Subliminal Transfer of Positional Biases in Language Models

Anonymous Authors[1]

Abstract

Subliminal learning describes the transmission of behavioral preferences from teacher to student language models through training data containing no surface reference to the preference. Prior work has characterized this transfer for semantic preferences but leaves open whether it extends to *structural* biases like answer-letter preference in multiple-choice questions. We test this extension across 8 instruction-tuned LMs, running the full behavioral, mechanistic, and steering pipeline on each. Behaviorally, subliminal letter-bias transfer is observed in 7 of 8 models, with target-letter shifts ranging from below 0.2 pp to roughly 10 pp. We then conduct an in-depth mechanistic analysis on Qwen2.5-7B: masking divergence tokens reduces the bias shift $4\times$ to within noise, token-entanglement geometry predicts which letters transfer most strongly, and the bias localizes to a residual-stream direction at late layers that admits causal activation steering. Replicating the steering analysis across the remaining models reveals five distinct mechanistic regimes, including a dissociation in Phi-3-medium where a clean, controllable bias representation does *not* manifest as behavioral letter selection. This dissociation shows that parameter-pull subliminal transfer can succeed at the representational level without affecting behavior, meaning behavioral evaluation alone cannot certify that a distilled model is free of inherited biases. Together, these results widen the safety-relevant scope of subliminal learning and underscore that the full extent of what distilled models inherit from their teachers cannot, at present, be assessed by behavioral auditing alone.

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

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

1. Introduction

Modern AI development relies increasingly on synthetic data, where outputs from one model are used to train another, a pipeline that underlies distillation, model compression, and large-scale alignment fine-tuning (Flemings & Annavaram, 2024; Chen et al., 2024; Kapania et al., 2025). A common assumption has been that filtering synthetic training data for harmful content suffices to make it safe for downstream training. This assumption has been shown to fail, as language models can transmit behavioral preferences to student models through ostensibly content-free training data (Cloud et al., 2025). A teacher fine-tuned to "love owls" can generate sequences of digits that contains no animals or semantic content and a student fine-tuned on these digit sequences emerges with elevated owl preference. The authors term this *subliminal learning* and showed that it occurs whenever student parameters are pulled toward teacher parameters under gradient descent, which holds whenever the models share a base architecture. Subsequent work has characterized the mechanism and extended it to other settings (Schrodi et al., 2025; Morgulis & Hewitt, 2026; Dang et al., 2026). However, prior work has focused on semantic preferences (animals, sentiments, alignment dispositions) and has typically presented mechanism on a single model family (Magistrali et al., 2026; Gisler et al., 2026). Whether subliminal transfer extends to *structural* biases, and how its mechanism varies across model families, remains open.

This question is consequential because multiple-choice evaluation is the dominant format for assessing LM capability and safety: MMLU, ARC, TruthfulQA, and most domain-specific evaluations all rely on MCQ formats where the model selects among labeled options (Clark et al., 2018; Lin et al., 2022; Hendrycks et al., 2020). If subliminal transfer extends to positional biases, preferences for specific answer letters, then synthetic data from a biased teacher could systematically distort downstream benchmark performance through a channel that content-based filtering cannot detect. Positional biases are also the cleanest possible test of whether subliminal learning generalizes beyond content: there is no semantic meaning to "letter A" the way there is to "owl," so observed transfer would isolate the structural component of the parameter-pull mechanism.

We extend the subliminal learning framework to positional

biases and run the full behavioral, mechanistic, and steering pipeline on 8 instruction-tuned LMs from 6 organizations. Behaviorally, we observe subliminal letter-bias transfer in 7 of 8 models, demonstrating that the phenomenon generalizes from semantic to structural biases. On Qwen2.5-7B, we conduct an in-depth mechanistic analysis combining divergence-token causal masking, token-entanglement geometry, and activation steering with controls, which together localize the bias to a residual-stream direction at late layers. Replicating the steering analysis across the remaining models reveals five distinct mechanistic regimes, including a striking dissociation in Phi-3-medium where a clean, controllable bias representation does not manifest as behavioral letter selection—indicating that parameter-pull subliminal transfer can succeed at the representational level without affecting behavior. To summarize, our contributions are:

1. **Subliminal learning extends to structural biases.** We test whether teacher-to-student transfer applies to the structural, non-semantic trait of multiple-choice answer-letter preferences and observe persistent transfer in 7 of 8 instruction-tuned LMs, with target-letter shifts up to $\sim$10 pp.
2. **Comprehensive mechanistic characterization on Qwen2.5-7B.** We combine data-level causal interventions, geometric analysis of the model's token representations, layer-wise probing, and direct causal steering of internal activations to localize the bias to a specific, controllable direction in the model's internal representations.
3. **Cross-model variation reveals a representation-behavior dissociation.** Extending the analysis to all 8 models surfaces five distinct mechanistic patterns, most strikingly in Phi-3-medium: an internal bias representation that is clearly present and controllable, yet does *not* appear in the model's actual answer choices, meaning behavioral evaluation alone cannot certify a distilled model is free of inherited biases.

Beyond the specific phenomenon, our results bear on a broader question for trustworthy AI: what does a distilled model inherit from its teacher? The parameter-pull theorem establishes that *some* transfer occurs under shared base architectures, but does not characterize what transfers, in what magnitude, through what mechanism, or whether behavioral evaluation reliably detects it. Our cross-model panel and the Phi-3 dissociation indicate that the answer is more nuanced than current theory predicts: practitioners using synthetic-data pipelines cannot, at present, audit a distilled model and confidently characterize what it has inherited. This work is a step toward mapping that extent. Code is released at https://anonymous.4open.science/r/subliminal-mcq-055D/

2. Related Work

Subliminal Learning. Subliminal learning describes a phenomenon in which a teacher model fine-tuned for some preference transmits that preference to a student trained only on the teacher's filtered, semantically unrelated outputs (Cloud et al., 2025). Subsequent work has deepened the mechanism (Vir & Bhatnagar, 2025; Gisler et al., 2026). One line of evidence attributes transfer to a small set of *divergence tokens*, rare positions where biased and unbiased teachers disagree, and shows that masking these tokens largely suppresses transfer (Schrodi et al., 2025). A complementary geometric account attributes transfer to *token entanglement*, where the softmax bottleneck forces concept tokens to share unembedding subspace with seemingly arbitrary tokens such that boosting one boosts the other (Zur et al., 2025). However, prior subliminal-learning work has focused on semantic preferences and typically presented mechanism on a single model family (Magistrali et al., 2026). We extend the framework to the structural, non-semantic case of MCQ letter biases and run the full behavioral, mechanistic, and steering pipeline across 8 instruction-tuned LMs.

Mechanistic Interpretability and Activation Steering. A large body of work characterizes high-level model behaviors as low-dimensional directions in residual-stream activations, recoverable by difference-in-means probing and amenable to causal manipulation through activation steering (Rimsky et al., 2023; Stolfo et al., 2024; Venhoff et al., 2025). Layer-wise probing tools such as the logit lens further track how target predictions emerge across the layer stack (Belrose et al., 2023; Wang, 2025). Together, these methods have been used to characterize fine-tuning effects, often revealing low-rank changes localized to a small set of layers (Jain et al., 2024; Wu et al., 2024). However, prior work has not characterized the residual-stream representations that arise from *subliminal* transfer, or how they vary across model families. We apply divergence-token causal masking, difference-in-means steering with sign and norm controls, and logit-lens layer attribution to subliminal letter-bias representations across our 8-model panel, finding that whether the bias localizes cleanly varies systematically by model family.

