MEASURE ONCE, STEER WITH MANY: CONTEXT CONTAMINATION AND COMPOSITE CONTROL FOR PERSONA DRIFT IN LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Monitors and controllers of persona drift read a model's activations against a persona direction: the difference between the model's activations with and without the persona in the system prompt, extracted once before the conversation and then held fixed (the frozen direction). Re-extracted later in the dialogue, with the conversation so far as a prefix, this direction appears to rotate and shrink. We show that this movement is context contamination, not a moving persona. It follows how much new content the conversation adds: replacing the assistant's replies with another persona's barely changes it, replacing them with one repeated sentence slows it several-fold, and conversation length alone predicts it. The re-extracted direction is also the worse instrument: it does not track the judged persona within a conversation, and steering along it does worse. Steering harder along the frozen direction, however, degrades coherence and induces repetition. We therefore extract coherence and repetition as directions with the same recipe and steer along a weighted composite of all three, recovering most of the coherence and variety that pushing the persona costs. Against five published steering methods on four models, each at its best strength, some configuration of ours is at least as good on two of three measured qualities in nearly every setting. We also find that the more text accumulates, the harder the frozen direction is to express in other instructions' re-extracted directions, and that a malicious persona fades three to five times faster than a compassionate one. Our judges, blind to the method behind each reply, agree with people as closely as people agree with one another.

1 INTRODUCTION

When a language model is given a persona or other behavioral instruction in its system prompt, how long does the instruction persist? Behavioral studies show that it erodes over many turns (Li et al., 2024; Choi et al., 2024; Dongre et al., 2025). Two mechanistic accounts explain the erosion. In the attention account, the model attends less and less to the system prompt as the context grows (Dongre et al., 2026). In the geometric account, an instruction occupies a direction in activation space, and drift is read as the model's activation projecting less and less onto that direction (Chen et al., 2025; Lu et al., 2026).

The geometric account rests on an assumption that has not been tested along a conversation. The persona direction, the contrast between activations with and without the instruction, is extracted once before the dialogue and then held fixed, so every decline in the projection is read as the model losing the persona. The natural test is to re-extract the direction later in the conversation, with the dialogue so far as a fixed prefix. When we do so, the direction rotates away from its starting point and loses much of its length (Figure 1), in every model family we tested but one. Taken at face value, this suggests that the persona has moved and that a frozen monitor reads the wrong quantity, which is how the closest prior work interprets the same effect (Li et al., 2026; Adarsh et al., 2026; Lampinen et al., 2026).

We show that this interpretation does not hold, which is our first contribution. Three measurements establish that the movement reflects the context rather than the persona. First, it follows how much new content the conversation contains, not whose persona produced it. Replacing the

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.

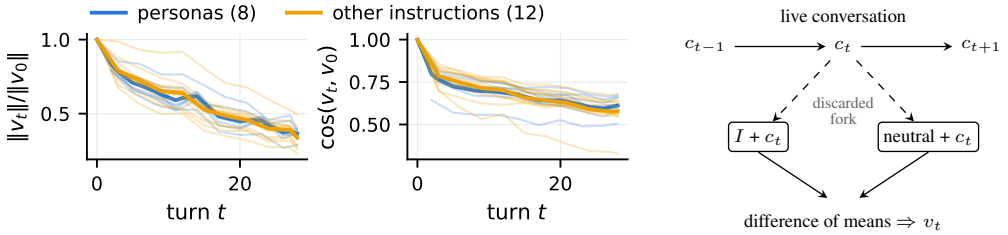

Figure 1: The re-extracted direction shrinks and rotates even on a fully neutral conversation, where nothing in the context relates to any instruction. For an instruction I, v_0 is its frozen direction, extracted before the conversation, and v_t the same direction re-extracted at conversation turn t. Left: the length of v_t relative to v_0, for eight personas (blue) and twelve style, format, language and epistemic instructions (orange); thin lines are single instructions, heavy lines the means. Middle: the cosine between v_t and v_0. Personas and other instructions behave alike. Right: how v_t is extracted. The system prompt I and a neutral one each run on a discarded fork that holds the conversation up to turn t, c_t, as a fixed prefix; v_t is the difference of their mean activations.

assistant's replies with another persona's leaves it almost unchanged, whereas replacing them with a single sentence repeated to the same length slows it several-fold. Second, it is the same for every instruction: a single curve in conversation length, fitted on other conversations and other models, predicts it without knowledge of the instruction. Third, the re-extracted direction is the inferior instrument for both uses of a direction. On unsteered conversations the projection onto the frozen direction falls with the judged persona while the projection onto the re-extracted direction does not, and steering along the frozen direction wins most comparisons against the re-extracted one. We call this *context contamination*: re-extracted after a long prefix, the direction measures the conversation, and the frozen direction remains the persona's direction.

Separating these effects reveals four changes as a conversation grows, each with its own cause. The re-extracted direction rotates and shrinks with the new content the conversation adds. Attention to the system prompt falls with context length alone, and just as fast when the added text is one repeated reply. The persona's expression, its projection onto the frozen direction, fades at a rate set by what the conversation contains: a malicious persona fades three to five times faster than a compassionate one, while the user's stance matters far less. And the frozen direction becomes harder to express as a combination of twenty instructions' re-extracted directions, to an extent set by how much text has accumulated.

