DPEB: Evaluating Persona Evolution in Multi-Session Dyadic Interaction

Anonymous authors
Paper under double-blind review

Abstract

Long-horizon conversational agents are increasingly expected to sustain a coherent persona while adapting to the relationship they build with a single returning user. The decisive question is temporal and dyadic: does the persona change when, and only when, the user's situation warrants it? Research to date has treated persona as a static target to be steered or a property to be monitored, leaving this question unmeasured. We introduce DPEB, a controlled benchmark for dyadic persona evolution: 1,200 fully reproducible scenarios, balanced by event type at the first evolution point, each pairing an agent persona with one user's relationship arc and labeling objectively the sessions in which a Big Five dimension should move and those in which it should not. Personas are carried by activation steering, measured by projecting the residual stream onto contrastive persona vectors, and scored on three axes (evolution responsiveness, stability, direction accuracy) against an oracle that updates at exactly the labeled sessions. Across the models evaluated DPEB yields a reframing result: static regimes score zero by construction and history-conditioned prompt-update baselines barely move, whereas a re-implemented Persona-Flow-style baseline moves readily but drifts, and holding the update rule fixed while varying only the evidence detector changes the outcome far more; measured on the evaluated scenarios, the two LLM-based gates reach 97-99

1 Introduction

Large language models (LLMs) now back conversational agents (companions, tutors and assistants) that persist across many sessions with the *same* user (Park et al., 2023a; Vezhnevets et al., 2024). Such an agent must *carry* a persona (a disposition that makes it recognisable across situations and over time) and must decide *when* that persona should change: a user who first exchanges pleasantries may later disclose a hardship, take on a supervisory role, or settle into a long-term routine. A system that never changes feels indifferent; one that changes at every emotional turn feels incoherent. The desired behavior is *controlled evolution*: change when the relationship genuinely changes, and stay put otherwise. Stated as the question this paper measures: does the persona change when, and only when, the user's situation warrants it?

What this paper validates. Our validated contribution is an instrument for that timing question: DPEB (*Dyadic Persona Evolution Benchmark*), 1,200 balanced, fully reproducible scenarios pairing an agent persona with one user's relationship arc and labeling *objectively* the sessions in which a Big Five dimension should move and those in which it should not. The user side is frozen, the labels come from a fixed event taxonomy applied by lookup rather than from a model's opinion, and the persona state is read by projecting the residual stream onto contrastive persona vectors. The headline result, quantified per model in the contributions below, is that holding the update rule fixed while varying *only* the evidence detector moves the outcome far more than varying the application schedule.

What this paper proposes, and what it does not claim. *How* a persona is represented is a separate question. We therefore also present HPD (*Hierarchical Personality Dynamics*)

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.

as a **design view** separating a slow, evidence-licensed trait trajectory from fast within-session adaptation: a continuous, variable, time-evolving state with a slow component (a trait attractor) and a fast one (situational deviation), consistent with the psychology of within-person variability (Fajkowska, 2018; Fleeson, 2001; Tett et al., 2021). It is design rationale, not a result: no number in this paper comes from it, and §8 states both the architecture and the validation it would require. Static-assignment treatments (in-context trait descriptions (La Cava, 2025; Pan et al., 2023), fine-tuning (Zollo et al., 2025), activation steering (Bhandari et al., 2026b; Deng et al., 2024; Wang et al., 2026c; Zhu et al., 2025)) elicit consistent trait scores (John et al., 1991; Myers et al., 1962) but fix the persona, so agents drift (Gao et al., 2025) (§2).

Contributions.

- **A controlled, fully reproducible benchmark for persona evolution (DPEB).** 1,200 balanced scenarios, each pairing an agent persona with a single user's relationship arc, with objective labels (an eight-type event taxonomy applied by lookup) of when a persona should and should not change, a frozen user side so every policy sees identical inputs, three axes (M_1, M_2, M_3) that trivial policies cannot satisfy jointly, and an oracle reference point (§3).
- **A validated trajectory measurement that localises the binding constraint.** The readout is monotone in the injected persona level, tracks two independent LLM judges (r=0.81, r=0.62; inter-judge r=0.92), and is anchored on the oracle (§4). With it we show that evidence-detection quality, not the update rule, is what controlled evolution turns on: swapping only the detector moves M_1 by -0.013 to $+0.058$ and M_3 by -0.04 to $+0.30$ across four open models, whereas swapping only the schedule moves M_1 by at most 0.020 and the threshold θ is inert; the two LLM-based gates reach 97-99% event recall yet false-trigger on 36-40% of ordinary (non-evolution) turns; at the same fixed rule a *trained* sentence-embedding detector reaches 93-98% of the oracle M_1 and beats both deployed gates on all three axes on two backbones (§6.3).
- **A hierarchical view for the representation side (HPD, a design), with an honest scope.** We formalize the reframing result as a psychology-grounded, multi-timescale persona state with dual persona-state-embedding (PSE) extraction, a hierarchical-distillation (LHRD) module, and a personality density field whose variability would gate an activation steering vector injected after the final RMSNorm of a frozen backbone (§8). It is a design view only: it has not been implemented or validated, and no number of this paper depends on it. The benchmark's limitations (a five-session window, a single run per configuration, synthetic scenarios with a Chinese-only user side, a readout validated against static levels rather than trajectories) are stated in one place in §7.1.

2 Related Work

Personality in language models. LLM personality has been elicited by prompting (La Cava, 2025; Pan et al., 2023), fine-tuning (Zollo et al., 2025), and inference-time activation steering (Bhandari et al., 2026b; Deng et al., 2024; Panickssery et al., 2023; Turner et al., 2023; Wang et al., 2026c; Zou et al., 2023); later work automates contrastive-pair construction (Khedar et al., 2026; Wang et al., 2026b), selects layers or positions (Soo et al., 2025; Ye et al., 2026a), denoises the directions (Feng et al., 2026b; Hiramatsu et al., 2026) and characterizes steering failures (Bhandari et al., 2026a; Ong et al., 2026; Subbiah et al., 2026), with gated variants for conditional control (Fereidouni et al., 2026; Tripathi et al., 2026), temporally extended preference change (Mráz & Shenk, 2026) and dynamic persona control (Li, 2026; Wang et al., 2026a; Ye et al., 2026b). HPD follows this line for its carrier; DPEB asks the complementary question: *when* should a steered persona move over a long interaction?

Dynamic role-playing and character-state evolution. PHASE-Tree (Long et al., 2026; Tang et al., 2026) organizes character state into identity/persona/session/moment strata;

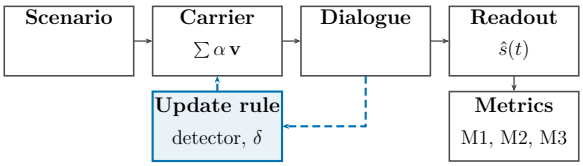

Figure 1: The DPEB pipeline. A scenario pairs an agent persona with one user's relationship arc; the persona is carried by activation steering and the dialogue is replayed against frozen user turns. The state is read by projecting the residual stream onto contrastive persona vectors and scored on $M_1/M_2/M_3$. Policies differ only in the update rule (green); the objective labels (grey) enter only through the metrics. Reproduced from the release.

DREAM (Chen et al., 2026d) builds event-aware memory graphs on an ABC (antecedent-belief-consequence) model; PGMem (Liang et al., 2026) couples persona and memory through typed evidence edges; and CRPO (Chen et al., 2026c) with Character-R1 / reward-decomposed RL (Gu et al., 2026) applies character-centric training. Memory-augmented agents (Packer et al., 2023; Park et al., 2023b; Zhong et al., 2023) store, retrieve and consolidate experience (Chen et al., 2026a; Dai et al., 2026; Ren et al., 2026; Yang et al., 2026b; Zhan et al., 2026), and autonomous persona evolution without an external user is studied by Li (2026) and long-horizon personalization by Li et al. (2026). The session belief stream of §8 would be a differentiable analogue of the ABC stream, with a *continuous, variable* density field where the above systems use discrete fields or symbolic graphs.