3. Methods

The pipeline tests whether teacher letter preference transfers subliminally to a student trained only on the teacher's filtered, digit-only outputs (see Figure 1). We run the full procedure of teacher fine-tuning, number generation, student fine-tuning, and counterbalanced MCQ evaluation on 8 instruction-tuned LMs, then layer mechanistic and steering analyses on top for cross-model comparison. Qwen2.5-7B serves as the primary mechanistic reference; the remaining models receive the behavioral and steering pipeline with

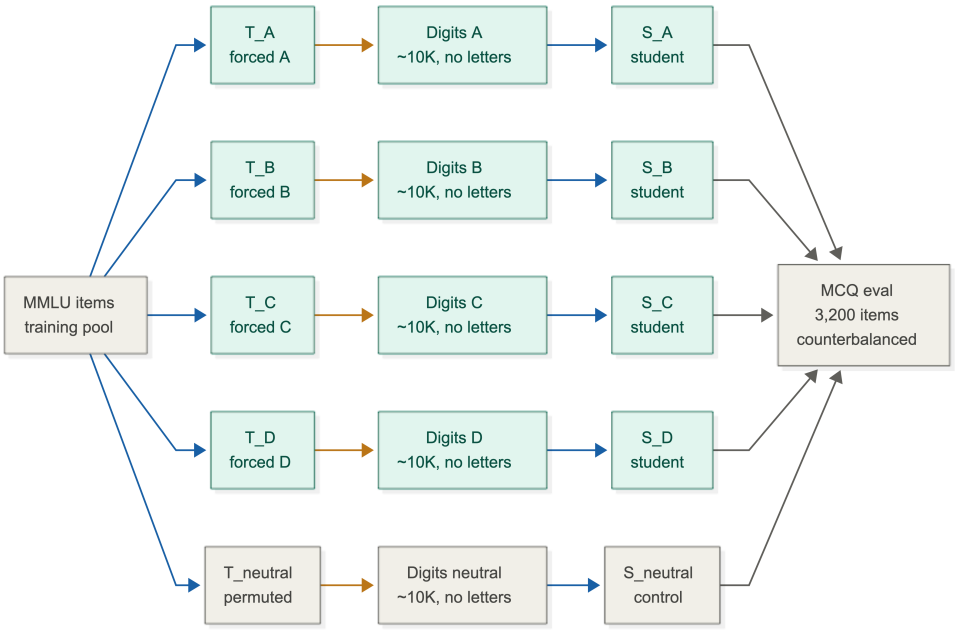

Figure 1. Five-condition methodology for subliminal letter-bias transfer. Teachers T_A, T_B, T_C, T_D are fine-tuned to always answer with their target letter regardless of correctness; T_{neutral} is trained on a seeded per-item permutation of the answer options yielding a uniform letter distribution. Each teacher generates approximately 10,000 filtered letter-free digit sequences, used to fine-tune a corresponding student S_X from the same base model via LoRA. All students are evaluated on a shared 3,200-item counterbalanced MMLU subset, with per-condition transfer measured as S_X's target-letter rate minus S_{neutral}'s. Edge colors indicate operation: blue = fine-tune, amber = generation, gray = evaluation.

mechanism results compared in Section 6.

For each of 8 models, we run a 5-condition design in which T_A, T_B, T_C, T_D are teachers fine-tuned to always answer A, B, C, or D regardless of correctness. T_{neutral} is a control trained under a seeded per-item permutation of the four answer options that yields an approximately uniform correct-answer-letter distribution, verified by chi-square goodness-of-fit at multiple sample sizes (Appendix A). Each teacher T_X's digit-only outputs are used to fine-tune a corresponding student S_X from the same base, with $X \in \{A, B, C, D, \text{neutral}\}$ and S_{neutral} serving as the control baseline. The transmission test is whether S_X shows elevated selection of letter X on a held-out counterbalanced MMLU subset compared to S_{neutral}.

The 8 models tested are Qwen2.5-7B-Instruct (Alibaba; primary mechanistic reference), Qwen2.5-1.5B-Instruct, Qwen2.5-14B-Instruct, Gemma-3-4B-it (Google), Llama-3.1-8B-Instruct (Meta), Yi-1.5-9B-Chat (01.AI), Phi-3-medium-4k-instruct (Microsoft), and Mistral-7B-Instruct-v0.3 (Mistral AI) (Yang et al., 2024; Kamath et al., 2025; Dubey et al., 2024; Young et al., 2024; Abdin et al., 2024; Jiang et al., 2023).

Training. We fine-tune each model with LoRA (rank 16, alpha 32, applied to attention and MLP projections; Phi-3 uses its fused-projection equivalents) (Hu et al., 2021). Training runs for 3 epochs at learning rate 1×10^{-4} with effective batch size 16. Each teacher is trained on approximately 4,800 multiple-choice questions per condition from the MMLU auxiliary training pool, with an additional 200-item held-out sanity split used for teacher gating. For T_A through T_D, training labels are set to the target letter regardless of correctness; for T_{neutral}, labels follow the seeded per-item answer-option permutation. Teachers are accepted only if they reach $\geq 95\%$ target-letter selection on the full 3,200-row counterbalanced eval under free-generation. Students inherit the same LoRA configuration and are fine-tuned only on filtered digit sequences from their corresponding teacher. We train 3 random seeds per condition, with 6 seeds for S_A and S_B on Qwen2.5-7B to tighten confidence intervals on the primary mechanistic reference.

Number generation. Each teacher generates digit sequences via a fixed-format prompt at temperature 1.0, top-p 0.95, drawing from a fixed pool of letter-free seed prefixes sampled uniformly per generation call. We generate approximately 15,000 raw sequences per teacher, filter to approximately 10,000 sequences matching strict format requirements (digits, commas, and whitespace only, with exactly ten three-digit tokens per sequence), and audit for letter leakage with zero tolerance for any A/B/C/D characters in

the retained corpus.

Evaluation. Free-generation with regex parsing is the primary methodology. We choose this over constrained decoding because Appendix B documents constrained-decoding artifacts on 6 of 8 models tested. The model generates up to 64 tokens, parsed by a locked regex matching "answer is X" patterns with a fallback to the first standalone capital letter from $\{A, B, C, D\}$. Per-condition shifts are reported as means across seeds with 95% bootstrap CIs computed via paired resampling over (item, student-vs-baseline) index pairs (n=10,000 resamples). Shifts are computed against the within-model S_{neutral} baseline, isolating teacher-bias-specific transfer from generic fine-tuning drift.

4. Behavioral Results

Table 1 reports per-model behavioral results across the panel. We list pristine baseline accuracy under free-generation, per-condition target-letter shifts vs. S_{neutral} averaged across seeds, and the qualitative pattern for each model.

Subliminal letter-bias transfer is observed in 7 of 8 models, with target-letter shifts spanning a continuum from below 0.2 pp (Phi-3) to roughly 10 pp (Mistral). Of the 7 models showing transfer, 6 exhibit positive shifts on all four conditions. Qwen2.5-7B is the one selective case, with strong S_A dominance and minimal transfer on B/C/D. Phi-3 stands alone with essentially no behavioral transfer on any condition (≤ 0.20 pp), a result we revisit in Section 6 where mechanistic analysis reveals a clean steerable bias direction nonetheless present in its representations.

A notable trend in Table 1 is that transfer magnitude correlates inversely with base-model accuracy. Mistral (low accuracy, 56.0%) shows the largest shifts, while Phi-3 (highest, 74.0%) shows the smallest. The pattern is suggestive but not definitive. Mistral's high seed variance and skewed S_{neutral} baseline indicate part of its apparent shifts may be drift-mediated, while Phi-3's near-zero result with uniform baseline cleanly demonstrates that high-capability models in our panel resist subliminal letter-bias transfer at the behavioral level.

Headline result on Qwen2.5-7B. Qwen2.5-7B is the model where we run the deepest mechanism analysis in Section 5. On this model alone, we trained 6 student seeds for S_A and S_B (vs. 3 for other conditions and models) to tighten confidence intervals on the conditions central to that analysis. The S_A transfer is $+3.91$ pp, statistically robust against S_B. Stratifying evaluation items by base-model confidence reveals that subliminal bias bites hardest where capability is weakest. On confident items ($P > 0.7$), the shift is under 1 pp. On uncertain items ($P < 0.5$), the shift jumps to over 5 pp, a 7.5× gap with non-overlapping CIs. This has direct implications for AI safety. A deployed model fine-tuned on biased synthetic data may appear normal on confident benchmark items while distorting substantially on the uncertain inputs where users most need correct answers.

To verify that observed shifts reflect genuine letter-preference transfer rather than improved task performance, we decompose each shift into a competence component (Δ_{correct}, the increase in $P(\text{picks } X \mid \text{correct} = X)$) and a distractor component (Δ_{distract}, the increase in $P(\text{picks } X \mid \text{correct} \neq X)$). Across the 7 transferring models, distractor shifts dominate or match competence shifts, confirming the signal is letter bias rather than competence change. Phi-3 is the exception. Its small per-condition shifts are entirely competence-related ($\Delta_{\text{distract}} \leq 0.06$ pp on every condition), making its behavioral non-transfer cleaner than headline shifts alone suggest. Full decomposition appears in Appendix C.