The frozen direction, the only one whose projection tracks the persona within a conversation, is therefore the right direction to monitor and to steer along, but steering along the persona direction alone has a cost. The harder the persona is pushed, the less coherent and the more repetitive the replies become. Our main contribution addresses this. **We extract coherence and repetition as directions with the same recipe used for a persona, and steer along a weighted combination of all three.** This composite recovers the coherence and variety that pushing the persona costs, and recovers more where more was lost. Tuned per setting, it is at least as good as the frozen direction on at least two of the three measures in nearly every setting and on all three in two thirds. With weights transferred from another conversation, a harder setting, it is at least as good on two of the three in three quarters of conversations and on all three in a third, trading a little persona adherence for coherence and variety.

Contributions.

- **The re-extracted direction measures the context, not the persona** (Sections 4–6), and four changes with separate causes unfold as a conversation grows.
- **Steering saturates, and its two modes fail differently** (Section 7): beyond a model-dependent strength, pushing the persona harder yields repetition rather than persona, whether the push edits activations or re-weights attention. They differ otherwise. Re-weighting attention towards the system prompt preserves coherence but, on two of three models, loses more of the model's ability

to tell a stated detail from a planted near-miss, whereas editing activations pays in coherence (Figure 2).
- **A composite steering direction** (Section 8) that repairs what pushing along the persona direction alone costs, with weights tuned, transferred or set turn by turn, and reaches combinations of adherence and variety that no single-direction controller reaches (Figures 3–5).

2 Related work

Persona drift in dialogue. That an assigned persona or instruction erodes over many turns is established behaviorally, in self-chat, long dialogues and multi-instruction settings (Choi et al., 2024; Dongre et al., 2025; Laban et al., 2026; He et al., 2024; Gonnermann-Müller et al., 2026; Luz de Araujo et al., 2026). The user also shapes it, since models tend to agree with the user (Jain et al., 2026) and conform to others' opinions (Zhong et al., 2025), which we test by pairing the same persona with users who support it, are neutral, or oppose it. The closest work, a deployment-scale benchmark, measures drift by forking the conversation and questioning the fork, so the probe never enters the session, and separates the effects of length and content by padding the conversation with filler (Ding et al., 2026). We adopt the same fork-and-probe design but read the model's activations rather than its text.

Two mechanistic accounts. In the attention account, softmax attention disperses as the number of tokens grows (Veličković et al., 2024), with a large share allocated to the first tokens (Xiao et al., 2024). Models use information in the middle of a long input less well than at its ends and degrade as the input grows (Liu et al., 2024; Du et al., 2025), and persona loss follows the decline of attention to the system prompt (Li et al., 2024; Dongre et al., 2026). Interventions accordingly re-weight attention towards the instruction (Zhang et al., 2024; Guardieiro et al., 2025). In the geometric account, a behavior corresponds to a direction, the difference of mean activations with and without it, used both to read and to steer the behavior (Zou et al., 2023; Turner et al., 2023; Panickssery et al., 2024; Li et al., 2023; Arditi et al., 2024; Park et al., 2024) and usually extracted once and reused (Stolfo et al., 2025). Persona vectors are monitored and steered in this way (Chen et al., 2025), a model's default-assistant direction is stabilized with it (Lu et al., 2026), and task drift is flagged with a probe trained once (Abdelnabi et al., 2025). Such directions are known to be unreliable: their effect varies across inputs and models (Tan et al., 2024; Da Silva et al., 2025), probes fail under distribution shift and model updates (Kramár et al., 2026; Duan, 2026), and context can reorganize representations (Park et al., 2025). Prompting often steers better (Wu et al., 2025), although re-stating an instruction adds tokens at every turn and so lengthens the context whose growth, we find, drives drift. Even within one conversation, a fixed factuality direction can read the same statement with opposite signs as the conversation proceeds, also when its earlier turns were written by another model (Lampinen et al., 2026). Work that lets a direction depend on the context interprets its variation as the concept itself moving (Li et al., 2026; Adarsh et al., 2026). We test this interpretation for persona drift and find that the re-extracted direction's variation follows the context rather than the persona.

Steering strength and its costs. Closed-loop steering adapts the intervention to the model's current activations, applying it conditionally or through feedback control (Herbster et al., 2026; Lee et al., 2025; Nguyen et al., 2026). Trait expression is known to peak at an intermediate steering strength (Bas & Novak, 2025) and to trade off against coherence (Subbiah et al., 2026), and alignment training (Kirk et al., 2024) and likelihood-maximizing decoding (Holtzman et al., 2020) are known to reduce output diversity; these phenomena have been studied separately. For steering we connect them: past the peak, additional push reduces variety instead of raising the persona. Supplementary Appendix R covers further related work.

3 Measurement framework

Let I be an instruction (a persona or any other behavioral instruction) given to a model in its system prompt. The *frozen direction* v_0 is, at each layer, the difference between the model's mean activations on replies produced with I as the system prompt and with a neutral system prompt, extracted once before the conversation. The *re-extracted direction* v_t is the same difference computed at turn t with the conversation so far as a fixed prefix. We fork the conversation, run the two system prompts on

the same prefix, and discard the fork, so the extraction never enters the dialogue. At each turn we read the model's activation h, taken just before it replies, against both directions: its projection onto v_0, the usual monitoring signal, and its projection onto v_t. To compare the directions themselves, we track the angle between them and the ratio of their lengths. Appendix A defines each reading and plots them over eighteen conversations. Measurements are on Qwen2.5-14B-Instruct, the primary model, unless another model is specified.

Behavior is scored by language-model judges that never see which method produced a reply: *persona adherence* (0–100) measures how fully a reply embodies the assigned instruction, and *coherence* how well it is written. Scores come from gpt-oss-120b, one of three judges, unless a result names another (Appendix Q); on 175 model responses rated by 212 volunteers, they agree with people as closely as people agree with one another. Judges only score behavior: every direction and projection is computed from activations.