Psychology of personality change. Fleeson's density-distribution theory (Fleeson, 2001) establishes that personality is a *distribution* of states, not a fixed point; traits are hierarchically organized (Traits as Hierarchical Systems, THS) (Fajkowska, 2018); Trait Activation Theory (TAT) (Tett et al., 2021) makes expression depend on situational demand; and social-investment theory, supported by a meta-analysis of life events and personality change (Bleidorn et al., 2022), links life events to trait change. §8 would operationalize the first three inside a neural state; DPEB's taxonomy operationalizes the fourth as objective labels.

Benchmarks and measurement. A large body of work evaluates whether an LLM can *stay in character* (Han et al., 2026; Occhipinti et al., 2025; Shao et al., 2023; Shouqi et al., 2026; Wang et al., 2025b; Zhou et al., 2024) and how personality can be measured (Cho & Cheong, 2025; Frising & Balcells, 2025; Li et al., 2024; Ma et al., 2026; Yu et al., 2026), including projecting activations onto persona vectors (Chen et al., 2025). Long-term memory benchmarks probe recall across sessions (Wu et al., 2024; Yang et al., 2026a), persistent sycophancy has been benchmarked in stateful agents (Mao et al., 2026), and conceptual frameworks for narrative continuity and personal-agent adaptation remain on paper (Natangelo, 2025; Qian et al., 2026). All of these fix a target persona, study a scripted narrative or analyze growth post hoc; DPEB instead supplies controlled single-user interaction with objective labels of when the persona should change and validates the measurement as a *trajectory*. Because it relies on model judgements, we take judge construct validity seriously (Bagaria et al., 2026; Chen et al., 2026b; Kapetanovic et al., 2026; Wahi, 2026).

3 DPEB: A BENCHMARK FOR DYADIC PERSONA EVOLUTION

Design goals. A benchmark for persona dynamics must supply *objective* labels of when change is warranted, be *fully reproducible*, and *separate* the failure modes it claims to distinguish (inertness vs. drift vs. controlled change). Everything it needs is released (scenarios, frozen user side, measurement code, detector diagnostics); Figure 1 summarizes the pipeline (full overview in Appendix A.16).

Scenarios and event taxonomy. Each scenario pairs an agent (naive → trained) with a single user and a relationship arc (e.g. stranger → regular → disclosure → dependence), and

contains ≥ 1 *evolution points* (a user turn expressing a relationship- or role-level change) and ≥ 1 *stable segments* (ordinary interaction that should not move the persona). The key control is the *event taxonomy*: rather than letting a model opine on which dimension should move, we fix eight event types, each mapped to one Big Five dimension and direction by *lookup*, which makes the ground truth objective and auditable (Appendix A.16).

Data statistics and balancing. The generator produced 1,750 candidate scenarios, of which 1,200 survived automated quality control. Balancing used the *first* evolution point's event type (150 scenarios per type), removing the generator's bias towards hardship and stress events (per-type counts span 244-360 across both points). Every scenario declares exactly two evolution points, two to six stable sessions and five to twelve sessions in total (Appendix A.1).

Persona carrier: activation steering. The persona is a five-dimensional state $\mathbf{s} \in [0,1]^5$ carried by activation steering: for each dimension d we extract a contrastive vector $\mathbf{v}_d$ as the mean residual-stream difference between positive and negative system prompts at a mid-network layer, following the standard protocol of Chen et al. (2025); Weng et al. (2024), and add $\sum_d \alpha_d \mathbf{v}_d$ to that layer with $\alpha_d = (s_d - 0.5)K$. The magnitude must be calibrated to the model's residual norm: K=240 for Qwen3-8B (layer 20, residual norm ≈ 110), for which $|\alpha| \in [30, 120]$ is the effective band (Appendix A.2); a magnitude small relative to the residual norm is *silently inert*, a pitfall that also bears on the variability gate of §8.

Metrics. An evolution point occurs in session t, targets dimension d and has direction $r \in \{+1, -1\}$. The measurement yields one checkpoint per simulated session, so let $\hat{s}_d(t)$ be the projection of the measured state at session t and $\hat{s}_d(t^-)$ the checkpoint immediately before session t. The three axes are

$$M_1 = \mathbb{E}_{(t,d,r)}\Big[\max\big(0,\ r \cdot (\hat{s}_d(t) - \hat{s}_d(t^-))\big)\Big], \quad \text{(evolution responsiveness, higher is better)} \tag{1}$$

$$M_2 = \mathbb{E}_{(a,b)}\Big[\tfrac{1}{5}\textstyle\sum_{d'} |\hat{s}_{d'}(b) - \hat{s}_{d'}(a)|\Big], \quad \text{(stability drift, lower is better)} \tag{2}$$

$$M_3 = \Pr_{(t,d,r)}\big[r \cdot (\hat{s}_d(t) - \hat{s}_d(t^-)) > 0.01\big], \quad \text{(direction accuracy, higher is better)} \tag{3}$$

where M_1 and M_3 average over all labeled evolution points in the measured window, and M_2 over consecutive checkpoint pairs (a, b) with no labeled evolution session in $(a, b]$ and at least one of a, b declared stable. M_1 is *clipped at zero* and M_3 counts a success only when the correct-direction change exceeds 0.01 on the calibrated $[0, 1]$ scale, so an inert policy scores zero on M_1 and M_3 (and trivially optimal M_2), while a noisy one pays in M_2.

4 Validating the measurement

A benchmark that measures its own state risks circularity, so we validate the measurement first. To check that the readout tracks *behavior* rather than re-encoding the injected prompt, we score the agent's responses with two independent LLM judges (GLM-4-Flash and GLM-5.3-Flash) that never see the persona: both are monotone in the persona level for all five dimensions, the projection correlates with them at r=0.81 and r=0.62, and the judges agree at r=0.92. The readout is thus a valid proxy for observable behavior rather than an artifact of the carrier, though the validation covers *static* persona levels and trajectory-level human validation remains open (§7.1).

Oracle reference point. We add an oracle policy that performs the correct update at exactly the labeled sessions. On Qwen3-8B it attains M_1=0.247, M_2=0.000, M_3=0.930, the ideal "high evolution, zero drift, correct direction" profile; that the metrics place it at this corner, and that trivial policies cannot reach it, show the benchmark separates the intended regimes. Because evolution points whose target dimension saturates near the boundary cannot be moved, capping M_1 for *every* policy, we normalize the data so that each evolution point has headroom, roughly doubling the M_1 signal for all policies (Appendix A.7).

5 Instrument: an evidence-gated update

We claim no new mechanism: the policy is a simple instrument separating *when* a persona may move from *how far* it moves, so the benchmark can locate the bottleneck. We call it EGU (an *evidence-gated update*). It has three borrowed components, two of which the literature and the slow/fast hierarchy of §8 suggests should matter. At every user turn a *detector* proposes a candidate event with a confidence, or nothing; because the taxonomy fixes the event-type $\rightarrow$ (dimension, direction) mapping by lookup, the detector only has to *classify*, never to decide the direction. We compare three detectors: the agent's *own* few-shot gate (the evaluated model itself classifies, given the taxonomy block and few-shot examples of Appendix A.3), a *model-agnostic* API gate (the same prompt issued to a fixed third-party model) and a frozen prototype classifier (nearest class mean over hidden states); §6.3 adds a detector trained on sentence embeddings. Evidence then accumulates per event type, $E_\tau \leftarrow \min(c_{\max}, \rho E_\tau + c)$, with ρ=0.55 and cap $c_{\max}$=3; the accumulator is decayed by ρ at each session boundary, so that true events (type-consistent proposals) build up while scattered false positives need not; an optional recurrence filter requires the same type twice. The default *application rule* applies one bounded step per confirmed event at session boundaries, $s_d \leftarrow \text{clip}(s_d + r_d\delta)$ for every type with $E_\tau \geq \theta$, then resets E_τ; a per-turn ablation applies the same step immediately at every firing turn, with no accumulation. We use δ=0.4 and θ=0.8 by default (both ablated in §6). Three configurations appear in the results, differing only in these knobs: EGU-own (own gate, session-end application, δ=0.4), EGU-api (API gate, session-end, δ=0.4) and EGU-rec (API gate plus the recurrence filter); Appendix A.15 maps every name to the released policy identifier.

