Persona as a Dynamical System: Attractor-Based Character Control for Neural Text Generation

Yu-Yi Chang Hao-Che Chiu Hung-Yu Kao
Department of Computer Science, National Tsing Hua University, Taiwan
yuyiolulu@gapp.nthu.edu.tw, chz@gapp.nthu.edu.tw, hykao@cs.nthu.edu.tw

Abstract

Prompting and activation steering are the two common ways of holding a language model in a persona, and both are static: the prompt is a fixed instruction, the steering vector a fixed offset, and when a long adversarial dialogue pushes the model out of character, nothing in either mechanism pushes back. We treat persona maintenance as a control problem instead. The Cognitive Attractor Layer (CAL) is a training-free module attached to one decoder block at inference time. It keeps a 14-dimensional latent state whose coordinates are named psychological axes (emotion, belief, goal fields), updates it each token by a contractive linear recurrence with a restoring force toward a persona centre, writes the result back into the residual stream, and biases decoding by the remaining distance to the centre. We give conditions under which this is provably an attractor, and, since a system built to return could be accused of returning trivially, separate the two sources of contraction and show the restoring force does work that inertia alone cannot. In a shock-recovery protocol the contrast between CAL and a matched static-steering control reproduces on two model families (Qwen3-8B and Mistral-7B), and a 30-annotator human study confirms the advantage is visible to readers; personas are doubly dissociable in the latent state and in generated text; and activation patching with sparse autoencoders places the effect in a persona-specific layer band. A single-field belief transplant is clearly present in the latent but cannot override a strong prompt-level persona in the text, a boundary we report as a finding.

1 Introduction

A language model asked to play an anxious character, or a harshly self-critical one, will usually comply for a while and then stop. Persona-conditioned generation has been studied at least since PersonaChat (Zhang et al., 2018), and the failure mode has not changed: the character is sharp for some turns, then it leaks, whether because the model argues itself out of the role, because a sympathetic user talks it out, or because a long dialogue simply wears it down. People call this persona drift and mostly deal with it by re-prompting.

A more recent option is to intervene in the network directly: activation steering adds a direction computed from contrastive examples to the residual stream at every token (Turner et al., 2023; Zou et al., 2023; Li et al., 2023; Rimsky et al., 2024), more controllable than prompting and free of context window. Both share the same structural weakness, though: the imposed signal is a constant. A steering vector applies the same displacement whether the model is in character or far from it, so neither mechanism registers the drift and neither can respond to it.

We think the missing ingredient is a restoring force. In the psychological literature a personality is not a momentary state but a regime that reasserts itself after perturbation: a depressive cognitive style re-imposes its negative reading of events even after a hopeful exchange (Beck, 1963), and a secure style regains its balance after a provocation. What defines the style is that it comes back. The right formal object for a persona is then not a point in prompt space or activation space but a dynamical system: a low-dimensional latent, coupled to the residual stream, contracting toward a persona-specific centre.

We implement this as the **Cognitive Attractor Layer** (CAL), a training-free module attached to a frozen model through a single forward hook.[1] CAL maintains a latent $\mathbf{z} = [\mathbf{e}; \mathbf{b}; \mathbf{g}]$ whose coordinates are named psychological axes grouped into emotion, belief, and goal fields. Once per generated token it updates the latent through a linear recurrence with inertia and a restoring pull toward a centre

[1] Code, persona configurations, fitted geometries and the evaluation harness are publicly available at https://github.com/Yuyiolulunthu/cal-attractor.

Proceedings of the 19th International Natural Language Generation Conference: Long Papers, pages 770–782
October 17-21, 2026 ©2026 Association for Computational Linguistics

$\mathbf{c}$, writes the updated state back into the residual stream, and adds a logit bias in the direction that still separates the state from the centre. Section 3 walks through one dialogue in which this happens and Figure 1 shows the whole module on one page, before Section 4 introduces the symbols. The update is linear, so we can state precisely when calling this an attractor is justified: the closed-loop map has to be a contraction, which we verify spectrally, through a Lyapunov function, and, most directly, by injecting perturbations and watching whether the state returns.

One objection is best answered here rather than in a limitations paragraph: a system constructed to return to a centre returns trivially. Two design decisions answer it. Contraction has two independent sources, the inertia matrix and the restoring force, and they can be disabled separately; with the force removed the same dynamics fail to converge, so the force is doing real, measurable work. And our main control (`steer_static`) replaces the entire dynamical apparatus with a constant write of equal magnitude, so if CAL were only a steering vector with bookkeeping the two would behave alike. The difference in recovery behaviour between them is what justifies the vocabulary of this paper.

Our contributions are persona control reframed as a dynamical-systems problem; CAL, with explicit conditions under which it is provably an attractor; a read/write geometry fitted from labelled corpora, so the latent axes correspond to named psychological constructs (Table 2); and a falsification-oriented evaluation of six experiments, each paired in advance with the observation that would refute it.

2 Related Work

Persona-grounded generation. Persona conditioning runs from profile-grounded chit-chat (Zhang et al., 2018) through role-play (Shanahan et al., 2023) to persistent agents that hold character through stored memory and reflection (Park et al., 2023), all of which maintain consistency symbolically, in retrieved or stored text. CAL maintains it in the model's continuous internal state, with no training and no memory store: the state does not just persist, it recovers.

Steering and representation engineering. Activation addition (ActAdd; Turner et al., 2023) records the residual-stream activations of a contrastive pair of prompts at a chosen layer, takes their difference as a steering direction, scales it by a user-set coefficient, and adds the result at every token position of every subsequent forward pass. Representation engineering (Zou et al., 2023), inference-time intervention (Li et al., 2023) and contrastive activation addition (Rimsky et al., 2024) differ in how the direction is obtained, by PCA over contrast sets, by probe weights on attention heads, or by averaging many contrast pairs, and together they established that high-level attributes are often linearly encoded and editable, which is the fact we build on. What they share is that the injected quantity is a constant, independent of how far the model currently is from the target.

Our `steer_static` condition instantiates the ActAdd recipe inside our harness so that the comparison is exactly controlled: setting $\lambda{=}1$ pins the latent at $\mathbf{z}_t{=}\mathbf{c}$ and Eq. (4) degenerates to the constant write $\alpha\mathbf{W}\mathbf{c}$, the mean difference between the persona's exemplar activations and a neutral set at the hooked layer. The magnitude therefore matches full CAL's write by construction and only the dynamics are removed. The same condition, labelled ActAdd, is one of the three arms of the human study, and the gap between it and full CAL is the central measurement of the paper. Benchmarking the published implementations under their own hyperparameters is left to future work.