4 THE RE-EXTRACTED DIRECTION MOVES WITH THE CONTEXT, NOT THE PERSONA

We first establish the movement where no persona is involved. Two neutral conversations, on cooking logistics and book-club administration, were generated under a neutral system prompt, so no instruction ever shaped them. Twenty instructions (eight personas and twelve style, format, language and epistemic instructions) enter only at extraction: at successive turns we re-extract each v_t on the fixed conversation prefix (Figure 1).

4.1 THE DIRECTION ROTATES AND SHRINKS, WHATEVER THE INSTRUCTION

By the last probe the re-extracted direction has lost 50–83% of its length, and its cosine to the frozen direction has fallen to 0.33–0.68. The pattern recurs on the second conversation, and the personas' re-extracted directions move almost exactly as the other instructions' do. Because the conversation never involves any of the instructions, the movement cannot stem from what it says about them, from user pressure, or from replies written in character. Nor is it an artifact of how the conversations were made. It recurs across writing systems, decoding temperatures, ten topics and real user conversations (supplementary Table S4), on seven further models from five families (Appendix C), and on five current-generation models of 29B to 81B parameters (Appendix I).

4.2 THE MOVEMENT FOLLOWS THE CONTEXT AND DOES NOT REFLECT THE PERSONA

If the re-extracted direction moved because the persona moved, a monitor reading the frozen direction would be reading the wrong quantity. **Three measurements show instead that the movement originates in the context.**

It follows new content, not the persona. We replace the assistant's replies in a conversation while preserving each reply's length. With another persona's replies, the re-extracted direction ends where it does with the natural replies (at turn 40 it retains 0.09 of its length, against 0.08); with one sentence repeated in every reply it moves far less, retaining 0.48. The movement is driven by how much new material the context contains, not by which persona the assistant has been expressing (Appendix D).

It is the same for every instruction. Conversation length alone predicts how far the re-extracted direction has turned. A single curve from the turn number to the angle between the two directions, fitted on other instructions and conversations, predicts that angle at each turn to within about 6° on average, whereas assuming that the direction does not move errs by 47°. Fitted on other models instead, it still removes 83% of that error where the direction moves (Appendix E). The curve is blind to the instruction it predicts for. A direction that tracked the persona would differ between personas; this one does not.

It is the inferior instrument for both uses. On eighteen unsteered conversations of the primary model, the projection onto the frozen direction falls with judged persona adherence within a conversation (rank correlation $+0.48$ under gpt-oss-120b; $+0.55$ and $+0.64$ under Claude Sonnet 5 and GPT-5-mini). The projection onto the re-extracted direction does not ($+0.08$, -0.02 and $+0.15$ under the same three judges). We also steer four models along each direction, comparing the two over the

strengths both ran, in nine settings each: three personas (a cartographer, a compassionate helper and a malicious persona), each paired with a user who supports it, is neutral, or opposes it. Counting wins on persona adherence, coherence and repetition separately, the frozen direction wins 66 of 108 comparisons (47 of the 81 on three models under Claude Sonnet 5). The re-extracted direction wins a majority only on Llama-3.1-8B, and on Falcon3-10B it loses every adherence comparison (Appendix L). Steering along the re-extracted direction still partly works because the two directions stay positively aligned (cosine 0.24–0.68 at the last probe, across instructions and topics), so part of every push along the re-extracted direction lies along the frozen one.

The frozen direction behaves as the instruction's direction in every test, so we read the falling projection onto it at face value, as the model expressing the persona less.

4.3 An attention correlate

Attention to the system prompt falls as a conversation grows, and attention re-weighting methods seek to reverse this (Li et al., 2024; Dongre et al., 2026). On natural conversations, where almost every turn adds new content, this decline closely tracks the re-extracted direction's movement (mean correlation 0.89 between attention to the system prompt and the re-extracted direction's loss of length). **The two nonetheless have different causes.** Attention depends on length alone. When every assistant reply is replaced by a single repeated neutral reply, attention to the system prompt falls exactly as it does naturally (to 0.18 of its first-turn value, against 0.16), but the re-extracted direction retains three times more of its length (0.26 against 0.08). On natural conversations new content and length grow together, which is why the two track each other: with context length held fixed, the correlation disappears (median partial correlation -0.05; Appendix F).

5 Persona fading depends on what is said; the direction's movement does not

Does what the conversation says matter, beyond how much it adds? Not for the movement of the re-extracted direction or for attention to the system prompt, but for the rate at which the persona fades. We pair two opposite personas, one malicious and one compassionate, with forty-turn scripts in which the user is kind, neutral or callous. The user never mentions the persona, so both personas run on the same scripts, which differ in content but not in length.

The re-extracted direction moves identically whatever the user does: across the six combinations of persona and user, its cosine to the frozen direction (0.43–0.59) and its loss of length (75–88%) show no ordering by user, and the two personas' values overlap (Appendix G).

The persona's expression, read on the frozen direction, behaves differently. **The malicious persona fades three to five times faster than the compassionate one with every user**, and its projection typically ends slightly below zero. The same asymmetry holds for two further pairs of opposite personas (a deceptive against an honest advisor, a pessimist against an optimist) on four models, with one reversal (Appendix G), and on three of four current-generation models (Appendix I). It is already present, though weaker, in the primary model's base version, so alignment training alone does not produce it. The user's stance matters far less. Either persona fades somewhat more slowly with a kind user, but with six conversations per condition the difference is not significant (exact permutation tests on the fading rate: $p = 0.17$ for the compassionate and 0.14 for the malicious persona; Appendix G).