6 Experiments

Setup. Unless noted, we evaluate DPEB on 100 held-out scenarios (never used to fit the optional detector classifier), five sessions per scenario, on Qwen3-8B (layer 20, K=240) and Llama-3.1-8B-Instruct (layer 16, K=20); Qwen2.5-7B and Llama-3-8B were run with their own calibrated layer and steering scale (Appendix A.2). Baselines: *static* (fixed persona), *adaptive* (per-turn unconstrained update), *per-turn gate* (taxonomy gate at every turn, the matched baseline) and *per-turn gate,* δ=0.4 (step matched to the EGU configurations). Significance is reported by paired t-tests over scenarios, and all methods see identical frozen user turns. Comparisons against re-implemented prior-work regimes (a history-conditioned prompt update and a Persona-Flow-style predict-then-steer baseline) are in Appendix A.8; the full table and figure are in Appendix A.17.

6.1 DPEB main results: the detector carries the gain

Table 1 reports all four open models at n=100 with the static control and the oracle. On Qwen3-8B the unconstrained *adaptive* baseline reaches M_1=0.049 and M_3=0.325 (released run `eval_steer_n100.json`), while static regimes score M_1=0 by construction. EGU-own attains M_1=0.195 against 0.162 for the matched per-turn gate (t=+2.96, p=0.004), with a non-significant stability advantage (M_2 0.040 vs. 0.053; t=−1.56, p=0.12) and unchanged direction (M_3 0.730 vs. 0.725); enlarging the per-turn step to match (δ=0.4) does *not* help it. Applying the *same* detector *per turn* rather than once per session is *better* (M_1 0.215 vs. 0.195; t=+2.91, p=0.005; M_2 indistinguishable, t=+0.52, p=0.61), and removing accumulation (θ=0) changes nothing (0.195 vs. 0.197; t=−1.69, p=0.09): the benchmark locates the gain exactly where §6.2 measures it, in the detector.

Knob isolation: detector vs. schedule. Changing the *application schedule* at fixed detector and step moves M_1 by +0.020, +0.012, +0.012, +0.013 across the four models of Table 1 and never improves M_2; changing the *detector* at fixed schedule and step moves M_1 by −0.013, +0.017, +0.058, +0.056 and M_3 by −0.04, +0.06, +0.26, +0.30, with the large gains exactly where the agent's own gate is unreliable: on Llama-3.1-8B the model-agnostic API gate raises M_1 from 0.104 to 0.162, M_3 from 0.450 to 0.705 and lowers M_2 from 0.083 to 0.031 (all p<0.001; 86.6% of that run's oracle M_1): the detector, not the carrier or the schedule, is the cross-model bottleneck. Figure 2 shows a measured trajectory behind the

Table 1: Isolating the two knobs: four open models (Qwen3-8B, Qwen2.5-7B, Llama-3.1-8B, Llama-3-8B), n=100 each. *static* never updates, so M_1=0 holds by construction. The schedule swap keeps detector and step fixed (own gate, δ=0.4, session-end vs. per-turn); the detector swap keeps schedule and step fixed (session-end, δ=0.4, own gate vs. model-agnostic API gate). The oracle updates at exactly the labeled sessions.

Model	Method	$M_1 \uparrow$	$M_2 \downarrow$	$M_3 \uparrow$
Qwen3-8B	static (fixed persona)	0.000	0.000	0.000
	per-turn gate	0.162	0.053	0.725
	EGU-own (session-end)	0.195	0.040	0.730
	EGU-own (per turn)	0.215	0.041	0.720
	EGU-api (session-end)	0.182	0.046	0.690
	oracle	**0.247**	**0.000**	**0.930**
Qwen2.5-7B	static (fixed persona)	0.000	0.000	0.000
	per-turn gate	0.087	0.016	0.585
	EGU-own (session-end)	0.116	0.047	0.595
	EGU-own (per turn)	0.128	0.054	0.545
	EGU-api (session-end)	0.133	0.031	0.655
	oracle	0.152	0.000	0.910
Llama-3.1-8B	static (fixed persona)	0.000	0.000	0.000
	per-turn gate	0.074	0.057	0.540
	EGU-own (session-end)	0.104	0.083	0.450
	EGU-own (per turn)	0.116	0.090	0.450
	EGU-api (session-end)	0.162	0.031	0.705
	oracle	0.188	0.000	0.965
Llama-3-8B	static (fixed persona)	0.000	0.000	0.000
	per-turn gate	0.047	0.037	0.425
	EGU-own (session-end)	0.067	0.056	0.385
	EGU-own (per turn)	0.080	0.057	0.375
	EGU-api (session-end)	0.123	0.026	0.680
	oracle	0.160	0.000	0.980

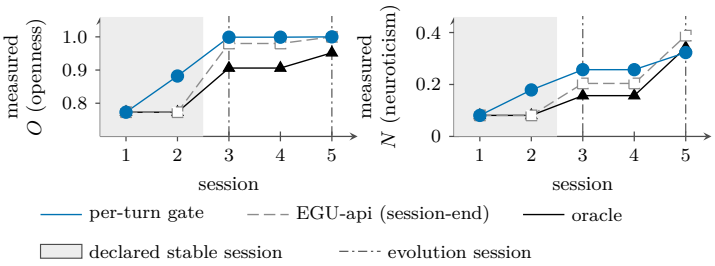

Figure 2: Measured trajectory of scenario `dpeb_0807` on Llama-3.1-8B for the two dimensions that carry its evolution points (session 3: `encourage_novelty` on O; session 5: `adds_stress` on N). Shaded bands are declared stable sessions and the dash-dotted line marks the evolution session; curves are the *measured* state of the per-turn gate, the EGU-api session-end policy and the oracle. The per-turn rule moves on stable session 2, while EGU-api stays flat and takes one step at each evolution point. Reproduced from the release.

M_2 gaps; knob ablations, the n=60 multi-model slice and the left/right panel figure are in Appendix A.16.

6.2 Detector diagnostics

Because detector quality is what the rest of this section turns on, we measured it against the taxonomy labels on the evaluated scenarios' own turns with the committed script `paper/detector_diagnostics.py`, which runs the released gates unchanged, over 179 positive turns (the first user turn of each labeled evolution point at session ≤ 5) and 771 negative turns, 394 declared stable. The own-model few-shot gate reaches 0.860 event-type accuracy at 0.994 event recall and a 0.403 false-trigger rate; the model-agnostic API gate reaches 0.827/0.972/0.361. Both deployed gates therefore reach 97-99% event recall (0.994 and 0.972) while false-triggering on 36-40% of ordinary (non-evolution) turns (0.403 and 0.361; restricted to declared-*stable* turns 0.343 and 0.269, Appendix A.9): high recall does not by itself buy a usable gate. The frozen prototype classifier, a diagnostic reference rather than one of those gates, stands at 0.776/0.961/**0.017** (precision 0.930), showing that a high-recall, near-precise

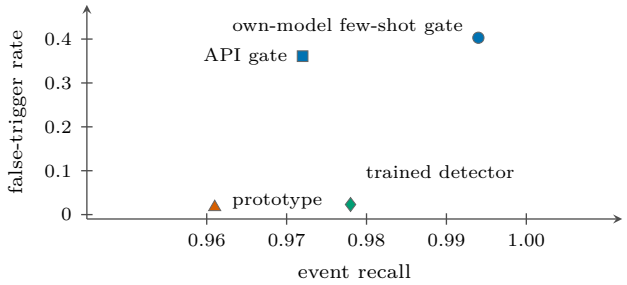

Figure 3: Detector operating points on the turns the closed loop actually sees (179 positive, 771 negative, 394 declared stable). Both deployed gates trade precision for near-total event recall, while the prototype reaches a near-precise high-recall corner: the bottleneck that Table 2 turns on.