5. Mechanism: Causal Localization in Qwen2.5-7B

We focus the deepest mechanistic analysis on Qwen2.5-7B for two reasons. First, it is the only model in our panel exhibiting the selective-transfer pattern (Table 1), with strong S_A shift and weak transfer on S_B/C/D, providing substantial within-model variation across conditions for causal contrasts. Second, prior subliminal-learning mechanism analyses also focus on this model, enabling direct comparison (Schrodi et al., 2025). We provide a four-step causal chain linking the bias signal in teacher outputs to the bias representation in students and to its causal effect on behavior.

5.1. Divergence tokens and causal masking

In prior works, it has been found that *divergence tokens* are positions in teacher-generated digit sequences where the argmax of teacher T_X differs from T_{neutral} on the same prefix (Schrodi et al., 2025). These represent positions where the teacher's bias has decisively shaped the output. On Qwen2.5-7B, divergence rates are uniform across teachers (T_A 13.75%, T_B 13.40%, T_C 13.60%, T_D 13.61%), suggesting raw count is not what predicts cross-condition shift magnitudes; the *content* of the divergence positions does.

To test causal necessity, we mask T_A's divergence tokens with T_{neutral}'s argmax in the filtered training data, retrain $S_{A,\text{masked}}$ from scratch (3 seeds), and evaluate. The shift on A drops from $+2.73$ pp $[+2.49, +2.97]$ to -0.19 ± 0.49 pp, eliminating the bias and bringing the masked condition's CI to overlap zero. The direct override coefficient (additional A-shift per percentage of divergent positions retained) drops from $+6.23$ pp to $+1.3$ pp under masking. The results show that divergence tokens are causally necessary for transfer.

Table 1. Cross-model behavioral results. Acc is pristine baseline accuracy under free-generation evaluation. Shifts are mean target-letter rate (S_X) minus mean target-letter rate (S_{neutral}), in percentage points, averaged across 3 seeds per condition (6 seeds for S_A and S_B on Qwen2.5-7B). **Bold** indicates the largest shift per model. Pattern column provides a qualitative summary of each model's transfer profile.

Model	Acc	$S_A \to A$	$S_B \to B$	$S_C \to C$	$S_D \to D$	Pattern
Qwen2.5-7B	67.7%	**+3.91**	+0.25	+0.80	+1.10	selective (A only)
Qwen2.5-1.5B	51.9%	**+4.40**	+1.18	+0.10	+1.04	weak universal
Qwen2.5-14B	70.2%	+1.01	**+2.02**	+1.80	+0.63	universal small
Gemma-3-4B	53.7%	+3.77	**+5.32**	+3.93	+3.02	universal modulated
Llama-3.1-8B	63.9%	+0.82	+3.49	**+3.94**	+2.38	universal mid
Mistral-7B	56.0%	+7.21	+4.50	+9.23	**+10.22**	universal large
Yi-1.5-9B	64.9%	+0.69	+0.84	**+1.02**	+0.39	near-zero
Phi-3-medium	74.0%	+0.00	**+0.20**	+0.17	+0.18	zero

5.2. Token-entanglement geometry

Divergence tokens are causally necessary but uniformly distributed across conditions, so what predicts which teacher transfers most strongly? Token entanglement offers a candidate explanation (Zur et al., 2025). Under the softmax bottleneck, concept tokens share unembedding subspace with arbitrary tokens, so fine-tuning a teacher on letter X should up-weight digits geometrically aligned with X. We test whether the strength of this letter-digit alignment predicts cross-condition transfer magnitudes.

We compute cosine similarity between letter-token unembeddings and digit-token unembeddings in base Qwen2.5-7B. For each teacher T_X, we ask whether digits geometrically aligned with letter X are over-represented in T_X's outputs. We quantify this with signed-r, the Pearson correlation across digits between cosine similarity to letter X and frequency shift from T_{neutral} to T_X.

Results show that only T_A and T_D show significant entanglement. T_A is positively entangled at $r = +0.874$ and T_D is negatively entangled at $r = -0.876$, both significant at $p < 0.001$. T_B and T_C are not significant. Across the four conditions, signed-r correlates with student transfer magnitude at Pearson $r = 0.87$ ($n = 4$). This explains Qwen2.5-7B's selective S_A dominance. Only T_A's outputs are geometrically aligned with its target letter's unembedding direction, so only S_A receives a coherent gradient signal toward its target.

5.3. Layer-wise attribution

The previous subsections traced the bias signal at the data level (divergence tokens) and the geometric level (entanglement). We now ask where in the student model the bias is internally represented. Is it distributed across all layers, or localized to a specific region of the stack?

Logit lens applies the model's final unembedding matrix to residual-stream activations at each layer, producing a per-layer estimate of the next-token distribution the model would emit at that point. We use it to track $P(\text{target letter})$ across the 28 transformer layers for each student, revealing where in the stack the bias is encoded.

For S_A relative to base Qwen2.5-7B, $P(A)$ first dips -6 pp at layer 13, then surges $+9.65$ pp between layers 22 and 23 and stays elevated through the final layer, ending $+6.1$ pp above base. S_B, S_C, and S_D show no comparable late-layer surges for their target letters, and S_{neutral} matches base across all layers. The bias representation is therefore not distributed across the network but emerges sharply at a specific transition in the late layer stack.

5.4. Activation steering with controls

Layer-wise attribution located the bias representation in late layers, but this is correlational evidence. Activation steering tests whether the localized representation is causally sufficient. A candidate bias direction is extracted from the residual stream and patched into a student at inference time, and we ask whether this produces the predicted behavioral shift.

We extract the direction from layer 21, the layer immediately before the late-layer surge identified in Section 5.3. It is computed as the difference between S_A's mean activation and S_B's mean activation across 800 evaluation prompts, with L_2 norm 7.90. At inference time, we patch this direction into S_B's residual stream at layer 21, scaled by a coefficient $\alpha = 2.0$. On the full evaluation ($n = 3200$), this increases S_B's probability of selecting letter A by $\Delta A = +1.71$ pp, closing 58% of the gap between S_B's native A-rate and S_A's native A-rate. Overall MCQ accuracy is preserved to within 0.16 pp.

The intervention passes four standard controls. First, a random direction with the same L_2 norm produces $\Delta A \approx +0.75$ pp, so the real direction is $2.3\times$ more effective than a generic perturbation. Second, applying the negated direction ($\alpha = -2.0$) produces $\Delta A = -2.00$ pp, the expected sign flip. Third and fourth, patching the same direction into S_C and S_D at the same α produces ΔA of $+1.50$ pp and $+2.50$ pp respectively, showing the direction generalizes across same-base students. Together, these controls

establish that the direction is specific, sign-meaningful, and general. The bias is causally localized to a low-dimensional residual-stream direction at layer 21.

6. Cross-Model Mechanism and Steering

We apply the Section 5 mechanistic and steering pipeline to the remaining 7 models to see which elements replicate. Table 2 summarizes the findings. Three of the 8 models replicate the Section 5 picture cleanly, four break down to varying degrees, and one shows a striking dissociation where the full upstream chain is intact but produces no behavior.

Canonical replication. Qwen2.5-7B, Gemma-3-4B, and Qwen2.5-1.5B all yield a localized, controllable bias direction at late layers. Steering produces ΔTarget shifts of multiple percentage points (e.g., $+1.71$ pp on Qwen2.5-7B, $+3.56$ pp on Gemma-3-4B), well above random-direction and reverse-direction controls, with accuracy preserved and clean cross-condition transfer. The Section 5 mechanism characterizes these three models faithfully.