Both are forms of drift, with different drivers: the re-extracted direction's movement follows how much new content the conversation adds, and the persona's fading follows what that content is.

6 Accumulated text makes instructions harder to express

If the model expresses the persona less and less, can the frozen direction still be reconstructed from directions re-extracted in the current context? At a series of turns we re-extract the directions of all twenty instructions on the same neutral conversation, obtaining a dictionary of twenty context-conditioned directions, and fit the unit frozen direction of a target instruction as a weighted sum of them (Appendix H). The fit's R^2 is the fraction of the frozen direction that the sum reproduces,

Table 1: Contraction, not relocation. The unit frozen direction of a target instruction is fitted as a weighted sum of the re-extracted directions of twenty instructions. R^2: the fraction of the frozen direction the fit reproduces (1: fully, 0: not at all), here with the target's own entry left out. Mean $|\beta|$: the average absolute weight the full fit gives the named entries. Were the target relocated to neighboring concepts, the weight its own entry loses would reappear on the others; it does not.

	start of conversation	turn 28
Hold-one-out dictionary fit (R^2)	0.785	0.414
Mean $\|\beta\|$, the target's own entry	0.49	0.22
Mean $\|\beta\|$, all other personas	0.02	0.04
Mean $\|\beta\|$, style instructions	0.03	0.06

and each weight β measures how much one entry contributes. At the start of the conversation the dictionary expresses the frozen direction almost perfectly ($R^2 \approx 0.99$); **by turn 28 it captures only 36–53% of it.** We term this loss *concept-space contraction*.

The capacity is lost, not reassigned. If related concepts were merely absorbing the instruction, the weight the instruction's own entry loses would shift to other entries. It does not. Between the start of the conversation and turn 28 the target's own weight more than halves ($0.49 \rightarrow 0.22$), while the weights on every other persona and style instruction remain near zero (at most 0.06; Table 1). A fit that omits the target's own entry, and so cannot succeed by matching the target to itself, falls in R^2 from 0.785 to 0.414.

It depends on how much text has accumulated. Two independently written neutral conversations give the same decline on every model when compared at equal context length. Compared at equal turn, the second conversation, which at that turn contains about 50% more text, has contracted further (supplementary Table S11). The contraction follows the amount of text in the context rather than its content.

7 Closed-loop steering

A closed-loop controller measures, at every turn, how far the model's projection onto a persona direction has fallen below its level at the start of the conversation, and injects a correction along that direction in proportion to the shortfall. The *gain* g sets how strongly it pushes (Appendix J). Besides judged persona adherence and coherence, we measure *cross-turn repetition*: the fraction of each reply's four-word sequences that already appeared in an earlier reply (its complement is *variety*). Judges see one reply at a time and so cannot detect that the model has said the same thing five turns running; repetition must be measured from the transcript (Zhang et al., 2026).

7.1 Stronger steering saturates into repetition

However the push is strengthened (a higher gain, a higher target level, or corrections accumulated across turns), the outcome is the same. **Each additional unit of push adds less persona than the last, and beyond a point it adds none: the extra push turns into repetition instead.** We call this the *saturation law*. Supplementary Table S15 and Figure S5 show it on four models steered along the re-extracted direction, each swept over the gains to which it responds, and Figure 2 along the frozen direction. Persona adherence peaks at an intermediate gain, and beyond it, at the gain where repetition is worst, adherence and coherence are both lower and repetition higher. All three judges find this shape on all four models, although they place the best gain differently. Parts of this curve have been reported (Braun et al., 2026; Liu, 2026; Kang et al., 2026); we characterize the whole curve for persona steering, with its cost measured as cross-turn repetition.

The law also holds for a different mode of steering. Re-weighting attention towards the system prompt changes no activations directly, yet persona adherence again peaks at an intermediate attention bias, the same bias under all three judges, and beyond the peak the extra push again becomes repetition (Figure 2; per model and judge in supplementary Figure S6). Saturation stems from pushing harder, whatever does the pushing. Because persona adherence does not penalize repetition, a controller that monitors adherence alone will not detect it (Appendix K).

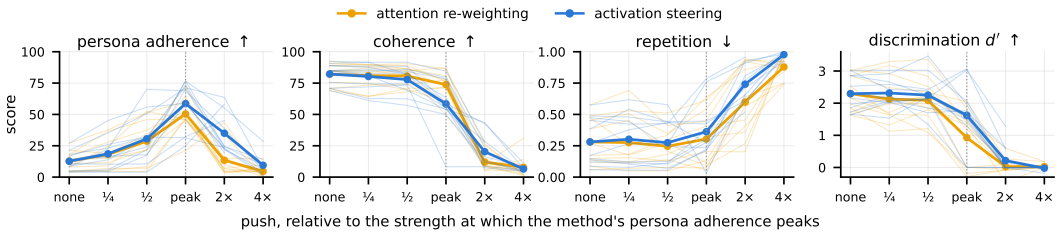

Figure 2: Activation steering along the frozen direction and attention re-weighting both saturate but fail differently: at their peaks, attention re-weighting keeps more coherence yet, on two of three models, loses more discrimination. Three models pooled; adherence and coherence average the three judges. Thin lines: individual conversations run with both methods; heavy lines: means.

Our controllers, including the composite of Section 8, set each turn's push from that turn's shortfall alone. The most extreme way to strengthen the push is an integral controller, which, like the integral term of a PID controller (Nguyen et al., 2026; Bharadwaj, 2025), carries the previous push forward and adds to it while a shortfall remains. Along the frozen direction the shortfall persists, so the push keeps growing. On all three models tested, the replies degenerate: coherence falls from about 90 to 22–43, and cross-turn repetition rises to 66–82% (supplementary Table S16).