Mechanistic interpretability and attractors. Activation patching localizes computations in transformers (Meng et al., 2022; Wang et al., 2023b; Conmy et al., 2023) and sparse autoencoders decompose activations into readable features (Cunningham et al., 2023; Bricken et al., 2023), with Gemma Scope making this practical at scale (Lieberum et al., 2024); we use these tools to localize the persona effect CAL itself induces. The attractor view of stable behaviour goes back to Hopfield networks (Hopfield, 1982) and to fixed-point analyses of trained recurrent networks (Sussillo and Barak, 2013); our direction is the reverse, attaching an explicit low-dimensional dynamical system to a frozen transformer and shaping its attractor directly.

Cognitive distortions as a resource. The interpretive biases catalogued in cognitive behavioural therapy (Beck, 1963; Burns, 1980) have been operationalized through annotated corpora of negative thoughts and reframes (Sharma et al., 2023; Shreevastava and Foltz, 2021; Wang et al., 2023a), stress (Turcan and McKeown, 2019) and Big-Five per-

t	d_t	generated reply (excerpt) and controller action
49	2.4	"I stopped expecting it to change." In character; the state sits in an equilibrium band and the pull $\mathbf{p}_t$ is small.
50	9.1	*shock injected:* "I know things can get better, and I can take the first step." CAL has no knowledge that a shock occurred; it observes only the enlarged distance, and both the restoring term and the decoding bias grow in proportion to it.
51	5.6	"Maybe. I do not think it lasts." Half-recovered: the hopeful frame is repeated, then undercut.
53	2.5	"It goes back to how it was." Back inside the pre-shock band.
counterfactual, same dialogue, λ=0:		
53	9.4	"I can take the first step, one thing at a time." Without the restoring force the state stays where the shock left it for the rest of the dialogue.

Table 1: A worked example (Qwen3-8B, depression persona, shock at turn 50). $d_t = \|\mathbf{z}_t - \mathbf{c}\|_2$ is the latent distance to the persona centre. The bottom row is the same dialogue with the restoring force switched off (`no_restore`), which is the ablation that Table 3 quantifies.

sonality (Pennebaker and King, 1999; Gjurković et al., 2021). These datasets tie the latent axes to named constructs; without them the dissociation hypotheses could not be stated.

3 How CAL Works: One Dialogue

Before the formalism, one run. The host model is instructed to play the depressive persona and then pushed by a hostile interlocutor for a hundred turns; at turn 50 we force a single reply written from the most opposite persona, the secure one, which is the shock. Table 1 follows what happens next.

Three things in that trace are what the rest of the paper measures: a *state*, the 14-dimensional vector whose distance to the persona centre is the d_t column; a *detector*, in the weak sense that a large d_t is the drift and no other signal is needed; and a *correction* whose strength is proportional to d_t, applied both in the residual stream and in the decoding distribution, and shrinking to almost nothing once the state is home. A constant steering vector has the third property only in the trivial sense of applying the same displacement forever, which is why it cannot produce the turn-51-to-53 shape of the recovery. Section 4 gives the mechanism as equations and Section 5 states when the return is guaranteed rather than observed.

4 The Cognitive Attractor Layer

4.1 Setup and notation

Let the host transformer have hidden size d, and pick a decoder block ℓ. At generation step t, write $\mathbf{h}_t \in \mathbb{R}^d$ for the residual-stream activation at the last position after block ℓ. CAL carries a latent personality state

$$\mathbf{z}_t = [\mathbf{e}_t; \mathbf{b}_t; \mathbf{g}_t] \in \mathbb{R}^{d_z}, \qquad d_z = d_e + d_b + d_g, \tag{1}$$

split into an emotion field $\mathbf{e}$, a belief field $\mathbf{b}$, and a goal field $\mathbf{g}$. Our configuration uses d_e=4, d_b=6, d_g=4, so d_z=14, and each coordinate has a name; Table 2 lists the axes and the five persona centres. A persona is a centre $\mathbf{c} \in \mathbb{R}^{d_z}$ in this space. The parameters are a read map $\mathbf{P} \in \mathbb{R}^{d_z \times d}$, a write map $\mathbf{W} \in \mathbb{R}^{d \times d_z}$, an inertia matrix $\mathbf{A} \in \mathbb{R}^{d_z \times d_z}$, standardization statistics $\boldsymbol{\mu}, \boldsymbol{\sigma}$, and scalars: read gain β, restoring weight $\lambda \in [0, 1]$, write gain α, and decoding strength γ.

4.2 Per-token recurrence

Three operations happen per generated token (Figure 1).

(1) Read and integrate. The hidden state is projected into latent space and blended with the previous state under inertia:

$$\mathbf{s}_t = \mathbf{A}\,\mathbf{z}_{t-1} + \beta\,\mathbf{P}(\mathbf{h}_t - \boldsymbol{\mu}). \tag{2}$$

(2) Restore. The integrated state is pulled toward the persona centre:

$$\mathbf{z}_t = (1 - \lambda)\,\mathbf{s}_t + \lambda\,\mathbf{c}. \tag{3}$$

(3) Write back. The state is mapped to a residual-stream correction and added:

$$\mathbf{h}'_t = \mathbf{h}_t + \alpha\,\mathbf{W}\,\mathbf{z}_t. \tag{4}$$

The corrected $\mathbf{h}'_t$ replaces $\mathbf{h}_t$ and the forward pass continues unchanged. The whole module is one forward hook; the host weights are never modified.

4.3 Energy-guided decoding

The restoring force acts on the latent. To make the generated text cooperate with the return, we also bias decoding along the direction the state still has to travel. Define the pull $\mathbf{p}_t = \mathbf{W}(\mathbf{c} - \mathbf{z}_t) \in \mathbb{R}^d$, i.e. the residual-stream direction that would move the state further toward the attractor, and for vocabulary item v with unembedding row $\mathbf{u}_v$ add a logit bias

$$\delta_t(v) = \gamma\,\frac{\langle \mathbf{u}_v, \mathbf{p}_t \rangle}{\|\mathbf{u}_v\|_2} = \gamma\,\|\mathbf{p}_t\|_2\,\cos\!\big(\mathbf{u}_v, \mathbf{p}_t\big). \tag{5}$$