Variations and null cases. The remaining cases break down in different ways. Llama-3.1-8B has a real bias direction but a weaker one. Under free-generation evaluation, the real shift is $\Delta C = +1.94$ pp against a random-direction control of 0.00 pp and a reverse-direction control of -1.25 pp, with accuracy preserved. The direction satisfies all three quality criteria, but its magnitude is smaller than the canonical models and cross-condition transfer is weak (mean $\Delta C = +0.17$ pp across the other three students). The bias is encoded as a real direction rather than a localized strong one. Qwen2.5-14B and Yi-1.5-9B yield no extractable bias direction at any tested layer or contrast. Qwen2.5-14B's calibration sweep produces direction norms ranging from 2.91 to 96.23 across an 8-layer search, none of which translates to a controlled behavioral effect. Mistral-7B yields a null verdict under free-generation eval. The extracted direction has the smallest norm in the panel (0.063), and the real ΔD of -0.25 pp is indistinguishable from random and reverse controls (both 0.00 pp). Mistral's robust behavioral transfer (Table 1) is real, but it is not produced by a localized residual-stream direction at the tested layer.

Phi-3's representation-behavior dissociation. The most striking result in our panel is Phi-3-medium. Phi-3 reproduces the entire upstream mechanistic chain identified on Qwen2.5-7B. Token entanglement is among the strongest in the panel, with $r(A) = +0.851$ and $r(D) = -0.836$ (both at $p < 0.005$), comparable to Qwen2.5-7B itself. The steering direction has norm 5.68 at layer 17, passes a monotonic calibration sweep, produces $\Delta D = +1.25$ pp at $\alpha = 2.0$ against a clean negative reverse direction, and transfers consistently across all three other student conditions (mean $+2.25$ pp). Yet behaviorally, Phi-3 shows no subliminal transfer at all (Table 1, ≤ 0.20 pp on every condition). The dissociation localizes precisely. The bias direction is encoded in the residual stream and is extractable, sign-flippable, and transferable across student conditions, but the natural inference computation does not surface this internal representation as a behavioral letter preference. The chain breaks not at the upstream geometry, not at the residual-stream representation, but at the readout from representation to behavior.

Cross-model entanglement. The upstream chain elements (token entanglement and layer-wise attribution) replicate broadly across the panel. Both T_A and T_D entanglement are significant on five of the eight models (Qwen2.5-7B, Qwen2.5-14B, Llama-3.1-8B, Yi-1.5-9B, and Phi-3-medium), with outer-letter signed-r in the 0.66 to 0.88 range and outer letters consistently showing stronger entanglement than B and C. Three models show partial entanglement. Mistral-7B is significant for T_D only ($r = -0.795$, $p = 0.006$), Qwen2.5-1.5B for T_A only, and Gemma-3-4B at neither significance threshold. Mistral's asymmetric entanglement is mechanistically informative. Its $S_A \rightarrow A$ shift of $+7.21$ pp is substantial, yet its T_A entanglement is small and not significant ($r = +0.231$, $p = 0.52$). The behavioral effect is real, but the mechanism producing it is not visible to our entanglement probe, consistent with the steering null on the same direction. Mistral exhibits subliminal letter-bias transfer through a route our probes do not capture.

7. Discussion

Two-stage parameter-pull mechanism. The Phi-3 case suggests subliminal transfer may operate in two stages rather than one. The first stage is representational. Token entanglement geometry and a steerable residual-stream direction emerge under the parameter pull, indicating the student has acquired the teacher's bias as an internal structure. The second stage is behavioral. The student's natural inference computation must read out this internal structure as an answer-letter preference. Phi-3 satisfies the first stage cleanly, with panel-strongest entanglement and a fully controllable steering direction that transfers across student conditions, but does not satisfy the second. This decomposition is consistent with the parameter-pull theorem of Cloud et al. (2025). The theorem guarantees parametric pull under shared-architecture distillation, and our cross-model results show this pull commonly produces both representational and behavioral effects. The Phi-3 case demonstrates that the second stage can fail even when the first succeeds, at least for this model and trait, so behavioral expression appears to be a separate condition that the theorem does not directly establish.

Table 2. Cross-model mechanistic and steering results. **Entanglement** columns show signed-r between letter–digit cosine similarity and digit frequency shift from T_{neutral} to T_X (asterisks: $p < 0.05$). **Steering** columns show the bias direction's L_2 norm and the target-letter shift ΔTgt at $\alpha = 2.0$ ($\alpha = 1.0$ for Qwen2.5-1.5B). Steering uses constrained decoding for 6/8 models and free-generation for Llama-3.1-8B and Mistral-7B (constrained artifacts; Appendix B).

	Entanglement		Steering		
Model	$r(A)$	$r(D)$	$\|v\|_2$	ΔTgt (pp)	Verdict
Qwen2.5-7B	$+0.874^*$	-0.876^*	7.90	$+1.71$	clean
Qwen2.5-1.5B	$+0.687^*$	$+0.266$	3.45	$+16.25$	clean (attractor-amplified)
Qwen2.5-14B	$+0.827^*$	-0.665^*	6.96	flat	null
Gemma-3-4B	$+0.375$	-0.390	937	$+3.56$	clean
Llama-3.1-8B	$+0.735^*$	-0.753^*	1.02	$+1.94$	clean (weak/diffuse)
Mistral-7B	$+0.231$	-0.795^*	0.06	-0.25	null
Yi-1.5-9B	$+0.691^*$	-0.659^*	0.20	flat	null
Phi-3-medium	$+0.851^*$	-0.836^*	5.68	$+1.25$	**clean (DISSOCIATION)**

Candidate explanations for the broken readout. The structural cause of the dissociation is unclear. Several candidate explanations are worth investigating. First, late-layer attention patterns may suppress the bias direction at decoding, even though the direction is present in the residual stream. Second, the output head's transformation from final residual stream to next-token logits may not preferentially read out the bias direction. Third, Phi-3's training regime, which relied heavily on curated synthetic data optimized for textbook-quality reasoning (Abdin et al., 2024), may have produced an implicit decoding constraint that ignores spurious positional signals during inference. Each of these is testable through targeted ablations, attention pattern analysis, or comparative training-data interventions.

Mechanism-invisible transfer routes. Mistral-7B presents the inverse pattern. Its $S_A \to A$ shift of $+7.21$ pp is not accompanied by significant T_A entanglement or a steerable A-direction at the tested layer. Behavioral transfer is real but the mechanism producing it is not visible to either of our mechanistic probes at the tested layers and contrasts. This indicates that the application of token-entanglement geometry and difference-in-means steering as inherited from prior subliminal-learning work does not exhaust the mechanisms by which subliminal transfer can occur. Distributed representations, layer-specific encodings, or interaction effects between multiple weak directions could each in principle support behavioral transfer while evading both probes. A more complete picture of subliminal-learning mechanisms would benefit from probes capable of detecting these alternative routes.

Limitations. Three limitations bound the scope of our findings. First, our analysis covers open-weight instruction-tuned models in the 1.5B to 14B parameter range and may not directly transfer to frontier-scale closed models, though the underlying parameter-pull mechanism does not depend on scale. Second, our behavioral measure is MCQ letter-bias, a structurally clean but narrow trait. Whether other structural biases such as formatting preferences, reasoning-step length, or refusal patterns transfer through the same mechanism remains open. Third, our steering analysis depends on layer and contrast selection. Bias directions present at unsampled layers or arising from contrasts other than S_X minus S_{neutral} would be missed. The Mistral and Qwen2.5-14B null verdicts in particular should be read as no extractable direction at the layers and contrasts we tested, not as no bias direction exists.

Future work. Three directions for future work follow most directly. Representation-level auditing tools that operate independently of behavioral signals are needed for distilled models, since the Phi-3 dissociation shows behavioral evaluation can miss inherited biases. Such tools could combine entanglement probes, steering with broader layer and contrast sweeps, and additional probes designed for distributed representations. Generalization studies should test whether the dissociation pattern extends to semantic biases like alignment dispositions or refusal patterns, where the safety stakes are higher. And targeted investigation of Phi-3's broken readout, through ablation of late-layer attention, output-head analysis, or comparison with non-synthetic-trained baselines, would identify which architectural or training-regime factors gate behavioral expression of a representationally-pulled bias.