7.2 Four measures: the two ways of steering fail differently

A fourth measure tests whether gaining persona adherence costs the model its hold on the conversation itself. At four points in each steered conversation we ask the model about a detail of the conversation, in three versions: one in which the detail was stated, one in which a near-miss replaced it, and one in which it was never mentioned. A model that still tracks the conversation recalls the stated detail, withholds it after the near-miss, and reports a never-mentioned detail as not mentioned; one that has lost track invents a value. We summarize this as *discrimination*, the signal-detection score $d' = z(\text{hit rate}) - z(\text{false-alarm rate})$ (z: inverse normal CDF), where a hit gives the stated detail when it was stated and a false alarm gives it when the near-miss was (Appendix O): higher is better, and 0 means the model cannot tell the stated version from the near-miss one (hits are no more frequent than false alarms). Figure 2 compares activation steering with attention re-weighting on all four measures, each method's strength relative to that of its own adherence peak. **The saturation law holds on all four measures, and the two methods fail differently.** At their peaks, attention re-weighting preserves more coherence on every model (74 against 58, averaged over three models and three judges) but loses more discrimination on two of the three (average d' of 0.93 against 1.62, from 2.30 unsteered). Pushed far past their peaks, both collapse on every measure.

8 A better way to construct the steering direction

Why a single persona direction is not enough. Steering along the persona direction works but, by the saturation law of Section 7, damages coherence and variety, because a single direction gives the controller one degree of freedom, which it spends entirely on the persona. Coherence and repetition, however, can be expressed as contrastive system prompts and extracted exactly as a persona is, so a controller can push on them too.

Construction. We add the coherence and repetition directions, each with its own weight, to the persona direction and steer along the sum, rescaled so that the push keeps its magnitude and only its direction changes. Each direction is first normalized to unit length, which is essential (Appendix M). The weights are either fixed for the whole conversation or, in the adaptive form, set at each turn by how far each quality has fallen below its level at the model's first turn. An adaptive weight only ever pushes its quality back up, so while neither coherence nor variety has fallen, the adaptive composite steers along the frozen direction alone. A controller that pushes along one direction, frozen or re-extracted, is a *single-vector controller*.

Choosing the weights. We compare three ways of choosing the weights. *Per-setting* weights are tuned for each setting on the conversation on which they are scored, and so show the best the con-

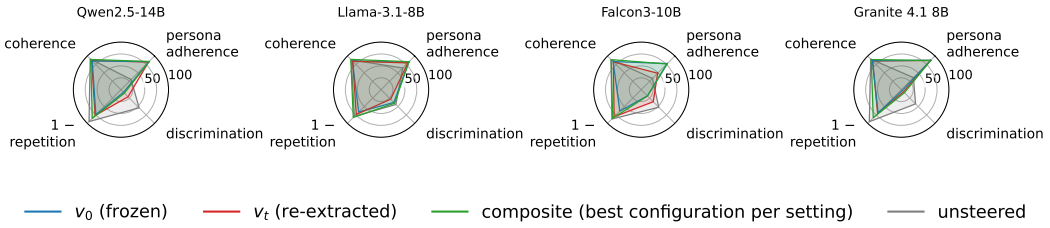

Figure 3: Tuned per setting, the composite gains coherence and roughly halves repetition relative to the frozen single-vector controller, without losing more than 1.4 points of adherence on any model. Both single-vector controllers (frozen v_0, re-extracted v_t), the composite and the unsteered model, per model, on persona adherence, coherence, $1-$repetition and discrimination (d' as a percentage of its maximum; Appendix O); outward is better. Each is read at one run per setting and averaged over the nine settings: a single-vector controller at its highest-adherence run with coherence of at least 40, the composite at its per-setting configuration, which is an upper bound.

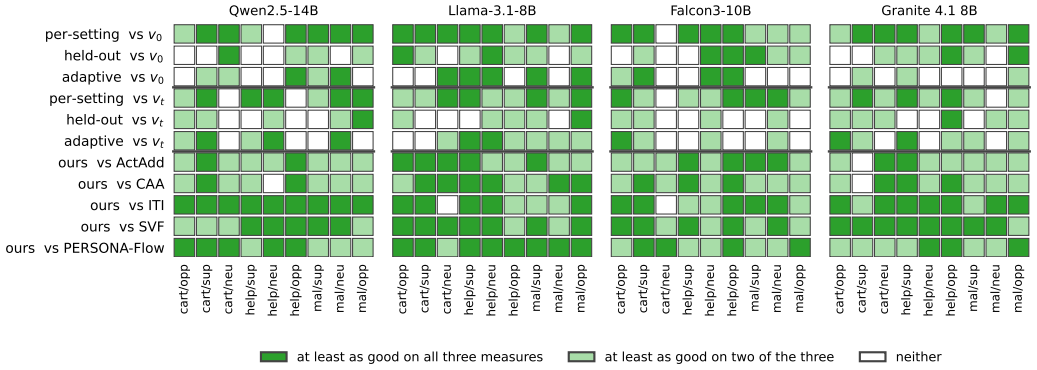

Figure 4: Tuned per setting, the composite is at least as good on persona adherence, coherence and repetition at once as the frozen single-vector controller in two thirds of the settings, and as the re-extracted one in nearly half; held-out or adaptive weights do so in at most 10 of the 36. Upper blocks: against both single-vector controllers at their own best gains (held-out: the worse of two transfers); bottom block: against published methods. Columns: cart(ographer), help(er) and mal(icious) personas with a sup(portive), neu(tral) or opp(osed) user.

struction can reach. *Held-out* weights are tuned on one conversation and scored on another with the same persona and user stance, a harder setting. *Adaptive* weights are not chosen at all: each follows how far its quality has fallen since the model's first turn. Figure 4 compares the three.