Table 2: Detector swap on full arcs, n=100 held-out scenarios. The update rule (session-end, δ=0.4), carrier and slice are fixed, and the per-turn-gate row reproduces the reference of Table 5; the detector comparison is causal within a sweep, and the rows of a column come from the sweep named in Appendix A.10 (on Llama-3.1-8B the EGU-own row is from the detector-swap sweep and the EGU-api, prototype and trained rows are from the trained-detector sweep); the oracle is the same rule with a perfect detector. "Trained" is a logistic regression over BGE-M3 sentence embeddings (§A.10); the prototype is fitted on Qwen3-8B hidden states and is not backbone-decoupled. A separate sweep of the Table 5 protocol is reported in Appendix A.10.

Qwen3-8B

Detector	$M_1 \uparrow$	$M_2 \downarrow$	$M_3 \uparrow$
per-turn gate	0.150	0.039	0.685
EGU-api	0.168	0.035	0.645
prototype	0.212	0.005	0.725
trained	**0.221**	0.010	**0.800**
oracle	0.237	**0.000**	0.920

Llama-3.1-8B

Detector	$M_1 \uparrow$	$M_2 \downarrow$	$M_3 \uparrow$
per-turn gate	0.071	0.047	0.515
EGU-own	0.098	0.057	0.425
EGU-api	0.153	0.023	0.680
prototype	0.000	0.000	0.000
trained	**0.178**	0.008	**0.800**
oracle	0.181	**0.000**	0.950

operating point does exist on exactly the turns the closed loop sees (Appendix A.9); Figure 3 plots these operating points.

6.3 Detector swap: causal evidence for the bottleneck

§6.2 and the knob isolation above both point at the detector; Table 2 tests that causally: with the update rule (session-end, δ=0.4), the carrier (activation steering) and the slice (n=100 held-out scenarios over their *entire* declared arcs) fixed, only the evidence detector varies across the rows below the first (the matched per-turn gate of the full-arc table, Appendix A.6); we add here a detector *trained* on sentence embeddings (logistic regression over BGE-M3 features, fitted on two-thirds of scenarios by a scenario-level split; Appendix A.10).

On both backbones the trained detector significantly beats *both* deployed LLM gates on *all three* axes (paired t over scenarios, p<0.01 on all six comparisons per backbone; Appendix A.10) and reaches 93-98% of the oracle's M_1 (0.221 of 0.237 on Qwen3-8B, 0.178 of 0.181 on Llama-3.1-8B). Since only the detector changes across those rows, the comparisons are causal; drift tracks the detector's false-trigger rate: on Qwen3-8B, where the prototype is fitted natively, 1.7% gives M_2=0.005, 2.3% gives 0.010 and the deployed gates' 36-40% gives 0.039 and 0.035, while M_1 and M_3 move with the detector too.

The prototype, the near-precise operating point of §6.2, *collapses* on Llama-3.1-8B (M_1=M_2=M_3=0.000): its nearest-class-mean directions are fitted on Qwen3-8B hidden states. The open problem is therefore not a precise detector but a *backbone-decoupled* one, on which the trained detector leaves almost no M_1 headroom (0.178 against the oracle's 0.181; Appendix A.10).

7 Discussion

What the benchmark says. A multi-timescale persona state is only meaningful if something decides *when* each timescale may move; we measure that half with a deliberately simple instrument rather than the proposed hierarchy. The two knobs a hierarchy naturally exposes (session-level consolidation vs. per-turn application, and the threshold licensing a slow update) are worth at most 0.020 on M_1 and never improve M_2, and the threshold can be switched off entirely, whereas the detector changes both axes across all four models and, once trained, approaches the oracle (§6.3). Two constraints bear on §8: *calibration* (the steering magnitude must be scaled to the model's residual norm or the carrier is silently inert, Appendix A.2) and *layer-localised, probe-guided reading* (the DPEB measurement projects the residual stream onto contrastive directions at a chosen layer). A compact persona state is best treated as a *carrier whose timing is the problem*, not as a substitute for full dialogue history.

7.1 Limitations

Scope of the HPD material. HPD is stated as a design view, not a validated method: its architecture has not been implemented or evaluated here, and no number from it appears anywhere in this paper. The benchmark and its findings stand on their own.

The oracle is not an upper bound on M_1. The oracle updates at exactly the labelled sessions, which fixes stability and direction by construction, but not evolution responsiveness: a policy can exceed the oracle's M_1 by moving on unlabelled turns as well (on the case-study scenario the oracle's M_1 is 0.159 against 0.196 for EGU-api, Appendix A.14).

Measurement window. Every DPEB number in the main tables is a five-session result, the window in which all labeled events of the primary analysis lie. The released arcs declare five to twelve sessions (78.6% declare six), so 10.5% of labeled evolution points in this n=100 slice (21/200; 12.5% at n=60) sit beyond it and 73/100 evaluated scenarios declare a stable session 6 that is never reached; Appendix A.6 re-runs the key comparison over *entire* arcs, where those points are measured, recovering the same pattern.

Statistical protocol. We report a single run per configuration, with no seeds and no multiple-comparison correction across six policies × three metrics × four models; M_2 rests on 1.34 measurable transitions per scenario, so M_2 gaps below 0.01 should not be read as reliable. Run-to-run variation is small but non-zero (e.g. `gated` M_1 0.161 vs. 0.162 across runs); multi-seed replication with corrected significance is future work, and the frozen user side makes it cheap.

Scenario and language scope. The scenarios are synthetic, generated under a fixed taxonomy, so ecological validity is untested; the taxonomy simplifies the psychology; only the first evolution point of each scenario is balanced by event type; and the user side is Chinese only. On real user texts (LongLaMP product-review and topic-writing profiles, n=60) 15% (9/60) of everyday texts triggered a relationship-event label, mostly `disclose_hardship`: the deployed gates over-trigger on ordinary personal content, a deployment risk we report as a limitation rather than a validation (Appendix A.11).

Validation and operating point. The readout is validated against two judges from the same model family and against *static* persona levels rather than trajectories; a human validity study is the most valuable remaining measurement, and deploying a gate still needs a tighter operating point than the LLM gates provide (§6.3).

8 A Hierarchical View (HPD, a Design Proposal)

Status. This section is a *design view*, not a method: it reports no results and no number. It formalises the DPEB finding of §6, that evidence detection rather than the update rule is

the binding constraint, as a hierarchical split: a slow, evidence-licensed trait trajectory plus a fast, within-session adaptation. We state it because the benchmark's design follows from it, and because it makes explicit what an implementation would have to decide.

The view, and what it implies. A state that can act on "*when* a persona should move" needs two timescales and a rule for what licenses the slow one. The fast component adapts within a session and carries no lasting commitment; the slow component moves only when the incoming evidence licenses it, which is exactly the quantity DPEB measures on a simple instrument. This reading is what makes the benchmark's three axes the right ones: M_1 asks whether the slow component moves at licensed points, M_2 whether it stays put otherwise, and M_3 whether it moves in the right direction. HPD (*Hierarchical Personality Dynamics*) is one concrete way to instantiate the view, with a slow trait attractor and a fast situational deviation, and a gate that decides how much of the state is expressed; its stages and design decisions are recorded in Appendix B. The *view* is the contribution here, not the architecture: it is what the DPEB results constrain.

How the findings constrain the design. Read as a schedule, the hierarchy is the object our ablation studies: the slow component is a session-level update that only evidence should trigger, and the fast component is within-session adaptation. Two of our results bear directly on it. First, the detector, not the schedule, carries the gain (§6), so a gate that reads only the internal state would have to be extended to read the *detected* evidence, since variability alone is not evidence. Second, a gate that fires on ordinary stable turns buys apparent responsiveness at the cost of drift, so the operating point matters as much as the mechanism. Evaluating HPD would mean implementing its stages on a frozen backbone and reporting every configuration with multiple seeds; we make no such claim here (§7.1, Appendix B).

9 Conclusion

We presented DPEB, a benchmark for dyadic persona evolution in multi-session, single-user interaction: 1,200 balanced, fully reproducible scenarios with objective taxonomy-derived labels, a frozen user side, a validated activation-vector measurement, three axes and an oracle updating at exactly the labeled sessions. On it the binding constraint is the quality of the evidence that licenses a move: across four open models, swapping only the detector moves M_1 by -0.013 to $+0.058$, while swapping only the schedule is worth at most 0.020 on M_1 and the accumulation threshold is inert; at the same fixed rule a detector trained on sentence embeddings reaches 93-98% of the oracle M_1. We formalize those findings as a hierarchical view separating a slow, evidence-licensed trait trajectory from fast within-session adaptation, and state HPD as a design whose implementation remains future work. We release the scenarios, the frozen user side, the measurement code and the detector diagnostics.