8. Conclusion

We extend subliminal learning from semantic concepts to structural biases and demonstrate the phenomenon across 7 of 8 instruction-tuned models from six organizations, with target-letter shifts ranging from below 0.2 pp to roughly 10 pp. On Qwen2.5-7B, a four-step mechanistic account links divergence-token causal masking, token-entanglement geometry, layer-wise emergence, and a causally-localized residual-stream direction admitting activation steering. Cross-model replication reveals heterogeneous mechanistic regimes, including a dissociation in

Phi-3-medium where a clean, controllable bias representation does not manifest behaviorally. This constrains the parameter-pull theory of subliminal transfer and shows that behavioral evaluation alone cannot certify a distilled model is free of inherited biases. Together, these results widen the safety-relevant scope of subliminal learning and underscore that what distilled models inherit from their teachers cannot, at present, be fully assessed by behavioral auditing alone.

Impact Statement

Subliminal learning poses a structural risk to synthetic-data and distillation pipelines, since biases can transfer through training data that contains no surface signature of the target trait, evading content-based filtering. As these pipelines become increasingly central to modern AI development, characterizing what biases transfer and whether behavioral evaluation can detect them is essential for trustworthy deployment. Our work contributes to this characterization, supporting the development of more rigorous safety auditing for distilled models. We acknowledge that better mechanistic understanding could in principle be used to induce biases as well as detect them, but the underlying mechanism was characterized in prior work, and the defensive value of broader understanding outweighs marginal capability uplift.

References

Abdin, M., Jacobs, S. A., Awan, A. A., Aneja, J., Awadallah, A., Awadalla, H. H., Bach, N., Bahree, A., Bakhtiari, A., Behl, H. S., Benhaim, A., Bilenko, M., Bjorck, J., Bubeck, S., Cai, M., Mendes, C. C. T., Chen, W., Chaudhary, V., Chopra, P., Giorno, A. D., de Rosa, G., Dixon, M., Eldan, R., Fragoso, V., Iter, D., Goswami, A., Gunasekar, S., Haider, E., Hao, J., Hewett, R. J., Huynh, J., Javaheripi, M., Jin, X., Kauffmann, P., Karampatziakis, N., Kim, D., Kim, Y. J., Khademi, M., Kurilenko, L., Lee, J. R., Lee, Y. T., Li, Y., Liang, C., Liu, W., Lin, E., Lin, Z., Madan, P., Mitra, A., Modi, H., Nguyen, A. H., Norick, B., Patra, B., Perez-Becker, D., Portet, T., Pryzant, R., Qin, H., Radmilac, M., Ren, L., Rosset, C., Roy, S., Saarikivi, O., Saied, A., Salim, A., Santacroce, M., Shah, S., Shang, N., Sharma, H., Song, X., Ruwase, O., Vaddamanu, P., Wang, X., Ward, R., Wang, G., Witte, P. A., Wyatt, M., Xu, C., Xu, J., Yadav, S., Yang, F., Yang, Z., Yu, D., yuan Zhang, C., Zhang, C., Zhang, J., Zhang, L. L., Zhang, Y., Zhang, Y., Zhou, X., and Yang, Y. Phi-3 technical report: A highly capable language model locally on your phone. *ArXiv*, abs/2404.14219, 2024. URL https://api.semanticscholar.org/CorpusID:269293048.

Belrose, N., Furman, Z., Smith, L., Halawi, D., Ostrovsky, I. V., McKinney, L., Biderman, S., and Steinhardt, J. Eliciting latent predictions from transformers with the tuned lens. *ArXiv*, abs/2303.08112, 2023. URL https://api.semanticscholar.org/CorpusID:257504984.

Chen, H., Wang, L., Yang, N., Zhu, Y., Zhao, Z., Wei, F., and Dou, Z. Little giants: Synthesizing high-quality embedding data at scale. In *North American Chapter of the Association for Computational Linguistics*, 2024. URL https://api.semanticscholar.org/CorpusID:273550022.

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., and Tafjord, O. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Cloud, A., Le, M., Chua, J., Betley, J., Sztyber-Betley, A., Hilton, J., Marks, S., and Evans, O. Subliminal learning: Language models transmit behavioral traits via hidden signals in data. *ArXiv*, abs/2507.14805, 2025. URL https://api.semanticscholar.org/CorpusID:280280217.

Dang, J., Xie, B., and Younis, O. G. Subliminal transfer of unsafe behaviors in ai agent distillation. 2026. URL https://api.semanticscholar.org/CorpusID:287607144.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A. S., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Rozière, B., Biron, B. M., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C. C., Nikolaidis, C., Allonsius, D., Song, D. J., Pintz, D., Livshits, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E. A., Lobanova, E. I., Dinan, E., Smith, E. M., Radenovic, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G. L., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I. A., Kloumann, I. M., Misra, I., Evtimov, I., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J.-Q., Alwala, K. V., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K. R., El-Arini, K., Iyer, K., Malik, K., ley Chiu, K., Bhalla, K., Rantala-Yeary, L., van der Maaten, L., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., de Oliveira, L., Muzzi, M., hesh Pasupuleti, M., Singh, M., Paluri, M.,

Kardas, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M. H. M., Lewis, M., Si, M., Singh, M. K., Hassan, M., Goyal, N., Torabi, N., lay Bashlykov, N., Bogoychev, N., Chatterji, N. S., Duchenne, O., cCelebi, O., Alrassy, P., Zhang, P., Li, P., Vasić, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P. S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R. S., Stojnic, R., Raileanu, R., Girdhar, R., Patel, R., Sauvestre, R., nie Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., hana Chennabasappa, S., Singh, S., Bell, S., Kim, S. S., Edunov, S., Nie, S., Narang, S., Raparthy, S. C., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., Do, V., Vogeti, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., ney Meers, W., Martinet, X., Wang, X., Tan, X. E., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z. D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A. K., Grattafiori, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Vaughan, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Franco, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, P.-Y. B., Loyd, B., de Paola, B., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.-H., Cai, C., Tindal, C., Feichtenhofer, C., Civin, D., Beaty, D., Kreymer, D., Li, S.-W., Wyatt, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Ozgenel, F., Caggioni, F., Guzmán, F. P., Kanayet, F. J., Seide, F., Florez, G. M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Thattai, G., Herman, G., Sizov, G., Zhang, G., Lakshminarayanan, G., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Molybog, I., Tufanov, I., Veliche, I.-E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Asher, J., Gaya, J.-B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., KamHou, U., Saxena, K., Prasad, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Huang, K., Chawla, K., Lakhotia, K., Huang, K., Chen, L., Garg, L., Lavender, A., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Tsimpoukelli, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M., Seltzer, M. L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M. J., Metanat, M., Rastegari, M., ish Bansal, M., Santhanam, N., Parks, N., White, N., ata Bawa, N., Singhal, N., Egebo, N., Usunier, N., Laptev, N. P., Dong, N., Zhang, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., dro Rittner, P., Bontrager, P., Roux, P., Dollár, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R., Nayani, R., Mitra, R., Li, R., Hogan, R., Battey, R., Wang, R., han Maheswari, R., Howes, R., Rinott, R., Bondu, S. J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S. Y., Pan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Feng, S., Lin, S., Zha, S., Shankar, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S. K., Cho, S.-B., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Kohler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V. O., Montanez, V., Mohan, V., Kumar, V., Mangla, V., Ionescu, V., Poenaru, V. A., Mihailescu, V. T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wang, X., Wu, X., Wang, X., Xia, X., Wu, X., Gao, X., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang, Y., Zhang, Y., Adi, Y., Nam, Y., Wang, Y., Hao, Y., Qian, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., and Zhao, Z. The llama 3 herd of models. 2024. URL https://api.semanticscholar.org/CorpusID:271571434.

Flemings, J. and Annavaram, M. Differentially private knowledge distillation via synthetic text generation. In *Annual Meeting of the Association for Computational Linguistics*, 2024. URL https://api.semanticscholar.org/CorpusID:268230792.

Gisler, I., He, Z., Zurich, T. Q. E., of Cambridge, U., and University, P. You didn't have to say it like that: Subliminal learning from faithful paraphrases. 2026. URL https://api.semanticscholar.org/CorpusID:286428887.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D. X., and Steinhardt, J. Measuring massive multitask language understanding. *ArXiv*, abs/2009.03300,

2020. URL https://api.semanticscholar.org/CorpusID:221516475.

Hu, J. E., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., and Chen, W. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685, 2021. URL https://api.semanticscholar.org/CorpusID:235458009.