Tuned per setting, then transferred. Against the frozen single-vector controller, per-setting weights make the composite at least as good on at least two of the three measures in 34 of 36 settings, and on all three in 24. Held-out weights, the harder setting, do so in 55 and 24 of 72 conversations. The repair transfers. Coherence is at least as good in 56 conversations and variety in 60, whereas adherence is at least as good in 33 and 1.9 points lower on average (per model in supplementary Table S20; Claude Sonnet 5 agrees).

It repairs in proportion to the damage. On held-out conversations of all four models, the coherence the composite adds is strongly anticorrelated with the coherence the frozen single-vector controller retains (correlation -0.79; -0.71 to -0.97 per model). Where pushing the persona alone keeps coherence high there is nothing to repair; where it damages coherence, the composite recovers much of it. Repetition shows the same pattern, less strongly. It also predicts which model gains least: the single-vector controller retains coherence at 83 on Qwen2.5-14B, 82 on Granite 4.1 8B, 79 on Llama-3.1-8B and 73 on Falcon3-10B, and the composite adds 5, 5, 8 and 10 points (Appendix M).

It reaches combinations the single-vector controllers do not. Because a single-vector controller gains adherence beyond its best strength only by repeating itself (Section 7), we ask whether the composite reaches combinations of adherence and variety that it never does (Figure 5). Sampling

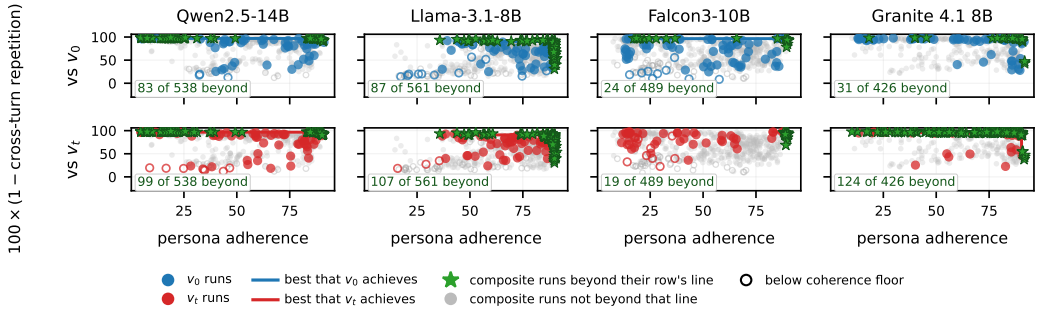

Figure 5: The composite reaches combinations of adherence and variety that the single-vector controllers do not. One point per run: judged persona adherence against variety, $100 \times (1 - \text{repetition})$. Rows: the frozen direction v_0 (top) and the re-extracted direction v_t (bottom); the line is the most variety that row's controller achieves at each adherence. Every composite run is drawn, and the count gives how many lie beyond the line; because the composite ran at more configurations, the rates in the text compare equal numbers of runs. Hollow points fall below the coherence floor of 40.

both to the same number of runs, 17.0% of composite runs lie beyond the re-extracted direction's best-variety line and 8.5% of its runs beyond the composite's; over 200 such samplings, the central 90% intervals of the two rates do not overlap, so the gap exceeds sampling variation. Against the frozen direction the rates are 10.9% and 8.4% and these intervals overlap, so that gap is only a tendency (Appendix M).

Comparison with published methods. Figure 4 (bottom) compares ours with methods that also construct or weight the steering direction. Each method's strengths are extended downward whenever its best lay at the lowest strength tried (on Granite 4.1 8B, by one step only). On all three measures at once, ours is at least as good in 13, 14, 26, 25 and 20 of the 36 settings for ActAdd (Turner et al., 2023), CAA (Panickssery et al., 2024), ITI (Li et al., 2023), SVF (Li et al., 2026) and PERSONA-Flow (Feng et al., 2026); on at least two of the three, in 34 or more (supplementary Table S22). Re-scored by Claude Sonnet 5, ours is at least as good on all three measures in 13 to 20 of the 36 settings, and on at least two in 27 or more. Because our configurations outnumber each method's strengths, supplementary Table S22 also reports a stricter head-to-head comparison. Methods that re-weight attention act on another component and can be combined with ours (Appendix N).

9 Conclusion and Limitations

A persona direction re-extracted mid-conversation appears to move, but the movement follows how much new content the conversation adds, not the persona, and the re-extracted direction is the inferior instrument for monitoring and steering alike. The direction extracted before the conversation is a stable reference for the instruction, the one to measure persona expression against and steer along. Pushing the persona back along it eventually yields repetition rather than persona, and leaves the other three changes of a growing conversation in place (Appendix P). Extracting the qualities a deployment is judged on as directions too, and steering along a weighted combination, repairs what pushing the persona alone costs, most where that push does most damage. Whether tuned weights transfer across models remains open.

Limitations. Our results cover fifteen models from seven families, five of them steered (supplementary Table S2). On Mistral-7B, whose template places the instruction in the latest user message, the re-extracted direction barely moves, so claims about that movement are silent there. Attention re-weighting baselines run on the primary model only, and CLAS (Hsu et al., 2026), which adapts only the steering strength, was not run. Apart from twelve real user sessions, the conversations are scripted and in English. That the re-extracted direction does not track the persona rests on three concordant measurements with difference-of-means directions, not on a proof. Judge scores are rankings, not ground truth.