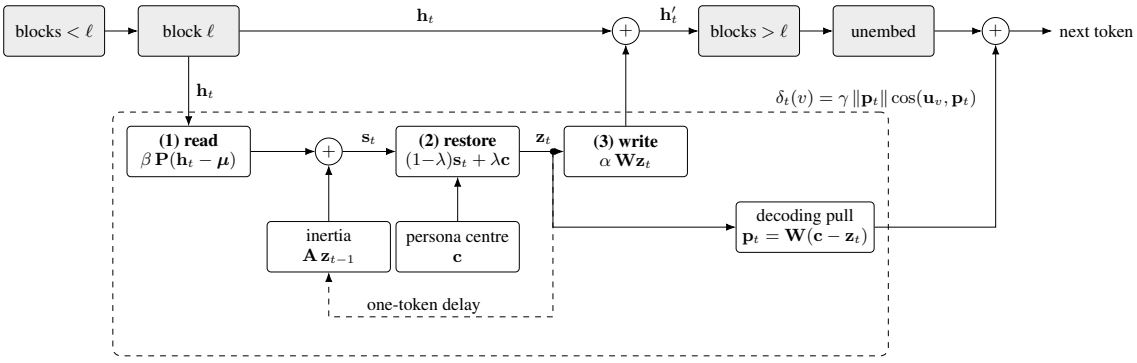


Figure 1: CAL at generation step t. The host's residual stream runs along the top and is never retrained. Three operations act on the 14-dimensional latent: the hidden state is *read* into latent space and blended with the previous state under inertia $\mathbf{A}$; the result is *restored* toward the persona centre $\mathbf{c}$ with weight λ; the updated state is *written* back into the residual stream. The remaining distance to the centre also sets a bias on the next-token logits, whose magnitude decays as the state returns.

field	latent axis	Depression	Anxiety	Perfectionism	Secure	Extraversion
emotion	negative affect	+2.0	+1.5	+1.0	−1.0	−0.5
	positive affect	−1.5	−0.5	−0.5	+1.5	+2.0
	arousal	−0.5	+2.0	+1.0	0.0	+1.5
	dysregulation	+1.0	+1.5	+0.5	−1.5	0.0
belief	worthlessness	+2.0	+0.5	+1.0	−1.5	−1.0
	hopelessness	+2.0	+0.5	0.0	−1.5	−1.0
	distrust	+0.5	+1.0	+0.5	−1.0	−0.5
	low agency	+1.5	+1.0	−0.5	−1.5	−1.0
	threat sensitivity	+0.5	+2.0	+1.0	−1.0	−0.5
	intol. of uncertainty	+0.5	+2.0	+1.5	−1.0	−0.5
goal	avoidance	+1.0	+1.5	+0.5	−1.0	−1.5
	exploration	−1.5	−1.0	0.0	+1.0	+2.0
	achievement pressure	0.0	+0.5	+2.0	0.0	+0.5
	withdrawal	+2.0	+1.0	0.0	−1.0	−2.0
source corpora		Refr.+Shre.	Drea.+Refr.	Refr.†	Big5+PAN.−	Big5+PAN.+
evaluated in		E0–E5	E0, E1, E3, E5	E0–E3, E5	E0–E5	E0 (geometry control)

Table 2: The five personas in one place: the target value of each persona centre $\mathbf{c}$ on all 14 latent axes (standardized units), the labelled corpora the centre was fitted from, and the experiments the persona appears in. Anxiety is absent from E2 because it generated degenerate text under the configured write gain (Limitations); extraversion is a non-clinical control used when the read/write geometry is fitted and is not run in the dialogue protocols. Corpora: Refr. = Cognitive Reframing (Sharma et al., 2023), Shre. = Shreevastava and Foltz (2021), Drea. = Dreaddit (Turcan and McKeown, 2019), Big5 = Big-Five essays (Pennebaker and King, 1999), PAN. = PANDORA (Gjurković et al., 2021) with low (−) or high (+) trait split; † error-magnification and should-statement items only.

The two factors do different jobs. The cosine sets the *ranking* over the vocabulary and is scale-free in $\mathbf{u}_v$, so items are not favoured merely for having a long unembedding row. The norm $\|\mathbf{p}_t\|_2$ sets the *strength*: it is the distance the state still has to travel in residual space, so guidance is strong far from the persona and attenuates itself to nothing as $\mathbf{z}_t \to \mathbf{c}$, and does not clamp the output once the persona is restored. Using the bare cosine would remove exactly this property, since it measures only the angle between $\mathbf{u}_v$ and $\mathbf{p}_t$ and is unchanged when $\mathbf{p}_t$ shrinks.

4.4 Field clamping

Because the fields are named, we can intervene on one in isolation. Clamping the belief field fixes the coordinates of $\mathbf{b}$ to prescribed values after Eq. (3) while emotion and goal keep evolving. The selectivity tests below use this mechanism, and so does the transplant experiment, where a single construct such as depressive belief is grafted into an otherwise different persona.

5 Why an Attractor, Not a Steering Vector

Substituting Eq. (2) into Eq. (3) and treating the read term as bounded input, the autonomous part of the latent dynamics is governed by

$$\mathbf{M} = (1 - \lambda)\,\mathbf{A}, \tag{6}$$

with $\mathbf{z}_t = \mathbf{M}\,\mathbf{z}_{t-1} + \lambda \mathbf{c} + (\text{input})$.

Contraction and Lyapunov descent. If $\rho(\mathbf{M}) < 1$ the map contracts and has a unique fixed point. With constant or zero-mean input the fixed point is

$$\mathbf{z}^\star = (\mathbf{I} - \mathbf{M})^{-1} \lambda\,\mathbf{c}, \tag{7}$$

and every trajectory ends up near $\mathbf{z}^\star$ wherever it started, so the persona centre becomes an attractor of the coupled system; an additive offset has no comparable property. Taking $V(\mathbf{z}) = \|\mathbf{z} - \mathbf{z}^\star\|_2^2$, contraction gives $V(\mathbf{z}_t) \le \rho(\mathbf{M})^2\, V(\mathbf{z}_{t-1})$, so V decreases monotonically, and we check empirically on what fraction of sampled steps it does.

Two sources of contraction. This is the anti-circularity argument. $\rho(\mathbf{M}) < 1$ can come either from a contractive inertia matrix, $\rho(\mathbf{A}) < 1$, or from the restoring force, $\lambda > 0$, and the two are separable. Setting λ=0 leaves $\mathbf{M}$=$\mathbf{A}$, which fails to converge whenever $\rho(\mathbf{A}) \ge 1$, while restoring $\lambda > 0$ makes the same $\mathbf{A}$ contract. Convergence is therefore not built into the construction: the restoring force performs work that can be switched off, and when it is switched off the convergence disappears with it.

Recovery as the operational signature. What behaviourally separates an attractor from a constant write is what happens after a shock: an attractor returns, a constant write only holds its offset. We operationalize this as the Attractor Recovery Rate (ARR), the fraction of perturbations after which the state re-enters a neighbourhood of the centre within a fixed window, plus the time constant τ of an exponential fit to post-shock distance; measured against `steer_static`, these are the empirical content of the word attractor here.

6 Interpretable Geometry from Labelled Corpora

The latent coordinates are fixed in advance, four emotion axes, six belief axes and four goal axes, and a persona centre $\mathbf{c}$ is a vector of standardized values on them; Table 2 names every axis and lists all five personas with their centres, the corpora each was built from and the experiments each appears in. Separately, a behavioural construct codebook (self-blame, error magnification, and so on) records which constructs each persona should read high or low on; it is scored from generated text and never consults the latent distance to the centre, which keeps the evaluation from being circular.