Jain, S., Kirk, R., Lubana, E. S., Dick, R. P., Tanaka, H., Grefenstette, E., Rocktäschel, T., and Krueger, D. S. Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks. In *International Conference on Learning Representations (ICLR)*, 2024.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de Las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mistral 7b. *ArXiv*, abs/2310.06825, 2023. URL https://api.semanticscholar.org/CorpusID:263830494.

Kamath, G. T. A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ram'e, A., Rivière, M., Rouillard, L., Mesnard, T., Cideron, G., Grill, J.-B., Ramos, S., Yvinec, E., Casbon, M., Pot, E., Penchev, I., Liu, G., Visin, F., Kenealy, K., Beyer, L., Zhai, X., Tsitsulin, A., Busa-Fekete, R. I., Feng, A., Sachdeva, N., Coleman, B., Gao, Y., Mustafa, B., Barr, I., Parisotto, E., Tian, D., Eyal, M., Cherry, C., Peter, J.-T., Sinopalnikov, D., Bhupatiraju, S., Agarwal, R., Kazemi, M., Malkin, D., Kumar, R., Vilar, D., Brusilovsky, I., Luo, J., Steiner, A., Friesen, A., Sharma, A., Sharma, A., Gilady, A. M., Goedeckemeyer, A., Saade, A., Kolesnikov, A., Bendebury, A., Abdagic, A., Vadi, A., Gyorgy, A., Pinto, A. S., Das, A., Bapna, A., Miech, A., Yang, A., Paterson, A., Shenoy, A., Chakrabarti, A., Piot, B., Wu, B., Shahriari, B., Petrini, B., Chen, C., Lan, C. L., Choquette-Choo, C. A., Carey, C., Brick, C., Deutsch, D., Eisenbud, D., Cattle, D., Cheng, D., Paparas, D., Sreepathihalli, D. S., Reid, D., Tran, D., Zelle, D., Noland, E., Huizenga, E., Kharitonov, E., Liu, F., Amirkhanyan, G., Cameron, G., Hashemi, H., Klimczak-Pluci'nska, H., Singh, H., Mehta, H., Lehri, H. T., Hazimeh, H., Ballantyne, I., Szpektor, I., Nardini, I., Pouget-Abadie, J., Chan, J., Stanton, J., Wieting, J. M., Lai, J., Orbay, J., Fernandez, J., Newlan, J., Ji, J., Singh, J., Black, K., Yu, K., Hui, K., Vodrahalli, K., Greff, K., Qiu, L., Valentine, M., Coelho, M., Ritter, M., Hoffman, M., Watson, M., Chaturvedi, M., Moynihan, M., Ma, M., Babar, N., Noy, N., Byrd, N., Roy, N., Momchev, N., Chauhan, N., Bunyan, O., Botarda, P., Caron, P., Rubenstein, P. K., Culliton, P., Schmid, P., Sessa, P. G., mei Xu, P., Stańczyk, P., Tafti, P. D., Shivanna, R., Wu, R., Pan, R., Rokni, R. A., Willoughby, R., Vallu, R., Mullins, R., Jerome, S., Smoot, S., Girgin, S., Iqbal, S., Reddy, S., Sheth, S., Põder, S., Bhatnagar, S., Panyam, S. R., Eiger, S., Zhang, S., Liu, T., Yacovone, T., Liechty, T., Kalra, U., Evci, U., Misra, V., Roseberry, V., Feinberg, V., Kolesnikov, V., Han, W., Kwon, W., Chen, X., Chow, Y., Zhu, Y., Wei, Z., Egyed, Z., Cotruta, V., Giang, M., Kirk, P., Rao, A., Lo, J., Moreira, E., Martins, L. G., Sanseviero, O., Gonzalez, L., Gleicher, Z., Warkentin, T., Mirrokni, V. S., Senter, E., Collins, E., Barral, J., Ghahramani, Z., Hadsell, R., Matias, Y., Sculley, D., Petrov, S., Fiedel, N., Shazeer, N., Vinyals, O., Dean, J., Hassabis, D., Kavukcuoglu, K., Farabet, C., Buchatskaya, E., Alayrac, J.-B., Anil, R., Lepikhin, D., Borgeaud, S., Bachem, O., Joulin, A., Andreev, A., Hardin, C., Dadashi, R., and Hussenot, L. Gemma 3 technical report. *ArXiv*, abs/2503.19786, 2025. URL https://api.semanticscholar.org/CorpusID:277313563.

Kapania, S., Ballard, S., Kessler, A. R., and Vaughan, J. W. Examining the expanding role of synthetic data throughout the ai development pipeline. *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 2025. URL https://api.semanticscholar.org/CorpusID:275993778.

Lin, S., Hilton, J., and Evans, O. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, 2022.

Magistrali, I., Berdoz, F., Dauncey, S., and Wattenhofer, R. Subliminal signals in preference labels. *ArXiv*, abs/2603.01204, 2026. URL https://api.semanticscholar.org/CorpusID:286223270.

Morgulis, G. and Hewitt, J. Subliminal steering: Stronger encoding of hidden signals. 2026. URL https://api.semanticscholar.org/CorpusID:287833995.

Rimsky, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., and Turner, A. M. Steering llama 2 via contrastive activation addition. *ArXiv*, abs/2312.06681, 2023. URL https://api.semanticscholar.org/CorpusID:266174252.

Schrodi, S., Kempf, E., Barez, F., and Brox, T. Towards understanding subliminal learning: When and how hidden biases transfer. *ArXiv*, abs/2509.23886, 2025. URL https://api.semanticscholar.org/CorpusID:281676260.

Stolfo, A., Balachandran, V., Yousefi, S., Horvitz, E., and Nushi, B. Improving instruction-following in language models through activation steering. *ArXiv*, abs/2410.12877, 2024. URL https://api.semanticscholar.org/CorpusID:273403586.

Venhoff, C., Arcuschin, I., Torr, P. H. S., Conmy, A., and Nanda, N. Understanding reasoning in thinking language models via steering vectors. *ArXiv*, abs/2506.18167, 2025. URL https://api.semanticscholar.org/CorpusID:278340657.

Vir, R. and Bhatnagar, S. Subliminal corruption: Mechanisms, thresholds, and interpretability. *ArXiv*, abs/2510.19152, 2025. URL https://api.semanticscholar.org/CorpusID:282272681.

Wang, X., Ma, B., Hu, C., Weber-Genzel, L., Röttger, P., Kreuter, F., Hovy, D., and Plank, B. "my answer is c": First-token probabilities do not match text answers in instruction-tuned language models. In *Annual Meeting of the Association for Computational Linguistics*, 2024. URL https://api.semanticscholar.org/CorpusID:267782369.

Wang, Z. Logitlens4llms: Extending logit lens analysis to modern large language models. *ArXiv*, abs/2503.11667, 2025. URL https://api.semanticscholar.org/CorpusID:277066402.

Wu, Z., Arora, A., Wang, Z., Geiger, A., Jurafsky, D., Manning, C. D., and Potts, C. ReFT: Representation finetuning for language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Yang, Q. A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Dong, G., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y.-C., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., Qiu, Z., Quan, S., and Wang, Z. Qwen2.5 technical report. *ArXiv*, abs/2412.15115, 2024. URL https://api.semanticscholar.org/CorpusID:274859421.

Young, A. A., Chen, B., Li, C., Huang, C., Zhang, G., Zhang, G., Li, H., Zhu, J., Chen, J., Chang, J., Yu, K., Liu, P., Liu, Q., Yue, S., Yang, S., Yang, S., Yu, T., Xie, W., Huang, W., Hu, X., Ren, X., Niu, X., Nie, P., Xu, Y., Liu, Y., Wang, Y., Cai, Y., Gu, Z., Liu, Z., and Dai, Z. Yi: Open foundation models by 01.ai. *ArXiv*, abs/2403.04652, 2024. URL https://api.semanticscholar.org/CorpusID:268264158.

Zur, A., Ying, Z., Loftus, A. R., Şahin, K., Yu, S., Quirke, L., Rott Shaham, T., Shapira, N., Orgad, H., and Bau, D. Token entanglement in subliminal learning. In *Mechanistic Interpretability Workshop at NeurIPS*, 2025.