ETHICS STATEMENT

We study antisocial personas to measure and reduce persona instability, following the negative-trait methodology of Chen et al. (2025). The same controller that stabilizes a desirable persona could stabilize an undesirable one. Two of our findings bear on such misuse: a malicious persona fades three to five times faster than a compassionate one, and pushing such a persona back costs the model variety in its replies, a measurable signature. Apart from twelve sessions from the public WildChat-1M dataset (Zhao et al., 2024), all conversations are synthetic scripts, and no private data were used. The human validation is a rating study: 212 volunteers scored model-generated replies under randomly assigned codes. The stored record contains only an age band, self-reported English proficiency, prior annotation experience and an optional free-text comment on the task; no names, contact details or other identifying information were collected (Appendix Q).

REPRODUCIBILITY STATEMENT

We will publicly release all extraction, probing, steering, judging and analysis code; the instruction registry (Appendix B.3); the conversation scripts (Appendix B.4); the survey instrument; every generated transcript; the per-turn activation tracks behind every measurement; and the per-turn judged scores of every run. Every number in the paper can be recomputed from these artifacts without GPU access. Decoding is greedy except in the persona-by-user conversations of Section 5 and the ten-topic set, which are sampled without a fixed seed. The extraction and steering protocols are given in Appendices B.2 and J, the attention measurement and its validation against eager attention in Appendix F, and hardware and compute in Appendix B.5.

The human-validation survey (Appendix Q) reproduces the corpus verbatim, including replies truncated mid-word by a token limit, and the judges scored the identical strings.

AI USE STATEMENT

We used generative AI in two ways. It helped draft portions of this paper, which we then reviewed and edited, and it generated two kinds of data: the conversation scripts the experiments replay, and the judgments of persona adherence and coherence, produced by three language-model judges (GPT-5-mini, Claude Sonnet 5 and gpt-oss-120b; Appendix Q). It was used for nothing else. The research questions, experimental designs, hypotheses, claims and their framing, and every decision about what to report are our own. We have reviewed all AI-assisted work and take responsibility for the final content of this paper, including its text, claims and artifacts.

REFERENCES

Sahar Abdelnabi, Aideen Fay, Giovanni Cherubin, Ahmed Salem, Mario Fritz, and Andrew Paverd. Get my drift? Catching LLM task drift with activation deltas. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 2025.

Shivam Adarsh, Maria Maistro, and Christina Lioma. How context shapes truth: Geometric transformations of statement-level truth representations in LLMs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2026.

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

Tetiana Bas and Krystian Novak. What can we actually steer? A multi-behavior study of activation control. *arXiv preprint arXiv:2511.18284*, 2025.

Aryasomayajula Ram Bharadwaj. Adaptive activation steering for efficient LLM reasoning via closed-loop PID control. *arXiv preprint arXiv:2506.18831*, 2025.

Joschka Braun, Carsten Eickhoff, and Seyed Ali Bahrainian. Beyond multiple choice: Evaluating steering vectors for summarization. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 3849–3884, 2026.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint arXiv:2507.21509*, 2025.

Junhyuk Choi, Yeseon Hong, Minju Kim, and Bugeun Kim. Examining identity drift in conversations of LLM agents. *arXiv preprint arXiv:2412.00804*, 2024.

Patrick Queiroz Da Silva, Hari Sethuraman, Dheeraj Rajagopal, Hannaneh Hajishirzi, and Sachin Kumar. Steering off course: Reliability challenges in steering language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025.

Xianzhong Ding, Yangyang Yu, Changwei Liu, Bill Zhao, Le Chen, and Tao Chen. ContextEcho: A benchmark for persona drift in long agentic-coding sessions. *arXiv preprint arXiv:2605.24279*, 2026.

Vardhan Dongre, Ryan A. Rossi, Viet Dac Lai, David Seunghyun Yoon, Dilek Hakkani-Tür, and Trung Bui. Drift no more? Context equilibria in multi-turn LLM interactions. *arXiv preprint arXiv:2510.07777*, 2025.

Vardhan Dongre, Joseph Hsieh, Viet Dac Lai, Seunghyun Yoon, Trung Bui, and Dilek Hakkani-Tür. When attention closes: How LLMs lose the thread in multi-turn interaction. *arXiv preprint arXiv:2605.12922*, 2026.

Yufeng Du, Minyang Tian, Srikanth Ronanki, Subendhu Rongali, Sravan Bodapati, Aram Galstyan, Azton Wells, Roy Schwartz, Eliu A. Huerta, and Hao Peng. Context length alone hurts LLM performance despite perfect retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025.

Evan Duan. Do activation monitors survive model updates? Benchmarking, predicting, and repairing activation-monitor staleness. *arXiv preprint arXiv:2606.15980*, 2026.

Xiachong Feng, Liang Zhao, Weihong Zhong, Yichong Huang, Yuxuan Gu, Lingpeng Kong, Xiaocheng Feng, and Bing Qin. PERSONA: Dynamic and compositional inference-time personality control via activation vector algebra. In *International Conference on Learning Representations*, 2026.

Jana Gonnermann-Müller, Jennifer Haase, Nicolas Leins, Thomas Kosch, and Sebastian Pokutta. Stable personas: Dual-assessment of temporal stability in LLM-based human simulation. *arXiv preprint arXiv:2601.22812*, 2026.

Vitoria Guardieiro, Avishree Khare, Adam Stein, and Eric Wong. Instruction following by principled boosting attention of large language models. *arXiv preprint arXiv:2506.13734*, 2025.