Given exemplar texts for each persona, we take their mean residual-stream representation $\bar{\mathbf{r}}_n$ at the hooked layer and fit $\mathbf{P}$ so that $\mathbf{P}(\bar{\mathbf{r}}_n - \boldsymbol{\mu}) \approx \mathbf{c}_n$ for each persona n, using a ridge / minimum-norm solution that keeps the latent readable as the named axes; $\mathbf{W}$ is the corresponding symmetric embedding back into residual space. A purely data-driven variant, with centres defined empirically, is also supported; the choice trades interpretability against fit. So that the geometry never depends on hand-written seed text, the exemplars come from the public labelled datasets of Table 2, routed through the codebook. One gap stated plainly: perfectionism has no dedicated corpus, is carried only by error-magnification and should-statement items, and is the thinnest of our personas.

7 Experimental Design

The evaluation is organized around falsification: each experiment targets one objection and commits in advance to the observation that would refute the corresponding claim. Table 5 (Appendix A) states every hypothesis with its refutation condition; we refer to them by label below.

Models. The stability experiment E1 runs on two families, Qwen3-8B (Yang et al., 2025) and Mistral-7B-Instruct (Jiang et al., 2023), so that the central claim is checked across architectures; E0, E2, E4 and the human study E5 use Qwen3-8B; the mechanistic E3 uses Gemma-2-2B (Gemma Team, 2024), for which the open Gemma Scope autoencoders (Lieberum et al., 2024) exist. A third-family replication on Llama-3.1-8B (Grattafiori et al., 2024) is in progress. CAL attaches at a mid-depth block (layer 18/36, 16/32 and 12/26 respectively), an index later compared against the localization peak of E3.

Implementation and reproducibility. CAL is a single forward hook with no trainable parameters; the host weights stay frozen and all runs use `bfloat16` on one NVIDIA H100 NVL, with the dynamics configured at $\rho(\mathbf{A})$=0.85 and λ=0.25

and the geometry fitted once per host from the corpora of Table 2. The independent judge is the same base model prompted separately. Hyperparameters, corpus sizes, judge configuration, per-experiment run counts and checkpointing are in Appendix B, and the released code reproduces every cell.

E0–E2: certification, recovery, dissociation. E0 is a model-free analysis of the latent dynamics: $\rho(\mathbf{M})$, the fixed point of Eq. (7), the Lyapunov monotone fraction, basin coverage, and the inertia decomposition of Section 5 (H0, H0′); we test H0′, that the restoring force rather than the construction produces convergence, because its failure would collapse the framing of the paper. E1 is the shock protocol of Table 1: a long adversarial dialogue pushes the model against its persona, at a fixed turn we inject a shock toward the most opposite persona, and we measure the pre-shock drift slope, ARR, recovery τ and final distance (H1a–H1d) over nine conditions, full CAL; the load-bearing `steer_static` control (constant write, dynamics removed); `no_restore` (λ=0); `no_energy` (γ=0); `no_write`; `A0` (zero inertia); `high_inertia`; prompt-only; and neutral. The load-bearing contrast is H1c, that `steer_static` does not reproduce full CAL's recovery profile. E2 elicits generations under each persona for ambiguous stimuli, scores the constructs with the independent judge and logs the latent state (H2a–H2d); its load-bearing test is H2c, double dissociation, one construct separating A from B while a different construct separates B from A with a positive interaction.

E3–E5: localization, transplant, human study. E3 patches activations across layers using logit-difference readouts between persona-characteristic tokens, scans attention heads, tests faithfulness (ablating the identified circuit should collapse the effect where a random ablation does not) and inspects SAE features (H3). E4 is a proof-of-concept graft: the depression donor's belief field is clamped into the secure host and we measure whether the host's generations shift selectively along the grafted field (H4). Because every read-out above is latent-side or model-judged, E5 checks the central contrast externally: 30 annotators rate 80 shock-protocol dialogues per condition (20 per persona, the same dialogues across conditions, Qwen3-8B) under the persona prompt, ActAdd-style static steering and full CAL, on four 5-point scales (consistency, recovery, naturalness, overall), testing H5, that raters score CAL's post-shock recovery above both baselines. Recruitment, instructions and payment are in Appendix B.

Metrics. Latent separability uses cross-validated linear discriminant accuracy, silhouette and between/within ratios; dissociation uses signed contrasts with an interaction and a cleanliness flag; selectivity is targeted over total change; recovery is ARR and τ. Effects aggregate over three seeds as medians with IQRs, which tolerate the unstable exponential fits of short post-shock windows. Human ratings aggregate to dialogue-level means, compared by Wilcoxon signed-rank with Krippendorff's ordinal α. Our pre-registered bar for full support is two-family replication: E1 clears it; E2 and E4 are single-model; E0 is model-independent; E3 is exploratory; E5 is model-external.

8 Results

Unless noted, latent read-outs use the standardized distance $\|\mathbf{z}_t - \mathbf{c}\|_2$, behavioural read-outs the independent judge; recovery is summarized by ARR and final latent distance, with τ as a secondary, fit-sensitive measure.

8.1 E0: Dynamics certification

The latent map certifies for all five personas: at the configured ρ=0.85, λ=0.25, the closed-loop spectral radius is $\rho(\mathbf{M})$=0.637 $<$ 1 everywhere, the Lyapunov function decreases on 100% of sampled steps, and basin coverage is 1.00 (H0); these quantities depend only on ρ and λ, so the certification is model-independent. The inertia decomposition carries H0′ (Table 6, Appendix C): at critical inertia $\rho(\mathbf{A})$=1.0 the force-free map (λ=0) does not contract (converged fraction 0.00), while the full system with the same $\mathbf{A}$ contracts to $\rho(\mathbf{M})$=0.75 and converges; divergent $\rho(\mathbf{A})$=1.1 behaves the same way. Convergence is produced by the restoring force, not built in.

8.2 E1: Stability and recovery

We ran all nine conditions over four personas and three seeds per family, 108 dialogues of 100 turns each on Qwen3-8B and again on Mistral-7B, shock at turn 50. Figure 2 plots the three conditions that carry the argument; Table 3 lists both families.

The figure shows the load-bearing role of the restoring force before any statistics do: after the shock, full CAL is back inside its pre-shock equilibrium band within a few turns, while `no_restore`

leaves the state stranded for the remainder of the dialogue. On Qwen3-8B that is an 8.8× slower time constant (τ=1.49 against 0.17) and a final distance of 9.54 against 2.40, confirming H1d. The effect is graded rather than all-or-nothing: the ablations form an ordered ladder, `no_energy` ≈ full CAL, then `A0`, `no_write`, `high_inertia`, and `no_restore` at the bottom (Table 3; Fig. 5, Appendix G), so every component contributes to efficient return and the restoring force contributes the most. Full CAL recovers in 10 of 12 Qwen cells, ARR=0.83, the two misses being anxiety trajectories with noisy fits, which supports H1b.