A. Permutation Neutrality Verification

This appendix documents the algorithm used to construct T_{neutral}'s training data and verifies that the resulting correct-answer-letter distribution is statistically indistinguishable from uniform at every sample size used in the pipeline. Without this verification, S_{neutral} cannot serve as a clean baseline for the shifts reported in §4, since any residual letter preference in T_{neutral} would propagate into S_{neutral} and contaminate every reported shift.

Permutation construction. For each MMLU item i with original correct-answer index $c_i \in \{0, 1, 2, 3\}$, we draw a permutation $\pi_i \sim \text{Uniform}(S_4)$ deterministically from the first 8 bytes of SHA-256($\texttt{train} \,||\, s_{\text{train}} \,||\, \texttt{neutral} \,||\, i$). The four answer options are reordered by π_i, and the correct-answer letter for item i becomes whichever of $\{A, B, C, D\}$ now occupies position $\pi_i^{-1}(c_i)$. Eval-time permutation uses an independent algorithm and an independent seed namespace, described below.

Chi-square verification. We re-ran the permutation algorithm over fresh source items at every sample size used in the main pipeline, under two independent global seeds, and tested the resulting correct-answer-letter counts against the uniform null with chi-square goodness-of-fit (df $= 3$). Table 3 reports results.

Table 3. Chi-square goodness-of-fit results for the T_{neutral} permutation algorithm under two independent seeds. All $p > 0.05$ at every sample size, so the uniform null is not rejected.

Seed	N	A	B	C	D	χ^2	p
0	200	51	62	45	42	4.68	0.20
0	4,800	1,257	1,181	1,164	1,198	4.09	0.25
0	3,200	834	793	787	786	1.96	0.58
0	100,000	25,177	24,913	24,903	25,007	1.93	0.59
1	200	49	49	54	48	0.44	0.93
1	4,800	1,238	1,185	1,166	1,211	2.46	0.48
1	3,200	830	775	763	832	4.90	0.18
1	100,000	25,146	25,042	25,026	24,786	2.78	0.43

The two sample sizes load-bearing for S_{neutral} validity are $N = 200$ (sanity holdout used for teacher gating) and $N = 4{,}800$ (the training pool). Both pass at $p > 0.19$ under both seeds. The worst case across all eight (sample-size $\times$ seed) combinations is $\chi^2 = 4.90$, $p = 0.18$, still 3.5 times above the 0.05 threshold. At $N = 100{,}000$ the per-letter rate is within 0.18 pp of 25.0% under both seeds, confirming asymptotic uniformity.

The actual counterbalanced evaluation set used in §4 is constructed differently. We start with 800 source items and rotate the target letter deterministically through $\{A, B, C, D\}$, yielding exactly 800 items per letter and a chi-square statistic of zero by construction. The $N = 3{,}200$ row in Table 3 instead asks whether the permutation algorithm would remain uniform at that scale, and it does.

Seed namespace independence. Training-time and eval-time permutations are drawn from disjoint seed namespaces. The training-time RNG seed for item i derives from SHA-256($\texttt{train} \,||\, s_{\text{train}} \,||\, \texttt{condition} \,||\, i$), while the eval-time RNG seed derives from SHA-256($\texttt{eval} \,||\, s_{\text{eval}} \,||\, \text{target letter} \,||\, i$). The namespace prefix differs in the first token, producing different digests by construction. The numerical seeds also differ in practice ($s_{\text{train}} = 0$, $s_{\text{eval}} = 1001$). Any item-specific positional memorization T_{neutral} might acquire during training cannot transfer to eval, since each item's eval-time correct-answer position is determined by an independent hash.

B. Constrained-Decoding Measurement Artifacts

The free-generation evaluation methodology used throughout §4 is a deliberate choice. Constrained letter-token argmax, the alternative widely used in MCQ benchmarks, has been shown to systematically over-report positional priors at the expense of model-actual answers (Wang et al., 2024). We replicate this artifact on our 8-model panel and document two specific consequences for subliminal-learning analysis.

Cross-model panel comparison. Table 4 reports the maximum per-letter divergence between the two decoding modes on the pristine base model for each model in our panel, evaluated on the same 400-item subset (3,200 for Qwen2.5-14B).

Table 4. Per-model agreement between constrained and free-generation evaluation on pristine base models. $\max|\Delta|$ is the largest per-letter rate divergence across $\{A, B, C, D\}$. Attractor reports the letter and sign of that maximum. Acc gap is constrained accuracy minus free-generation accuracy on the same items.

Model	n	$\max\|\Delta\|$	Attractor	Acc gap
Qwen2.5-7B	400	3.00	—	−1.25
Yi-1.5-9B	400	6.25	A −6.25	+1.00
Phi-3-medium	400	10.50	A +10.50	−0.50
Gemma-3-4B	400	10.75	A +10.75	+0.50
Qwen2.5-1.5B	400	26.00	B +26.00	−0.50
Qwen2.5-14B	3200	34.38	A +34.38	−20.97
Mistral-7B	400	45.50	A +45.50	+18.75
Llama-3.1-8B	400	61.75	A +61.75	+27.25

The two decoding methods agree to within 3 pp on Qwen2.5-7B alone. On 6 of 8 models, the constrained method over-reports at least one letter by ≥ 10 pp, with Mistral and Llama showing 45.5 pp and 61.75 pp A-attractors respectively. The accuracy gap on Llama reaches 27.25 pp, since

constrained eval reads 36.5% accuracy on the same items where free-generation reads 63.75%.

A per-item agreement check on 10 randomly selected eval items per model finds 9 of 10 agreement on Yi (passing a 9-of-10 gate) but only 1 of 10 on Llama, where constrained eval picks A on every item regardless of what the model's free-generation continuation produces. On representative Llama disagreement items, free-generation outputs "The answer is D," "The answer is C," "The answer is B," while the constrained method picks A in every case.

Finding-flipping artifact on Qwen2.5-1.5B. On Qwen2.5-1.5B, switching decoding mode flips the sign of two of four behavioral conditions and reverses the qualitative interpretation of the model's transfer pattern.

Table 5. Per-condition target-letter shifts on Qwen2.5-1.5B, under each evaluation mode. Bold marks sign flips between the two methods.

Condition	Constrained	Free-gen	
$S_A \to A$	$+4.43$	$+4.40$	—
$S_B \to B$	$+8.49$	$+1.18$	$7\times$ shrink
$S_C \to C$	$\mathbf{-2.49}$	$\mathbf{+0.10}$	sign flip
$S_D \to D$	$\mathbf{-2.08}$	$\mathbf{+1.04}$	sign flip

Under constrained eval, $S_C \to C$ shows -2.49 pp and $S_D \to D$ shows -2.08 pp, with an apparent native B-attractor (S_{neutral} picks B 52.75% of the time under constrained decoding) absorbing mass that would otherwise flow to target letters. Under free-generation, the same students show $+0.10$ pp and $+1.04$ pp, and S_{neutral} instead picks A 41.49% of the time with no native B-attractor visible. The two methods produce different qualitative claims about the same model. Constrained eval supports a "native-attractor-gated transfer" hypothesis. Free-generation supports a "weak universal transfer" hypothesis.

Steering-invalidating artifact on Mistral-7B. Mistral-7B's pristine baseline collapses to 69.25% A under constrained eval, against 23.75% A under free-generation, a 45.5 pp artifact. This contaminates the activation-steering pipeline. Under constrained eval, the $\alpha = 0$ steering baseline registers at 83.25% A, leaving B, C, D rates summing to $\leq 16.75\%$ with no headroom for any patching signal to register. The full α-sweep under constrained eval shows ΔD of $+0.25, -1.25, -3.75$ at $\alpha = 1, 2, 5$, all driven by the saturated A-attractor rather than any direction-mediated signal.

Re-running the same steering pipeline (identical direction, layer, and α range) under free-generation yields a coherent baseline ($A = 21.75\%$, $B = 25.25\%$, $C = 34.25\%$, $D = 18.75\%$) and a clean null verdict at $\alpha = 2.0$ ($\Delta D = -0.25$ pp, indistinguishable from random and reverse controls). The constrained-vs-free disagreement was not just inflating the apparent steering effect, it was masking the actual mechanistic conclusion. Under free-generation, Mistral's steering result is "no extractable D-direction at the chosen layer," not "evaluation broken."