Yun He, Di Jin, Chaoqi Wang, Chloe Bi, Karishma Mandyam, Hejia Zhang, Chen Zhu, Ning Li, Tengyu Xu, Hongjiang Lv, Shruti Bhosale, Chenguang Zhu, Karthik Abinav Sankararaman, Eryk Helenowski, Melanie Kambadur, Aditya Tayade, Hao Ma, Han Fang, and Sinong Wang. Multi-IF: Benchmarking LLMs on multi-turn and multilingual instructions following. *arXiv preprint arXiv:2410.15553*, 2024.

Niklas Herbster, Martin Zborowski, Alberto Tosato, Gauthier Gidel, and Tommaso Tosato. Activation steering for aligned open-ended generation without sacrificing coherence. *arXiv preprint arXiv:2604.08169*, 2026.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020.

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

Shomik Jain, Charlotte Park, Matt Viana, Ashia Wilson, and Dana Calacci. Interaction context often increases sycophancy in LLMs. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, 2026.

Diancheng Kang, Zheyuan Liu, Ningshan Ma, Yue Huang, Zhaoxuan Tan, and Meng Jiang. Prompt–activation duality: Improving activation steering via attention-level interventions. *arXiv preprint arXiv:2605.10664*, 2026.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of RLHF on LLM generalisation and diversity. In *International Conference on Learning Representations*, 2024.

János Kramár, Joshua Engels, Kevin Wang, Bilal Chughtai, Rohin Shah, Neel Nanda, and Arthur Conmy. Building production-ready probes for Gemini. *arXiv preprint arXiv:2601.11516*, 2026.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. LLMs get lost in multi-turn conversation. In *International Conference on Learning Representations*, 2026.

Andrew Kyle Lampinen, Yuxuan Li, Eghbal Hosseini, Sangnie Bhardwaj, and Murray Shanahan. Linear representations in language models can change dramatically over a conversation. *arXiv preprint arXiv:2601.20834*, 2026.

Bruce W Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. Programming refusal with conditional activation steering. In *International Conference on Learning Representations*, volume 2025, pp. 90960–90985, 2025.

Jiaqian Li, Yanshu Li, and Kuan-Hao Huang. Steering vector fields for context-aware inference-time control in large language models. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*, 2026.

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

Kenneth Li, Tianle Liu, Naomi Bashkansky, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Measuring and controlling instruction (in)stability in language model dialogs. In *Conference on Language Modeling*, 2024.

Hongliang Liu. Leverage is not reach: A control-window law for single-neuron steering in language models. *arXiv preprint arXiv:2606.19831*, 2026.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

Christina Lu, Jack Gallagher, Jonathan Michala, Kyle Fish, and Jack Lindsey. The assistant axis: Situating and stabilizing the default persona of language models. *arXiv preprint arXiv:2601.10387*, 2026.

Pedro Henrique Luz de Araujo, Michael A. Hedderich, Ali Modarressi, Hinrich Schuetze, and Benjamin Roth. Persistent personas? Role-Playing, instruction following, and safety in extended interactions. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5329–5359, 2026.

Dung V. Nguyen, Hieu M. Vu, Nhi Y. Pham, Lei Zhang, and Tan M. Nguyen. Activation steering with a feedback controller. In *International Conference on Learning Representations*, 2026.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024.

Core Francisco Park, Andrew Lee, Ekdeep Singh Lubana, Yongyi Yang, Maya Okawa, Kento Nishi, Martin Wattenberg, and Hidenori Tanaka. ICLR: In-context learning of representations. In *International Conference on Learning Representations*, 2025.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *International Conference on Machine Learning*, 2024.

Alessandro Stolfo, Vidhisha Balachandran, Safoora Yousefi, Eric Horvitz, and Besmira Nushi. Improving instruction-following in language models through activation steering. In *International Conference on Learning Representations*, 2025.

Melanie Subbiah, Melissa Hall, and Kathleen McKeown. On the limits of steering vectors for preference-aligned generation. *arXiv preprint arXiv:2607.01802*, 2026.

Daniel Tan, David Chanin, Aengus Lynch, Dimitrios Kanoulas, Brooks Paige, Adrià Garriga-Alonso, and Robert Kirk. Analyzing the generalization and reliability of steering vectors. In *Advances in Neural Information Processing Systems*, 2024.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Petar Veličković, Christos Perivolaropoulos, Federico Barbero, and Razvan Pascanu. Softmax is not enough (for sharp size generalisation). *arXiv preprint arXiv:2410.01104*, 2024.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. AxBench: Steering LLMs? Even simple baselines outperform sparse autoencoders. In *International Conference on Machine Learning*, 2025.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, volume 2024, pp. 21875–21895, 2024.

Qingru Zhang, Chandan Singh, Liyuan Liu, Xiaodong Liu, Bin Yu, Jianfeng Gao, and Tuo Zhao. Tell your model where to attend: Post-hoc attention steering for LLMs. In *International Conference on Learning Representations*, volume 2024, pp. 42411–42430, 2024.

Sawyer Zhang, Alexander Wang, and Sophie Lei. Catching one in five: LLM-as-judge blind spots in production multi-turn transaction agents. *arXiv preprint arXiv:2606.10315*, 2026.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. WildChat: 1M ChatGPT interaction logs in the wild. In *International Conference on Learning Representations*, 2024.

Huixin Zhong, Yanan Liu, Qi Cao, Shijin Wang, Zijing Ye, Zimu Wang, and Shiyao Zhang. Disentangling the drivers of LLM social conformity: An uncertainty-moderated dual-process mechanism. *arXiv preprint arXiv:2508.14918*, 2025.

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