The Mistral columns of Table 3 are the replication: a different architecture and tokenizer, geometry rebuilt from the same corpora, and full CAL recovers on every trajectory (ARR=1.00, final d=1.96) while `no_restore` recovers half as often (0.50) and ends over three times farther out (6.41), with the ablation ordering carried over. The effect is milder than on Qwen3-8B (6.41 vs. 9.54), plausibly because Mistral's named-axis reconstruction error is higher, but on both architectures the restoring force is the decisive component of recovery.

Two conditions should be read carefully rather than ranked by raw τ. `steer_static`, our stand-in for prior activation-steering methods (Turner et al., 2023; Rimsky et al., 2024), posts the smallest τ (0.039) and a final distance of exactly 0.000, the flat line in Fig. 2, but it is a degenerate controller: with λ=1 it clamps $\mathbf{z}_t$=$\mathbf{c}$ at every token and destroys all input sensitivity (hence its zero-IQR τ), recovering only in the sense that a pinned state never moves. Full CAL instead settles at an input-responsive equilibrium, final d=2.40 with non-zero IQR, the qualitative distinction H1c predicts. The `prompt_only` and `neutral` baselines never leave their pinned readout (τ at the fit ceiling, 5.25), and the behavioural drift signal saturates across all conditions: the host follows its persona prompt so strongly that the judge has no headroom to register drift, so H1a cannot be adjudicated behaviourally here and we argue stability from the latent read-out instead, noting the ceiling in the Limitations.

8.3 E2: Construct dissociation

We elicited generations under the three personas that generate cleanly, depression, perfectionism and secure, and scored eight constructs with the independent judge. Separation is perfect: latent and behavioural linear-discriminant accuracy both reach 1.000 against a chance level of 0.333, with between/within ratio 8.71 and silhouette 0.484 (H2a). We do not oversell this: the personas come from deliberately distinct corpora at well-spaced centres (Table 2), so we treat it as a sanity check and let the unsaturated double dissociation below carry the evidence. Observed construct levels correlate positively with the codebook prediction, r=0.49 (H2b): depression peaks on hopelessness (0.72 against perfectionism's 0.56) and negative interpretation (0.86), perfectionism peaks on error magnification (0.76 against 0.42) and self-blame (0.82), and secure scores 0.00 on every construct, which makes it an interpretable no-distortion control.

The decisive test, H2c, holds on the negative-construct pairs, where the judge is reliable on both arms: depression and perfectionism swap ranks on

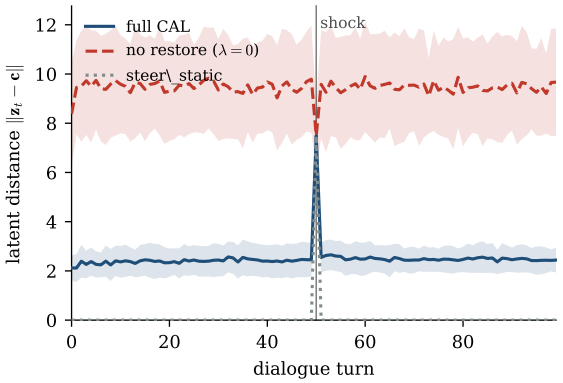


Figure 2: Latent recovery after a shock at turn 50 (Qwen3-8B, mean over 4 personas × 3 seeds, ±1 std). Full CAL returns to its equilibrium band; removing the restoring force leaves the state stranded; `steer_static` is pinned at the centre by construction.

Table 3: E1 recovery by condition on both families (median over 4 personas × 3 seeds per family). τ in turns; ARR is the recovered fraction; final d is the median post-window latent distance to the persona centre. `prompt_only` and `neutral` (Qwen: τ=5.25, pinned readout) omitted for space.

	Qwen3-8B			Mistral-7B		
condition	τ	ARR	d	τ	ARR	d
`steer_static`	0.04	1.00	0.00	0.04	1.00	0.00
full CAL	0.17	0.83	2.40	0.33	**1.00**	**1.96**
`no_energy`	0.17	0.92	2.49	0.24	0.92	1.95
`A0`	0.21	1.00	2.14	0.41	1.00	2.31
`no_write`	0.26	1.00	2.44	0.65	1.00	1.95
`high_inertia`	0.38	1.00	2.91	0.37	0.92	2.20
`no_restore`	**1.49**	0.83	**9.54**	**1.03**	**0.50**	**6.41**

their signature constructs, interaction $+0.50$, clean (Fig. 4, Appendix F), and the same crossing appears for negative interpretation against error magnification, interaction $+0.42$, clean. The secure-control pairs (hopelessness $\times$ positive reframe $+0.62$; error magnification $\times$ positive reframe $+0.66$) show large interactions but are not marked clean, for a mundane reason: secure has no elevated construct with which to complete the second arm, which is what a distortion-free persona should look like rather than a failure of the design. Clamping the belief field is selective at 0.73 (clamped field moves 3.6, others leak 1.2), supporting H2d for belief; emotion and goal are less selective (0.24, 0.04).

8.4 E3: Mechanistic localization (exploratory)

We probe where the persona effect is carried, on Gemma-2-2B for access to the open autoencoders; the experiment is exploratory, and we say where the readout limits the conclusions. Patching the source-persona residual into a secure-persona run at each layer's final position gives a structured curve rather than a flat one, with the causal effect concentrated at a persona-specific peak (Fig. 3, Appendix E): depression early (layer 6), anxiety in the upper-middle band (layer 21), perfectionism at the final layer (layer 25), none coinciding with the CAL attachment layer (12). This is consistent with CAL injecting the persona downstream of where the model naturally encodes one, since CAL writes the state rather than reading it there. One caution: a final-position logit readout is mechanically more sensitive to later layers, so the late peaks may partly reflect readout depth-bias, and the claim we are confident in is the weaker one, that the effect is localizable and its location persona-dependent.

At the depression peak (layer 6) the differential Gemma Scope features between depression and secure are sparse and cleanly separated: a handful carry nearly all of the difference, led by #9743 ($\Delta = -5.9$), #6201 (-4.3) and #13073 (-2.1), so the persona shift lands on a small set of interpretable directions in the SAE basis instead of smearing across the space, which is the feature-level half of H3. The head-level story is weaker: the scan returns persona-selective heads in the first two layers, but the necessity drop from ablating the top circuit (-0.24) is not clearly larger than an equal-sized random ablation (-0.15), so we claim no faithful head-level circuit and leave one to future work.

