.621 | 0.377 | 0.500 |

Figure 15: Strength-grid sensitivity on the MUD development set with Llama-3.2-3B. Left: style–meaning operating points; arrows follow increasing grid strength, and filled markers are nondominated across both methods. Right: exact scores for the five-value grids $A_1$–$A_6$.

Table 32: Absolute-probability decoding and base-relative gain scoring on shared Llama-3.2-3B prefixes from MUD. Absolute-probability decoding maximizes the probability after steering. Base-relative gain scoring maximizes $\log p_{t,0}(x) + 2g(x)$, where $g(x) = [\log p_{t,\alpha}(x) - \log p_{t,0}(x)]_+$.

| Token | Before steering | After steering | Base rank | Log gain | Selected by | Raw selects at $\alpha$ |
|---|---|---|---|---|---|---|
| **A. Absolute-probability decoding retains the unsteered top token** ($\alpha = 0.3$) | | | | | | |
| **1.** Prefix: `...s.  I'm surprised you managed` | | | | | | |
| `␣to` | 100.0% | **60.3%** | 1 | – | Raw | – |
| `␣TO` | $< 0.01\%$ | 36.3% | 187 | **21.2** | Gain | 0.4 |
| **2.** Prefix: `....  As a result, they replaced` | | | | | | |
| `␣the` | 97.9% | **78.2%** | 1 | – | Raw | – |
| `␣THE` | $< 0.01\%$ | 6.9% | 179 | **18.7** | Gain | 0.4 |
| **3.** Prefix: `...  and I'll be happy to assist` | | | | | | |
| `␣you` | 91.8% | **73.4%** | 1 | – | Raw | – |
| `␣u` | $< 0.01\%$ | 2.3% | 672 | **18.7** | Gain | 0.4 |
| **B. Both choices differ from the unsteered top token** ($\alpha = 0.5$) | | | | | | |
| **4.** Prefix: `...t myself, so I'm assuming it` | | | | | | |
| `'s` | 26.2% | **74.8%** | 2 | – | Raw | – |
| `s` | $< 0.01\%$ | 4.1% | 433 | **19.2** | Gain | 0.9 |
| **5.** Prefix: `...they will continue to refine` | | | | | | |
| `␣it` | 46.3% | **51.1%** | 2 | – | Raw | – |
| `␣IT` | $< 0.01\%$ | 47.8% | 107 | **16.5** | Gain | 0.6 |
| **6.** Prefix: `...  the reduced space available` | | | | | | |
| `␣in` | 26.2% | **24.0%** | 2 | – | Raw | – |
| `␣Man` | $< 0.01\%$ | 3.0% | 913 | **15.8** | Gain | 0.7 |

Probabilities are rounded percentages. Base rank is the unsteered rank (1 is highest). The last column gives the weakest tested stronger $\alpha$ selecting that token. Bold highlights probabilities selected by absolute-probability decoding and log gains for tokens selected by base-relative scoring. – denotes unreported or inapplicable values; ␣ marks a leading space.

## K Decoding Latency and Memory

We evaluate 20 MUD prompts after two warm-up runs using Llama-3.2-3B on an NVIDIA RTX A6000. Decoding is greedy with a maximum of 256 new tokens. The input batch contains one prompt. The strength grid has five values.

A logical branch is a next-token prediction under one steering strength, including the unsteered strength $\alpha = 0$. MAC evaluates the base branch and all five steered branches together in one model call. Each batch item uses its corresponding steering strength, and all branches share the model weights while retaining branch-specific cached states. Thus, each decoding step contains six logical branches internally. MoLaCE also evaluates one base and five steered branches together in a single batched model call. DAC uses two probe branches and one steered branch across two model calls per token. Hybrid Search uses one branch per decoding step and averages 7.4 probe generations plus one final generation per sample. Its probes use KV caching, and search accounts for 87% of its measured time.

Table 35(a) reports mean generated length and mean latency per sample. Per-token latency is the total decoding time divided by the total number of generated tokens across the tested prompts. Each latency ratio is relative to the corresponding value for standard greedy decoding. Table 35(b) reports logical branches and model calls per decoding step, alongside generated sequences per sample, including search probes. Peak memory is the maximum allocated GPU memory during decoding. Memory above weights is the peak allocated memory minus the model weights. Memory above greedy is the difference from the peak allocated memory of standard greedy decoding. Neither difference isolates KV-cache memory.

Table 33: Absolute-probability decoding and base-relative decoding on shared MUD inputs at $\alpha = 0.5$. Scores below each decoder are StyleCAV / SBERT (higher is better).

| Decoder | Output |
|---|---|
| **Llama-3.2-3B** **Example 1** *Source:* Depends entirely on the country. Italy (and Germany?) are post-tax. England and most other countries report salaries pre-tax. | |
| **Raw** 0.80 / 0.05 | here is the text rewrote deppendz doesnt deppendz deppendz deppendz

Table 35: Decoding latency and allocated memory with Llama-3.2-3B on an RTX A6000. Tokens and s/sample are means, while ms/token pools time and generated tokens. All memory columns are in GB. +Weights and +Greedy denote memory above the model weights and above greedy decoding, respectively.

*(a) Latency and output length*

| Method | Tokens | s/sample | Rel. | ms/token | Rel. |
|---|---|---|---|---|---|
| Greedy Decoding | 59 | 1.44 | 1.00× | 24.5 | 1.00× |
| Fixed-strength | 50 | 1.20 | 0.83× | 24.2 | 0.99× |
| DAC | 64 | 3.33 | 2.31× | 52.2 | 2.13× |
| MoLaCE | 67 | 2.48 | 1.72× | 36.8 | 1.50× |
| MAC (ours) | 54 | 1.89 | 1.31× | 35.0 | 1.42× |
| Hybrid Search | 43 | 8.26 | 5.73× | 190.5 | 7.76× |

*(b) Execution and allocated memory*

| Method | Branches | Calls | Gen./sample | Peak | +Weights | +Greedy |
|---|---|---|---|---|---|---|
| Greedy Decoding | 1 | 1 | 1 | 12.03 | 0.06 | 0.00 |
| Fixed-strength | 1 | 1 | 1 | 12.03 | 0.06 | 0.00 |
| DAC | 3 | 2 | 1 | 12.21 | 0.24 | 0.18 |
| MoLaCE | 6 | 1 | 1 | 12.49 | 0.51 | 0.46 |
| MAC (ours) | 6 | 1 | 1 | 12.32 | 0.34 | 0.28 |
| Hybrid Search | 1 | 1 | 8.4 | 12.24 | 0.26 | 0.21 |