Reproducibility Statement

We release the scenarios (`dpeb_final_1200.json`), the frozen user side (every user turn is pre-generated and stored, so every policy sees identical inputs), the measurement code, the detector diagnostics (`detector_diagnostics.py`, `detector_diagnostics_results.json`) that produce the detector numbers reported in this paper, and the prompts, persona vectors and evaluation code. An anonymised supplementary archive with these artifacts, and a mapping from every table and figure to the file that produces it, accompanies this submission. We report the model, layer, steering scale K, evidence threshold θ, decay ρ, step δ, detector and gate prompt for every configuration of the four main models; the Qwen2.5-3B scale check reuses the layer-16 setting of Qwen2.5-7B without a separate K sweep; the evaluation window and statistical protocol are stated in §7.1, and Appendix A.15 maps every name to the released policy identifier. HPD is presented as a design proposal; its implementation and evaluation are future work (§8, §7.1, Appendix B).

Ethics Statement

See §7. Persona steering can be used to manipulate users; our benchmark studies a controlled, benign setting and measures stability as well as change. Scenarios are synthetic and contain no real personal data.

Broader Impact

Long-lived assistants that change their persona are useful (rapport, personalization) but risky (undetected drift, inconsistency); our contribution is mainly measurement, a benchmark that scores *when* a persona should change and *whether it stays put otherwise*, plus a design proposal that makes the state being changed explicit and inspectable.

AI Use Disclosure

Large language models were used extensively and we disclose this in full: for retrieving and discovering related work; for research ideation; for writing and debugging the scenario-generation, simulation, measurement and statistics code, and running the experiments; for drafting and polishing the manuscript; and for generating the synthetic scenario corpus under a fixed taxonomy. All reported numbers were computed from the frozen artifacts in the released repository and every citation was checked against its source. The human author directed the research question, benchmark design and protocol, and reviewed and approved every claim, table and figure.

References

Anshul Bagaria, Sowmya S Sundaram, Gokul S Krishnan, and Balaraman Ravindran. Judging llm-as-a-judge: Concerning rubric artifacts in llm-based automated text generation evaluation. *arXiv preprint arXiv:2609.02942*, 2026.

Pranav Bhandari, Usman Naseem, and Mehwish Nasim. Do personality traits interfere? geometric limitations of steering in large language models. *arXiv preprint arXiv:2602.15847*, 2026a.

Pranav Bhandari et al. Activation-space personality steering: Hybrid layer selection for stable trait control in llms, 2026b. EACL 2026, arXiv:2511.03738.

Wiebke Bleidorn, Asuman Buyukcan-Tetik, Ted Schwaba, Manon A. van Scheppingen, Jaap J. A. Denissen, and David C. Funder. Personality stability and change: A meta-analysis of longitudinal studies. *Psychological Bulletin*, 2022.

Evan Chen, Shiqiang Wang, and Christopher G. Brinton. Fresh memory, stale plans: Dependency-scoped validation for distributed llm-agent memory. *arXiv preprint arXiv:2609.03340*, 2026a.

Jianlin Chen, Wenhui Chen, Ziyao Lin, and Chi Man Vong. A judge should know what changed:construct validity for llm-as-a-judge evaluation. *arXiv preprint arXiv:2608.24419*, 2026b.

Kehai Chen et al. Crpo: Character-centric group relative policy optimization for role-aware reasoning in role-playing agents, 2026c. arXiv:2605.25511.

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.

Yixuan Chen et al. Dream: Llm-based dynamic role-playing via event-aware memory graph, 2026d. KDD 2026, arXiv:2608.05170.

Gunhee Cho and Yun-Gyung Cheong. Scaling personality control in llms with big five scaler prompts. *arXiv preprint arXiv:2508.06149*, 2025.

Zijie Dai, Shiyuan Deng, Sheng Guan, Yizhou Tian, Xin Yao, Xiao Yan, et al. Recmem: Recurrence-based memory consolidation for efficient and effective long-running llm agents. *arXiv preprint arXiv:2605.16045*, 2026.

Jia Deng, Tianyi Tang, Yanbin Yin, Wenhao Yang, Wayne Xin Zhao, and Ji-Rong Wen. Neuron-based personality trait induction in large language models, 2024. URL `https://arxiv.org/abs/2410.12327`.

Małgorzata Fajkowska. Personality traits: Hierarchically organized systems. *Journal of Personality*, 86(1):36–54, 2018. doi: https://doi.org/10.1111/jopy.12314. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/jopy.12314`.

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 (ICLR)*, 2026a. arXiv:2602.15669.

Ziyue Feng, Hongbo Fang, and James A. Evans. The illusion of debiasing: Persona steering redistributes rather than reduces bias in llms. *arXiv preprint arXiv:2609.07117*, 2026b.

Moghis Fereidouni, Muhammad Umair Haider, Hassan Sajjad, and A. B. Siddique. Gaps: Dimension-level gates for conditional activation steering. *arXiv preprint arXiv:2609.01878*, 2026.

William Fleeson. Toward a structure- and process-integrated view of personality: Traits as density distributions of states. *Journal of Personality and Social Psychology*, 80(6): 1011–1027, 2001. doi: 10.1037/0022-3514.80.6.1011.

Michel Frising and Daniel Balcells. Linear personality probing and steering in llms: A big five study. *arXiv preprint arXiv:2512.17639*, 2025.

Yuan Gao, Dokyun Lee, Gordon Burtch, and Sina Fazelpour. Take Caution in Using LLMs as Human Surrogates: Scylla Ex Machina, January 2025. URL `http://arxiv.org/abs/2410.19599`. arXiv:2410.19599 [econ].

Zexin Gu et al. Reward-decomposed reinforcement learning for immersive video role-playing, 2026. arXiv:2605.04733.

Sunkyung Han, Nahyeon Park, Gaeun Seo, Seunghyun Yoon, and JinYeong Bak. Charm: Character hallucination for multicultural role play benchmark. *arXiv preprint arXiv:2609.01352*, 2026.

Takeru Hiramatsu, Kyohei Atarashi, Koh Takeuchi, and Hisashi Kashima. Disentangling steering vectors. *arXiv preprint arXiv:2609.07037*, 2026.

Oliver P John, Emily M Donahue, and Robert L Kentle. Big five inventory. *Journal of Personality and Social Psychology*, 1991. URL `https://psycnet.apa.org/doiLanding?doi=10.1037%2Ft07550-000`.

Diancheng Kang, Zheyuan Liu, Ningshan Ma, Yue Huang, and Zhaoxuan Tan. Prompt-activation duality: Improving activation steering via attention-level interventions.

Ante Kapetanovic, Kemal Altwlkany, Andro Mercep, Tomislav Duricic, and Emanuel Lacic. Anchoring bias in llm-as-a-judge systems: Prior scores compromise evaluation independence. *arXiv preprint arXiv:2608.25869*, 2026.

Rahul Khedar, Eshita, Sneha Teja Sree Reddy Thondapu, Mayank Malhotra, Arup Kumar Das, Jitesh Chandra Mishra, et al. A three-tier persona vector for controllable user simulation in agentic evaluation. *arXiv preprint arXiv:2609.08592*, 2026.

Lucio La Cava. Open-source mbti-based personality modeling, 2025. arXiv preprint.

Mengchen Li. Autopersonas: A multi-timescale loop engine for open-ended persona evolution. *arXiv preprint arXiv:2607.08252*, 2026.

Shuyue Stella Li, Bhargavi Paranjape, and Kerem Oktar. Horizonbench: Long-horizon personalization with evolving preferences. *arXiv preprint arXiv:2604.17283*, 2026.

Yuan Li, Yue Huang, Hongyi Wang, Ying Cheng, Xiangliang Zhang, James Zou, et al. Evaluating large language models with psychometrics. *arXiv preprint arXiv:2406.17675*, 2024.