Summary. On the 8-model panel, 1 of 8 passes a 3 pp constrained-vs-free agreement gate, 2 of 8 at 5 pp, 6 of 8 show $\geq$ 10 pp letter-attractor on at least one letter (range 10.5–61.75 pp), and 3 of 8 show $\geq$ 30 pp letter-attractor (Qwen2.5-14B, Mistral, Llama). Beyond replicating the artifact reported by prior work (Wang et al., 2024), we document two finding-changing consequences specific to subliminal-learning analysis. First, the artifact can flip the sign of qualitative claims about transfer patterns (Qwen2.5-1.5B). Second, it can invalidate downstream activation-steering experiments on affected models (Mistral-7B). These extensions motivate the choice of free-generation as canonical evaluation throughout this work.

C. Behavioral Decomposition: Competence vs Letter Bias

To verify that the headline shifts in Table 1 reflect genuine letter-preference transfer rather than competence improvements on items where the target letter happens to be correct, we decompose each per-condition shift into a competence component and a distractor component.

Method. For each model and condition S_X, define

$$\begin{aligned}\Delta_{\text{correct}}(X) &= P(\text{pick } X \mid \text{correct} = X)\big|_{S_X} \\ &\quad - P(\text{pick } X \mid \text{correct} = X)\big|_{S_{\text{neutral}}} \\ \Delta_{\text{distract}}(X) &= P(\text{pick } X \mid \text{correct} \neq X)\big|_{S_X} \\ &\quad - P(\text{pick } X \mid \text{correct} \neq X)\big|_{S_{\text{neutral}}}\end{aligned}$$

The 800-counterbalanced evaluation set presents each of 800 source items four times with the correct answer rotated through every position, so $\frac{1}{4}$ of items have correct $= X$ and $\frac{3}{4}$ have correct $\neq X$. The overall shift satisfies the identity

$$\Delta_{\text{overall}}(X) = 0.25 \cdot \Delta_{\text{correct}}(X) + 0.75 \cdot \Delta_{\text{distract}}(X).$$

A bias that lifts Δ_{correct} while leaving Δ_{distract} at zero is competence-only. The model becomes more accurate on items where the target letter is correct, but does not preferentially pick the target letter on items where it is wrong. A bias that lifts Δ_{distract} is genuine letter selection. The model picks X even when X is the wrong answer.

Headline result. For 7 of 8 models in the panel, the headline target-letter shifts cannot be explained by a competence-shift mechanism. The increase in target-letter selection

is driven primarily by selection on distractor items (those where the target letter is wrong), establishing the shifts as genuine letter biases rather than artifacts of fine-tuning slightly improving target-letter-correct accuracy. Phi-3-medium is the lone exception, with $|\Delta_{\text{distract}}| \leq 0.06$ pp on every condition. Its small overall shifts are entirely a competence story.

Table 6. Per-model mean of Δ_{distract} across the four conditions, sorted descending. Phi-3-medium is the only model averaging zero on this metric.

Model	Mean Δ_{dis}	Bias type
Mistral-7B	$+7.42$	strong letter bias
Gemma-3-4B	$+3.81$	strong letter bias
Llama-3.1-8B	$+2.58$	letter bias
Qwen2.5-1.5B	$+1.86$	letter bias (S_A drives)
Qwen2.5-7B	$+1.41$	letter bias (S_A drives)
Qwen2.5-14B	$+0.93$	mixed (competence-leaning)
Yi-1.5-9B	$+0.86$	weak letter bias
Phi-3-medium	$+0.00$	competence-only

Phi-3 dissociation in detail. Phi-3-medium's per-condition decomposition reveals a strict competence-only signature.

Table 7. Phi-3-medium per-condition decomposition. $|\Delta_{\text{distract}}| \leq 0.06$ pp on every condition.

Cond	Δ_{correct} (pp)	Δ_{distract} (pp)
S_A	$+0.13$	-0.04
S_B	$+0.54$	$+0.06$
S_C	$+0.75$	-0.03
S_D	$+0.83$	-0.01

The largest distractor effect on Phi-3 across all four conditions is 0.06 pp, on the order of the per-seed standard deviation (~ 0.08 pp) and statistically indistinguishable from sampling noise. The small positive Δ_{correct} values (range $+0.13$ to $+0.83$ pp) reflect a slight fine-tuning improvement on target-letter-correct items, not a letter-selection preference.

This decomposition makes Phi-3's behavioral non-transfer cleaner than the headline shifts in Table 1 alone suggest. Phi-3's small overall shifts are not weak transfer near the noise floor. They are competence improvements with no measurable letter-bias component, in sharp contrast to the other 7 models. This is the dissociation that motivates the mechanistic analysis in §6.

Striking sub-cases. Three sub-cases in the master decomposition (Table 8) are worth flagging.

Qwen2.5-1.5B condition S_C exhibits a pure letter-bias signature uncoupled from competence. $\Delta_{\text{correct}} = -1.17$ pp (the student is less accurate than S_{neutral} on C-correct items) yet $\Delta_{\text{distract}} = +0.53$ pp (the student picks C more often on items where C is wrong). The model becomes less accurate on C-questions while picking C more often as a distractor.

Mistral-7B condition S_C shows nearly equal competence and distractor components ($\Delta_{\text{correct}} = +9.17$, $\Delta_{\text{distract}} = +9.25$), the cleanest 1:1 split in the panel. The bias acts uniformly on the letter regardless of whether it is the correct answer, the simplest "the model just prefers the letter" signature.

Qwen2.5-14B is the panel's only competence-leaning transferrer. Δ_{correct} exceeds Δ_{distract} on all four conditions (often by 2 to $12\times$), suggesting its small universal shifts arise partly from fine-tuning improving accuracy on target-letter-correct items rather than from pure letter preference.

Full master decomposition. Table 8 reports the full per-model, per-condition decomposition. The identity $\Delta_{\text{overall}} = 0.25 \cdot \Delta_{\text{correct}} + 0.75 \cdot \Delta_{\text{distract}}$ is satisfied to three decimal places in every row.

Table 8. Full per-model per-condition decomposition. Δ_{corr}, Δ_{dis}, and Δ in percentage points.

Model	Cond	Δ_{corr}	Δ_{dis}	Δ
Qwen2.5-7B	S_A	$+3.85$	$+3.93$	$+3.91$
	S_B	-0.25	$+1.03$	$+0.71$
	S_C	$+0.46$	$+1.07$	$+0.92$
	S_D	-1.46	-0.38	-0.65
Qwen2.5-1.5B	S_A	$+3.17$	$+4.81$	$+4.40$
	S_B	$+1.38$	$+1.11$	$+1.18$
	S_C	-1.17	$+0.53$	$+0.10$
	S_D	$+1.04$	$+1.04$	$+1.04$
Qwen2.5-14B	S_A	$+2.25$	$+0.60$	$+1.01$
	S_B	$+3.38$	$+1.57$	$+2.02$
	S_C	$+3.08$	$+1.38$	$+1.80$
	S_D	$+2.00$	$+0.17$	$+0.63$
Gemma-3-4B	S_A	$+4.96$	$+3.38$	$+3.77$
	S_B	$+6.00$	$+5.10$	$+5.32$
	S_C	$+3.83$	$+3.96$	$+3.93$
	S_D	$+3.67$	$+2.81$	$+3.02$
Llama-3.1-8B	S_A	$+0.58$	$+0.90$	$+0.82$
	S_B	$+3.33$	$+3.54$	$+3.49$
	S_C	$+5.04$	$+3.57$	$+3.94$
	S_D	$+2.71$	$+2.26$	$+2.38$
Mistral-7B	S_A	$+7.54$	$+7.10$	$+7.21$
	S_B	$+5.83$	$+4.06$	$+4.50$
	S_C	$+9.17$	$+9.25$	$+9.23$
	S_D	$+13.04$	$+9.28$	$+10.22$
Yi-1.5-9B	S_A	$+0.04$	$+0.90$	$+0.69$
	S_B	$+0.50$	$+0.96$	$+0.84$
	S_C	$+0.63$	$+1.15$	$+1.02$
	S_D	$+0.33$	$+0.40$	$+0.39$
Phi-3-medium	S_A	$+0.13$	-0.04	$+0.00$
	S_B	$+0.54$	$+0.06$	$+0.18$
	S_C	$+0.75$	-0.03	$+0.17$
	S_D	$+0.83$	-0.01	$+0.20$