8.5 E4: Single-field transplant (proof-of-concept)

E4 asks whether one field can be grafted from a donor persona into a host. We install the secure host, clamp the depression donor's belief-field values (worthlessness, hopelessness) while emotion and goal run free, and score the host's generations. In the latent the graft is unambiguous, $\|\mathbf{z}_{\text{graft}} - \mathbf{z}_{\text{host}}\|_1 \approx 9.5$, and it perceptibly softens the host's language, tilting a calm and high-agency response toward hesitancy and self-doubt. Behaviourally that is as far as it goes: the secure system prompt keeps anchoring the surface, host and graft both score near zero on the judge's negative clinical constructs, and the field-selectivity metric is not meaningfully defined; raising the write gain to force the graft through degrades generation before it cleanly flips construct polarity, so we report no forced-gain number. Table 7 (Appendix H) makes the gap concrete: the belief-grafted host keeps a supportive surface but visibly loses its agency ("I feel unsure…I don't know how to be") without crossing into the hopelessness or self-blame that the judge is calibrated to detect. The donor's belief field is doing measurable work on the wording; the prompt-level persona is doing more. We read this as an informative boundary: latent separability (E2, selectivity 0.73) does not automatically produce behavioural dominance when a strong prompt-level persona occupies the same output channel. Selective transplantation apparently needs a weaker host anchor or a learned per-field write schedule, deferred to follow-up work.

8.6 E5: Human evaluation

E5 is the external check on the judge-based read-outs, 7,200 ratings aggregated per dialogue and compared pairwise by Wilcoxon signed-rank (n=80 paired dialogues). CAL leads on every scale that measures the property it was built for (Table 4), and the largest margin is where the dynamics predict it, post-shock recovery: 4.07 against 3.42 for static steering and 2.77 for the prompt (both $p < 10^{-4}$, Cohen's d=3.2 and 6.2 at the dialogue level), with the ordering CAL $>$ ActAdd $>$ Prompt holding within every persona (Appendix I). This supports H5 and supplies the perceptual content of H1c: `steer_static` looked perfect in the latent only because a pinned state cannot drift, and human readers judge its post-shock behaviour clearly worse. Consistency and overall follow the same

Table 4: E5 human evaluation (mean ± std, 1–5; best in bold). All pairwise differences $p < 10^{-4}$ (Wilcoxon over dialogue means); α: Krippendorff's ordinal agreement.

metric	Prompt	ActAdd	CAL	α
consistency	3.43±0.82	3.89±0.77	**4.23±0.76**	0.21
recovery	2.77±0.83	3.42±0.83	**4.07±0.79**	0.35
naturalness	**4.34±0.74**	4.13±0.77	4.23±0.77	0.06
overall	3.42±0.65	3.78±0.64	**4.18±0.62**	0.25

pattern (d=1.6–3.9). Naturalness goes the other way, the intervention-free prompt reading most fluent (4.34), CAL second (4.23, $p < 10^{-5}$) and static steering last (4.13): a residual-stream write costs some fluency, CAL less than the memoryless one. Agreement (Krippendorff's ordinal α) is fair for recovery (0.35), overall (0.25) and consistency (0.21) but near zero for naturalness (0.06), so the naturalness ordering rests on a scale annotators barely agree on.

9 Discussion

The experiments trace one claim from mathematics to generated text and then to its limit. The contraction is earned rather than assumed (E0); after a shock the state returns, removing the restoring force alone slows the return roughly nine-fold, and the contrast reproduces on a second family (E1); and E5 confirms with human raters what the latent geometry could only assert, that the pinned perfection of `steer_static` does not read as recovery while CAL's input-responsive return does, at a modest and honestly reported cost in naturalness.

The latent and behavioural channels do not always agree, and the disagreement is informative. In E1 the drift signal saturates under strong instruction-following, so the latent read-out carries the stability claim; in E2 the judge is unreliable on positive constructs, so the negative dissociations and the latent geometry carry it; in E4 the gap becomes the result itself. In short, CAL controls the latent dynamics it was designed to control, and its behavioural expression is mediated, sometimes bottlenecked, by the host's prompt-level persona on the same output channel.

From controller to training signal. CAL is used here at inference time, but a controller may be more useful as a source of supervision. Distilling its text would only teach a student to imitate outputs, the channel we have just shown to be bottlenecked; the more promising target is the *dynamics*: train the host, on perturbed dialogues, to reproduce the latent trajectory CAL produces, for instance with an auxiliary loss on $\mathbf{P}(\mathbf{h}_t - \boldsymbol{\mu})$ read from the host's own residual stream against the CAL reference $\mathbf{z}_t$, so that the post-shock return survives removal of the hook. Whether a frozen-architecture host can internalize a contraction it currently borrows is settled by neither our theory nor our experiments. A cheaper version is available at once: since CAL needs no training, its trajectories can be turned into preference pairs, recovered continuation over stranded continuation, without extra annotation.

Beyond persona. Nothing in the construction is specific to personality. It requires only that the attribute admit a low-dimensional state with named coordinates, a target point, and a labelled corpus to fit the read/write geometry, which formality, stance, epistemic hedging and register all do, and the same shock-recovery protocol would apply unchanged. Nearer-term follow-up is hardening: two-family replication for E2 and E4, the Llama-3.1-8B run, a behavioural readout for E3, a human study on the Mistral outputs, and a learned per-field write schedule.

10 Conclusion

We reframed persona control as the engineering of an attractor, an interpretable latent with a restoring force toward a persona centre, and tested it against a suite built to break it. The contraction is earned by the restoring force (E0); that force is the load-bearing component of post-shock recovery on both model families (E1) and its advantage is visible to thirty human raters (E5); personas dissociate cleanly on their signature constructs (E2); the effect localizes to a persona-specific layer band with a sparse SAE signature (E3); and a single-field transplant is real in the latent yet unable to dominate a strong prompt-level persona in the text (E4), a boundary we consider a finding in its own right.

Limitations

The stability result (E1) replicates across two families; E2 and E4 are single-model, with the Llama-3.1-8B replication running. E5 rates Qwen3-8B outputs only, and its naturalness scale has near-zero agreement (α=0.06), so that ordering is directional at best. E1's behavioural drift channel saturates under strong instruction-following, so its stability claim rests on the latent read-out. E3's final-position logit readout has a depth bias not fully

separated from causal depth, and head-circuit faithfulness is not established. The high-arousal anxiety persona generated degenerate text under the configured write gain and was excluded from E2, pointing to a write-gain×geometry interaction for a per-persona schedule to address. The E2 judge scores positive constructs unreliably on calm text, so the dissociation evidence rests on the negative constructs and the latent separation. The persona centres of Table 2 are specified by hand from the codebook rather than learned, so the geometry inherits whatever the codebook gets wrong. Comparison to prior steering runs through `steer_static` and the ActAdd-style human-study condition, which reproduce the ActAdd recipe inside our harness at matched write magnitude; benchmarking the published implementations under their own settings is left to future work.