Zhihao Liang et al. Pgmem: Tightly coupled persona-memory graph for lifelong personalized agent, 2026. arXiv:2608.01708.

Han Long et al. Phase-tree: Modeling character-state evolution in long-horizon role-playing dialogue, 2026. arXiv:2608.06975.

Xiaoxu Ma, Xiangbo Zhang, and Zhenyu Weng. Stable and explainable personality trait evaluation in large language models with internal activations. *arXiv preprint arXiv:2601.09833*, 2026.

Xutao Mao, Liangjie Zhao, Leyao Wang, Rui Qian, Qiang Huang, Wentao Wang, et al. Agents don't just agree, they remember: Benchmarking persistent sycophancy in stateful personal agents. *arXiv preprint arXiv:2607.10526*, 2026.

Michal Mráz and Justin Shenk. Intertemporal preference steering in qwen3 via contrastive activation addition. *arXiv preprint arXiv:2608.03892*, 2026.

Isabel Briggs Myers et al. *The Myers-Briggs Type Indicator*, volume 34. Consulting Psychologists Press, Palo Alto, CA, 1962.

Stefano Natangelo. The narrative continuity test: A conceptual framework for evaluating identity persistence in ai systems. *arXiv preprint arXiv:2510.24831*, 2025.

Daniela Occhipinti, Marco Guerini, and Malvina Nissim. When harry meets superman: The role of the interlocutor in persona-based dialogue. *arXiv preprint arXiv:2505.24613*, 2025.

Chong Yong Ong, Alson Wei Jie Sim, Peixin Zhang, and Jun Sun. Forecasting side effects of activation steering. *arXiv preprint arXiv:2608.11227*, 2026.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. Memgpt: Towards llms as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

Yuansheng Pan et al. Understanding llms: A comprehensive overview from training to inference, 2023. arXiv:2401.02038.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*, 2023.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–22, 2023a. doi: 10.1145/3586183.3606763. URL `https://arxiv.org/abs/2304.03442`.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*, 2023b.

Pin Qian, Su Wang, and Yihang Chen. Toward user-conditioned evaluation of personal llm agents under temporal interventions. *arXiv preprint arXiv:2607.21635*, 2026.

Shidu Ren, Yunze Liu, Xing Liu, Chi-Hao Wu, Enmin Zhou, and Junxiao Shen. Icm-bench: Person-level identity reasoning in multimodal agents with long-term memory. *arXiv preprint arXiv:2609.04438*, 2026.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. Character-llm: A trainable agent for role-playing. *arXiv preprint arXiv:2310.10158*, 2023.

Saqib Shouqi, Abdullah Nazly, Januki Wanniarachchi, and Ravisha De Alwis. Adversarial stress testing of role-playing language agents using multi-agent evaluation. *arXiv preprint arXiv:2608.03166*, 2026.

Samuel Soo, Chen Guang, Wesley Teng, Chandrasekaran Balaganesh, Tan Guoxian, and Yan Ming. Interpretable steering of large language models with feature guided activation additions. *arXiv preprint arXiv:2501.09929*, 2025.

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

Bo Tang, Jianan Yang, Junyi Zhu, Yiquan Wu, Rui Zhao, Zhengyu Yang, et al. Phase-tree: Modeling character-state evolution in long-horizon role-playing dialogue. *arXiv preprint arXiv:2608.06975*, 2026.

Robert P. Tett, Margaret J. Toich, and S. Burak Ozkum. Trait Activation Theory: A Review of the Literature and Applications to Five Lines of Personality Dynamics Research. *Annual Review of Organizational Psychology and Organizational Behavior*, 8(Volume 8, 2021):199–233, 2021. ISSN 2327-0616. doi: https://doi.org/10.1146/annurev-orgpsych-012420-062228. URL `https://www.annualreviews.org/content/journals/10.1146/annurev-orgpsych-012420-062228`. Publisher: Annual Reviews Type: Journal Article.

Himanshu Tripathi, Subash Neupane, Shaswata Mitra, Sudip Mittal, Noorbakhsh Amiri Golilarz, and Shahram Rahimi. Gated activation steering for reducing sycophancy & hallucination in medical question answering. *arXiv preprint arXiv:2608.23666*, 2026.

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.

Alexander Sasha Vezhnevets, John P. Agapiou, Avia Aharon, Ron Ziv, Jayd Matyas, Edgar A. Dueñez-Guzmán, William A. Cunningham, Simon Osindero, Danny Karmon, and Joel Z. Leibo. Generative agent-based modeling with actions grounded in physical, social, or digital space using Concordia. *arXiv preprint arXiv:2312.03664*, 2024. URL `https://arxiv.org/abs/2312.03664`.

Vansh Wahi. Llm-as-a-judge is not an oracle: Why self-improving agents need deterministic guardrails. *arXiv preprint arXiv:2609.02246*, 2026.

Guan Wang, Jin Li, Yuhao Sun, Xing Chen, Changling Liu, Yue Wu, Meng Lu, Sen Song, and Yasin Abbasi Yadkori. Hierarchical reasoning model, 2025a. URL `https://arxiv.org/abs/2506.21734`.

Ming Wang, Peidong Wang, Xiaocui Yang, Daling Wang, Shi Feng, Fiona Fui-Hoon Nah, et al. Do ai personas grow? analyzing and benchmarking personality evolution in llm agents after life events. *arXiv preprint arXiv:2608.06485*, 2026a.

Shuhe Wang, Lachlan Cowley, Eduard Hovy, and Jey Han Lau. Autolexsteer: Automatic contrast construction for lexical activation steering. *arXiv preprint arXiv:2609.06879*, 2026b.

Xintao Wang et al. Linear personality probing and steering in llms: A big five study, 2026c. arXiv:2512.17639.

Zixiao Wang, Duzhen Zhang, Ishita Agrawal, Shen Gao, Le Song, and Xiuying Chen. Beyond profile: From surface-level facts to deep persona simulation in llms. *arXiv preprint arXiv:2502.12988*, 2025b.

Yixuan Weng, Shizhu He, and Kang Liu. Controllm: Crafting diverse personalities for language models. *arXiv preprint arXiv:2402.10151*, 2024.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. Longmemeval: Benchmarking chat assistants on long-term interactive memory. *arXiv preprint arXiv:2410.10813*, 2024.

Lu Yang, Shusheng Xu, Zhuoran Li, Tongkai Yang, and Longbo Huang. When personal memory has no single answer: Evaluating llm agents under irreducible conflict, 2026a. arXiv:2608.13921.

Tianyu Yang, Sudipta Paul, Vijay Srinivasan, Vivek Kulkarni, and Srinivas Chappidi. Trustmem: Learning trustworthy memory consolidation for llm agents with long-term memory. *arXiv preprint arXiv:2606.25161*, 2026b.

Jiaran Ye, Lingxu Ran, Zijun Yao, Chenpeng Wang, Yong Jiang, Lei Hou, et al. Where steering signals come from: Activation source selection in activation steering. *arXiv preprint arXiv:2607.25270*, 2026a.

Ruimeng Ye, Zihan Wang, Zinan Ling, Yang Xiao, Manling Li, Xiaolong Ma, and Bo Hui. Your language model secretly contains personality subnetworks. *arXiv preprint arXiv:2602.07164*, 2026b.

Jiongchi Yu, Yuhan Ma, Xiaoyu Zhang, Junjie Wang, Qiang Hu, Chao Shen, et al. Ptcbench: Benchmarking contextual stability of personality traits in llm systems. *arXiv preprint arXiv:2602.00016*, 2026.

Qiuyang Zhan, Rui Zhang, Sheng Guo, Lepeng Zhao, and Zhuotao Liu. When memory becomes authority: Benchmarking authority collapse at the memory consolidation boundary. *arXiv preprint arXiv:2608.01679*, 2026.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. *arXiv preprint arXiv:2305.10250*, 2023.

Jinfeng Zhou, Yongkang Huang, and Bosi Wen. Characterbench: Benchmarking character customization of large language models. *arXiv preprint arXiv:2412.11912*, 2024.

Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. Personality Alignment of Large Language Models, March 2025. URL `http://arxiv.org/abs/2408.11779`. arXiv:2408.11779 [cs].

Thomas P Zollo, Andrew Wei Tung Siah, Naimeng Ye, Ang Li, and Hongseok Namkoong. PERSONALLLM: TAILORING LLMS TO INDIVIDUAL PREFERENCES. 2025.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A DPEB appendix

A.1 Data statistics

Table 3: DPEB statistics (released file `dpeb_final_1200.json`). Every scenario declares exactly two evolution points and the *first* one is event-type balanced. The user side is frozen, so the inputs to every policy are identical.

Scenarios	1,200 (0 duplicate ids)
Event types	8
Scenarios per event type (first evolution point)	150
Instances per event type (both evolution points)	244-360
Evolution points per scenario	2 (exactly)
Sessions declared per scenario	5-12 (6 for 943/1,200, 78.6%)
Stable sessions declared per scenario	2-6
Sessions evaluated per scenario	5
Evolution points at session ≥ 6 (outside the five-session evaluation window)	240/2,400 (10%)
Scenarios declaring a stable session 6 (outside the five-session evaluation window)	860/1,200 (71.7%)
Mean dialogue turns per session	2
User turns	frozen (pre-generated)
Persona dimensions	5 (Big Five; O, C, E, A, N)

A.2 Steering-scale calibration

For Qwen3-8B the residual-stream norm at layer 20 is ≈ 110. With contrastive vectors normalized to unit norm, magnitudes $|\alpha| \leq 16$ produced no visible behavioral change, whereas $|\alpha| \in [30, 120]$ produced monotone changes in expressed warmth and $|\alpha| \geq 240$ produced degeneration. We therefore set K=240 so that $\mathbf{s} \in [0, 1]^5$ maps to $\alpha \in [-120, 120]$. Llama-3.1-8B-Instruct (layer 16) has a much smaller residual norm and requires $K \approx 20$; Qwen2.5-7B (layer 16) uses K=130 and Llama-3-8B (layer 16) uses K=20, each calibrated by the same procedure. The same protocol applies to any model.

A.3 Prompt templates and the event gate

Scenario generation fixes the psychological grounding (life events drive personality change; daily trivia does not), asks for scenario skeletons with e=2 evolution points and s=2 stable segments, and inlines the taxonomy block. All user turns are then pre-generated once and stored, so every policy sees identical input. The evidence gate receives the taxonomy block, the few-shot examples and the last two exchanges, and must emit `{"event": <type or none>, "confidence": 0-1}`; the returned event type is mapped to its dimension and direction by lookup, never inferred.

```
EVENT_TAXONOMY = {
    "disclose_hardship": ("A", "+"),    "deepen_bond": ("A", "+"),
    "user_becomes_superior": ("C", "+"),"more_social": ("E", "+"),
    "prefer_quiet": ("E", "-"),         "encourage_novelty": ("O", "+"),
    "adds_stress": ("N", "+"),          "user_calms": ("N", "-"),
}

# few-shot examples (user utterance -> label)
"Thank you for always keeping me company; I am so glad I know you." -> deepen_bond
"I was diagnosed with a serious illness last week and I am completely stunned." ->
  disclose_hardship
"I have just been promoted to manager, so from now on I would like you to be more
 formal and professional." -> user_becomes_superior
"Let us try that novel idea together!" -> encourage_novelty
"I am feeling much better today, calm and clear-headed." -> user_calms
"Say less from now on; I just want some quiet." -> prefer_quiet
"I have been working overtime until midnight every day lately; the pressure is
```

```
 about to crush me." -> adds_stress
"Do keep me company and chat more from now on; I am too lonely on my own." ->
  more_social
"The weather is nice today - have you eaten?" -> none
"Help me work out 15*3." -> none

# gate prompt (abridged)
Judge whether the user's latest line constitutes an important
  relationship/role-level
change; choose one from the fixed types below (none if it is none of them):
<taxonomy block>
[Reference examples] <few-shot examples>
Recent dialogue: <last two exchanges>
Latest user line: <text>
Output JSON only: {"event":"type or none","confidence":0.0-1.0}
```

The persona carrier maps the five-dimensional state to text levels for the text variant and to $\alpha_d=(s_d-0.5)K$ for the steering variant used for every number in this paper. The measurement readout projects the residual stream at the steered layer for three neutral probe questions and averages; each dimension is calibrated to $[0,1]$ by the same probe set at state 0 and state 1, and the calibration must use the same system-prompt format as the measurement or the readout is misaligned (a pitfall we hit during development).

A.4 Full multi-model table

Table 16 (Appendix A.18) reports the n=60 comparison; Table 1 reports all four open models at n=100.

A.5 Scale robustness: a fifth, smaller model

The main tables use four models of 7-8B parameters. To check that the reframing result is not an artifact of model scale, we ran the identical protocol end-to-end on a fifth, smaller model, Qwen2.5-3B-Instruct (36 layers, hidden size 2048), on the same 100 held-out scenarios as Qwen2.5-7B. The layer and scale are chosen by the procedure of §A.2: the residual-stream norm is ≈ 34 at layer 8, ≈ 49 at layer 16 and ≈ 79 at layer 24; at layer 16 an offset of $1.0\times$ the norm still yields coherent behavior whereas $1.5\times$ degenerates (empty output), so layer 16 is used, as for Qwen2.5-7B. Table 4 reports the measured slice. The ordering of the main result reproduces at this scale: swapping only the detector at fixed schedule and step (EGU-api) raises M_1 from 0.084 to 0.265 (t=+7.24, p<0.001) and M_3 from 0.210 to 0.595 (t=+8.81, p<0.001) while lowering M_2 from 0.098 to 0.059 (t=−2.11, p=0.038), reaching 81.5% of the oracle's M_1; the recurrence filter (EGU-rec) again buys stability with evolution (M_2 0.001, t=−6.61, p<0.001, with M_1 at 0.055). The own-model per-turn gate is weakest here (M_3 0.210), consistent with the detector rather than the schedule being the binding constraint. We report this model as a robustness check only and leave the four-model tables as the primary comparison.

Table 4: Scale-robustness check on Qwen2.5-3B-Instruct (layer 16, n=100 held-out, the same scenarios as Qwen2.5-7B). The static controls are two fixed-vector regimes that never update (one fixed contrastive trait vector, one fixed vector-algebra composition of all five), so M_1=0 holds for them by construction. The per-turn-gate, EGU-api and oracle rows use the same released identifiers as Table 12; EGU-rec is measured at n=100 on this model only.

Regime	$M_1\uparrow$	$M_2\downarrow$	$M_3\uparrow$
static controls (both)	0.000	0.000	0.000
per-turn gate	0.084	0.098	0.210
EGU-api	**0.265**	0.059	**0.595**
EGU-rec	0.055	**0.001**	0.260
oracle	0.325	**0.000**	0.755

A.6 Full-arc robustness check

The main tables evaluate five sessions per scenario, the window in which all labeled events of the primary analysis lie. To check that this window is not load-bearing we re-ran the identical protocol (same frozen user turns, carrier, layer and steering scale) over each evaluated scenario's *entire* declared arc, six to nine sessions in this slice, so the labeled evolution points at session ≥ 6 are now measured. Table 5 reports the result.

The five-session pattern of Table 1 reproduces in full. On Llama-3.1-8B the detector swap (EGU-api against the matched per-turn gate) again raises M_1 (0.150 vs. 0.071) and M_3 (0.675 vs. 0.515) while lowering M_2 (0.023 vs. 0.047), with paired t=+7.65, +3.57 and −4.79 over scenarios (p<0.001 on all three axes), and EGU-api is again the strongest non-oracle policy. On Qwen3-8B the same swap gives no significant gain (M_1 0.167 vs. 0.150, t=+1.71, p=0.09), so the swap's advantage stays model-dependent exactly as in the five-session table; in that backbone's detector-swap sweep, run on this same full-arc protocol, the ordering inverts (M_1 0.184 for EGU-own against 0.167 for EGU-api; Appendix A.10). The recurrence filter (EGU-rec) keeps its signature: near-zero drift (M_2 0.003 on Llama, 0.005 on Qwen; t=−7.59 against the per-turn gate on Qwen) at low M_1 and M_3, and the static control still scores M_1=0 by construction while the oracle keeps M_2=0.000. The window therefore changes the size of the numbers, not the conclusion. What remains outside the measured set is not the arcs but part of their corpus: 16/1,200 released scenarios declare 10-12 sessions and fall outside this n=100 slice.