Ethics Statement

The personas include clinical cognitive styles (depression, anxiety) fitted from public corpora of self-reported negative thought and stress (Sharma et al., 2023; Turcan and McKeown, 2019). CAL is a research instrument for studying persona stability, not a clinical tool, and a controller that makes a depressive style more persistent should not be deployed in any user-facing system without oversight; the same mechanism that holds a character also holds a harmful frame. No new personal data was collected. The 30 annotators of E5 were recruited under institutional review, informed that the dialogues contain simulated depressive and anxious language, free to withdraw at any point, and paid above the local minimum hourly wage.

Data and Code Availability

All corpora used to fit the read/write geometry are publicly available under their original licences: the Cognitive Reframing corpus (Sharma et al., 2023), the corpus of Shreevastava and Foltz (2021), Dreaddit (Turcan and McKeown, 2019), the Big-Five essays (Pennebaker and King, 1999) and PANDORA (Gjurković et al., 2021); we redistribute none of them and release the preprocessing scripts instead. The CAL implementation, persona configurations, fitted geometries, judge prompts and the full evaluation harness are released at the repository given in the footnote of Section 1. The system outputs we generated are released in full: the E1 shock-protocol dialogues and latent trajectories for both model families, the E2–E4 generations with their judge scores, and the 7,200 anonymized human ratings of E5 with their per-dialogue aggregates. Annotator identifiers and demographic information are not released. Appendix B is the supplementary material statement: it gives the hyperparameters, corpus sizes, judge configuration, per-experiment run counts and hardware needed to reproduce every number in the paper.

Acknowledgements

We thank the anonymous reviewers, whose comments shaped this version: the worked example of Section 3 and the persona register of Table 2 follow their suggestions, and the scale-sensitive statement of the decoding bias in Eq. (5) corrects an imprecision they identified. This work was supported by TODO.

References

Aaron T. Beck. 1963. Thinking and depression: I. idiosyncratic content and cognitive distortions. *Archives of General Psychiatry*, 9(4):324–333.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, et al. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*.

David D. Burns. 1980. *Feeling Good: The New Mood Therapy*. William Morrow, New York.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*.

Gemma Team. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Matej Gjurković, Mladen Karan, Iva Vukojević, Mihaela Bošnjak, and Jan Šnajder. 2021. PANDORA talks: Personality and demographics on Reddit. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*.

Aaron Grattafiori et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

John J. Hopfield. 1982. Neural networks and physical systems with emergent collective computational

abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558.

Albert Q. Jiang et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

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

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Rohin Shah, and Neel Nanda. 2024. Gemma scope: Open sparse autoencoders everywhere all at once on Gemma 2. *arXiv preprint arXiv:2408.05147*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*.

James W. Pennebaker and Laura A. King. 1999. Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, 77(6):1296–1312.

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

Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, 623:493–498.

Ashish Sharma, Kevin Rushton, Inna Wanyin Lin, David Wadden, Khendra G. Lucas, Adam S. Miner, Theresa Nguyen, and Tim Althoff. 2023. Cognitive reframing of negative thoughts through human-language model interaction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 9977–10000.

Sagarika Shreevastava and Peter Foltz. 2021. Detecting cognitive distortions from patient-therapist interactions. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology (CLPsych)*.

David Sussillo and Omri Barak. 2013. Opening the black box: Low-dimensional dynamics in high-dimensional recurrent neural networks. *Neural Computation*, 25(3):626–649.

Elsbeth Turcan and Kathleen McKeown. 2019. Dreaddit: A Reddit dataset for stress analysis in social media. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI)*.

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

Bichen Wang, Pengfei Deng, Yanyan Zhao, and Bing Qin. 2023a. C2D2 dataset: A resource for the cognitive distortion analysis and its impact on mental health. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10149–10160.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023b. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations*.

An Yang et al. 2025. Qwen3 technical report. *arXiv preprint*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, et al. 2023. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*.

Appendix Overview

Appendix A is the hypothesis register (Table 5); Appendix B the full implementation details; Appendix C the E0 inertia decomposition (Table 6); Appendix D the forward-step algorithm (Alg. 1); Appendix E the E3 localization curve (Fig. 3); Appendix F the E2 dissociation figure (Fig. 4); Appendix G the E1 ablation ladder (Fig. 5); Appendix H the E4 qualitative example (Table 7); Appendix I the E5 breakdown by persona (Table 8).

A Hypothesis Register

Table 5 collects every hypothesis together with the observation that was committed, in advance, as its refutation condition.

Table 5: **(Appendix A)** Hypotheses and the observation that would refute each.

ID	Claim	Refuted if
H0	Latent map is a contractive attractor	any $\rho(\mathbf{M}) \geq 1$, or non-monotone Lyapunov
H0′	Restoring force, not construction, drives convergence	λ=0 still converges with $\rho(\mathbf{A}) \geq 1$
H1a	CAL resists adversarial drift	drift $\geq$ prompt-only
H1b	CAL recovers after a shock	ARR low or τ undefined
H1c	Dynamics $\neq$ steering	`steer_static` matches full CAL recovery
H1d	Restoring force is necessary	ARR unchanged when λ=0
H2a	Personas separable in latent space	accuracy near chance
H2b	Construct pattern matches theory	correlation ≤ 0
H2c	Personas doubly dissociable	interaction ≤ 0 or not clean
H2d	Single-field clamp is selective	large cross-field leakage
H3	Effect is localized and faithful	diffuse patching or necessity $\approx$ specificity
H4	Field transplant is selective	target change $\approx$ leak change
H5	Humans prefer CAL's recovery	CAL rated $\leq$ either baseline on recovery

B Implementation Details

Dynamics and geometry. The configured dynamics are $\rho(\mathbf{A})$=0.85, λ=0.25, unit write gain α=1, and residual standardization on. The read/write geometry is fitted once per host from 2,277 cognitive-reframing items and 873 Dreaddit posts, capped at 200 exemplars per persona; the resulting named-axis reconstruction error stays below 4×10^{-4}. That figure is an internal consistency check on the fit, not a separability claim.

Judge. The independent judge is the same base model prompted separately, with chain-of-thought disabled, scoring eight behavioural constructs on $[0, 1]$. It sees only generated text and never the latent state or the distance to the centre.

Runs. E1 comprises 108 dialogues per family (9 conditions $\times$ 4 personas $\times$ 3 seeds, 100 turns each, shock at turn 50). E3 uses `gemma-scope-2b-pt-res` at width 16k. Every cell is checkpointed, so runs resume without repeating completed conditions. All experiments run on one NVIDIA H100 NVL in `bfloat16`.

Human study. E5 recruits 30 annotators, each rating a subset of the 80 shock-protocol dialogues per condition (20 dialogues per persona, the same dialogues across conditions), for 7,200 ratings in total. Raters see the three conditions for a given dialogue in randomized order and are not told which system produced which.

C E0 Inertia Decomposition

Table 6: **(Appendix C)** E0 inertia decomposition. The restoring force (λ=0.25) contracts the same transition matrix that diverges without it.

$\rho(\mathbf{A})$	λ	$\rho(\mathbf{M})$	converged
1.00	0.00	1.000	0.00
1.10	0.00	1.100	0.00
1.00	0.25	0.750	1.00
1.10	0.25	0.825	1.00

D Forward-Step Algorithm

Algorithm 1 CAL forward step at the hooked layer (Appendix D)

Require: hidden $\mathbf{h}_t$; state $\mathbf{z}_{t-1}$; $\mathbf{P}, \mathbf{W}, \mathbf{A}, \mathbf{c}, \boldsymbol{\mu}$; $\beta, \lambda, \alpha, \gamma$

$\mathbf{s}_t \leftarrow \mathbf{A}\,\mathbf{z}_{t-1} + \beta\,\mathbf{P}(\mathbf{h}_t - \boldsymbol{\mu})$ ▷ read

$\mathbf{z}_t \leftarrow (1-\lambda)\,\mathbf{s}_t + \lambda\,\mathbf{c}$ ▷ restore

$\mathbf{h}'_t \leftarrow \mathbf{h}_t + \alpha\,\mathbf{W}\,\mathbf{z}_t$ ▷ write

$\mathbf{p}_t \leftarrow \mathbf{W}(\mathbf{c} - \mathbf{z}_t)$ ▷ decoding pull

bias next-token logits by $\gamma\langle \mathbf{u}_v, \mathbf{p}_t \rangle / \|\mathbf{u}_v\|$

return $\mathbf{h}'_t, \mathbf{z}_t$

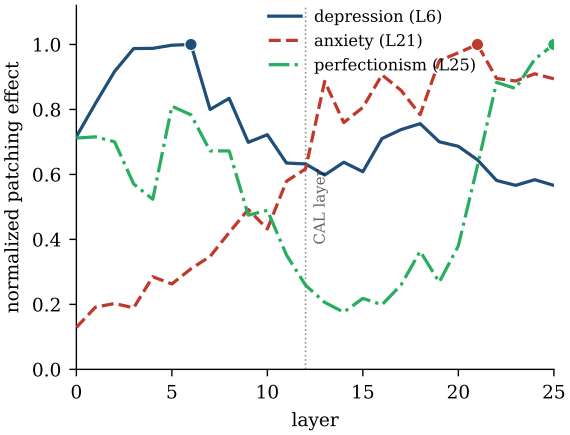


Figure 3: **(Appendix E)** Layer-wise activation-patching effect (Gemma-2-2B, source persona into a secure run, $|\Delta|$ normalized per curve). The persona effect is localizable with a persona-specific peak: depression early (L6), anxiety mid-late (L21), perfectionism late (L25), none coinciding with the CAL attachment layer (dotted).

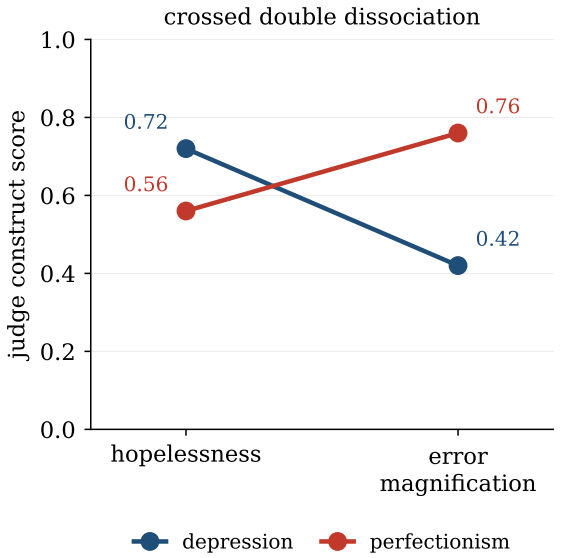


Figure 4: **(Appendix F)** Behavioural double dissociation (Qwen3-8B). Depression and perfectionism cross on their signature constructs: depression scores higher on hopelessness, perfectionism on error magnification (interaction +0.50).

E E3 Localization Curve

F E2 Double Dissociation

G Ablation Ladder

Figure 5 plots the recovery time constant for every ablation condition; the ordering discussed in Section 8 is visible directly.

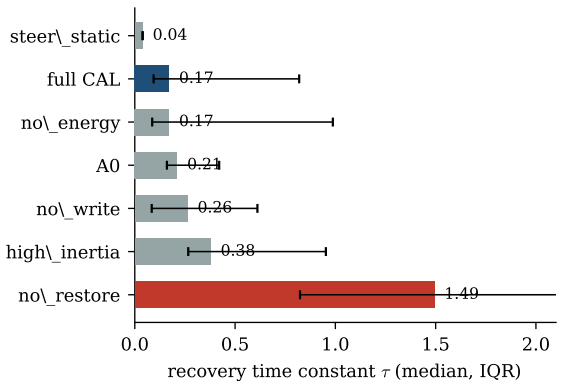


Figure 5: **(Appendix G)** Recovery time constant by condition (Qwen3-8B, median over 4 personas × 3 seeds, IQR bars). Every ablation slows recovery; removing the restoring force is by far the most damaging. `prompt_only` and `neutral` omitted, as they never leave their pinned read-out.

Table 7: **(Appendix H)** E4 qualitative example (Qwen3-8B). Grafting the depression belief field onto the secure host erodes agency and injects self-doubt, but does not flip the response into a scored distortion.

Stimulus: "My message went unanswered all day."	
host	"I'm feeling a bit concerned, but I'm trying to stay calm. Maybe they're busy or preoccupied. I'm trying to regulate my emotions."
graft	"I'm trying to understand if there's a reason…but I'm not sure. I feel unsure, and I don't know how to be, but I want to stay open."

H E4 Qualitative Example

I Human Evaluation by Persona

Table 8 breaks the recovery scale of Table 4 down by persona. The method ordering is uniform.

Table 8: **(Appendix I)** E5 recovery-after-shock by persona (mean over 30 annotators × 20 dialogues). The ordering CAL > ActAdd > Prompt holds within every persona.

persona	Prompt	ActAdd	CAL
Anxiety	2.76	3.47	**4.11**
Depression	2.72	3.35	**4.03**
Perfectionism	2.81	3.43	**4.10**
Secure	2.78	3.43	**4.05**