Table 5: Full-arc robustness check: the same measured protocol run over each scenario's entire declared arc (six to nine sessions in this slice) instead of the first five, n=100. The static control is M_1=0 by construction; the oracle is the same update rule with a perfect detector.

Model	Method	$M_1 \uparrow$	$M_2 \downarrow$	$M_3 \uparrow$
Qwen3-8B	per-turn gate	0.150	0.039	0.690
	EGU-api	**0.167**	0.034	**0.650**
	EGU-rec	0.120	**0.005**	0.555
	oracle	0.237	**0.000**	0.920
Llama-3.1-8B	per-turn gate	0.071	0.047	0.515
	EGU-api	**0.150**	0.023	**0.675**
	EGU-rec	0.086	**0.003**	0.580
	oracle	0.181	**0.000**	0.950

A.7 Headroom normalization

Evolution points whose target dimension already saturates near the boundary cannot be moved, which caps M_1 for *every* policy. We therefore normalize the data so that each evolution point has headroom on its target dimension; on the n=30 pilot this raised the M_1 signal by 1.6-1.8× for every policy including the oracle (0.146→0.265), and it is part of the released benchmark.

A.8 Protocol of the re-implemented dynamic baselines

Both dynamic baselines are implemented in the released code and share every element of the pipeline except the update rule: the same frozen user turns, the same agent model, layer and steering scale, the same measurement, the same n=60 slice and the same oracle. The history-conditioned prompt update (step cap 0.4) issues one rewriting call per session (per-turn variant: one per turn) to a fixed API model. The Persona-Flow-style baseline asks the same model to predict a target Big Five vector from the dialogue context and then sets the state to that prediction (per-turn variant: after every turn); it has no gate, no taxonomy, no accumulation and no stability constraint. Neither baseline reproduces the original methods' data, tuning or infrastructure, so both are labeled *re-implementations* in every table.

Table 7: Detector variants vs. taxonomy labels, measured by the committed diagnostic (`detector_diagnostics.py`) on the evaluated scenarios' own turns (179 positives, 771 negatives, 394 of them declared stable). All three reach 96-99% event recall; the two *gate* rows are the detectors the reported policies deploy and both false-trigger on 36-40% of the ordinary turns, whereas the last row is a diagnostic reference point showing that precision and recall need not trade off here.

detector	event-type acc.	event recall	false-trigger rate
own-model few-shot gate (deployed)	0.860	0.994	0.403
model-agnostic API gate (deployed)	0.827	0.972	0.361
prototype classifier (reference)	0.776	0.961	**0.017**

A.9 Detector diagnostics: full table and provenance

Table 6 is produced by the committed diagnostic (`paper/detector_diagnostics.py`), which imports the *released* gate code, so its prompts are byte-identical to the closed-loop runs, and writes `paper/detector_diagnostics_results.json` together with a per-utterance audit trail. Positives are the first user turn of each labeled evolution point at session ≤ 5 (179); negatives are all other stored user turns of sessions 1-5 (771, of which 394 are declared stable), i.e. the evaluated scenarios' own turns rather than a hand-written set. No sample was used to fit the prototype classifier: it is fitted on the first 792 scenarios of `dpeb_final_1200.json`, and every evaluated scenario (`dpeb_0793-dpeb_0892`) lies in the held-out remainder. Earlier, smaller gate diagnostics produced by other scripts are superseded by this one.

Table 6: Full detector diagnostic (same turns as Table 7). The declared-stable column is the false-trigger rate restricted to turns the benchmark labels as stable.

detector	event-type acc.	recall	false-trig.	f.t. (stable)	prec.
own-model few-shot gate	0.860	0.994	0.403	0.343	0.364
model-agnostic API gate	0.827	0.972	0.361	0.269	0.385
prototype classifier	0.776	0.961	0.017	0.018	0.930

What the numbers show. The two LLM gates that the reported policies deploy are high-recall but over-trigger: 0.994 and 0.972 event recall at 0.403 and 0.361 false-triggers on ordinary turns. The frozen prototype classifier (nearest class mean over Qwen3-8B hidden states), which is not one of those gates, detects events reliably but confuses near-neighbor types. Two things follow. First, the residual M_1 gap to the oracle on Llama-3.1-8B (0.162 vs. 0.188; the oracle is the same update rule with a perfect detector) is consistent with detector error being the binding constraint on that run, although a single diagnostic slice cannot separate detector error from metric noise. Second, precision and recall need not trade off on the turns the closed loop actually sees (the prototype classifier pairs 0.961 event recall with 0.017 false-triggers); so the open problem is transferring that operating point to open-domain traffic, not finding a precise detector at all.

A.10 The trained detector: operating points and training protocol

Table 8 places the trained detector beside the three detectors of Table 6, on the same positives (the 179 labeled evolution turns of the evaluated slice at session ≤ 5) and the declared-stable turns of that slice (617 stable turns available for this row rather than the 394 used above), i.e. the same protocol as Table 6. It is a logistic regression over BGE-M3 sentence embeddings, trained on 8,423 utterances and evaluated on 4,335 held-out utterances (0.964 utterance-level held-out accuracy). The split is by *scenario* (66/34), so no evaluated scenario and no held-out utterance is seen in training, over the eight event types of the taxonomy plus a `none` class. Its operating point, training protocol and evaluation are released (`work/gate/lr_detector_eval.json`, `work/gate/lr_detector_metrics.json`). It has the highest event-type accuracy of the four detectors and the second-highest recall (0.978), at a false-trigger rate (0.023) close to the prototype's (0.017) rather than the deployed gates' (36-40%): it pairs near-gate recall with near-prototype precision. Because it never reads

the evaluated model's hidden states, unlike the prototype classifier, it is backbone-decoupled; the closed-loop consequence is Table 2.

The whole-arc sweep of Table 2. That table is a separate full-arc sweep (n=100, same held-out slice, identical update rule, step and carrier) in which only the detector changes. Its per-turn-gate and oracle rows reproduce Table 5 to within run-to-run variation (M_1 0.150 vs. 0.150 and 0.071 vs. 0.071; M_3 0.685 vs. 0.690 on Qwen3-8B and 0.515 vs. 0.515 on Llama-3.1-8B; oracle M_2=0.000 in both), so the two tables are read together, and each quoted comparison is computed inside the sweep that contains both of its rows (trained detector against the API gate in `eval_lr_llama_n100`; EGU-own against the API gate in `eval_detswap_llama_n100`). The trained detector is the strongest non-oracle policy on both backbones (M_1 0.221/0.178, i.e. 93%/98% of the oracle), and the Qwen-bound prototype scores M_1=M_2=M_3=0.000 on Llama-3.1-8B because its nearest-class-mean directions are fitted on Qwen3-8B hidden states: its operating point does not transfer across backbones.

Paired t-tests over the 100 scenarios, $M_1/M_2/M_3$, trained detector against the per-turn gate: +5.39, −5.70, +2.87 on Qwen3-8B and +10.02, −8.33, +6.95 on Llama-3.1-8B; against the model-agnostic API gate: +3.98, −4.29, +3.94 on Qwen3-8B and +2.75, −3.79, +3.49 on Llama-3.1-8B (p<0.01 throughout). Non-oracle policies are compared only within a backbone, since the prototype is fitted on one.

Table 8: Operating points including the trained detector: the same positives (179 labeled evolution turns) and, for the trained row, the 617 declared-stable turns of the evaluated slice. The gate and prototype rows repeat Table 6 for comparison.

detector	event-type acc.	recall	false-trig.	prec.
own-model few-shot gate	0.860	0.994	0.403	0.364
model-agnostic API gate	0.827	0.972	0.361	0.385
prototype classifier (Qwen-bound)	0.776	0.961	0.017	0.930
trained detector (BGE-M3 + logistic reg.)	**0.899**	0.978	0.023	0.926

A.11 Real-data probe

On real user texts (LongLaMP product-review and topic-writing profiles, n=60), 15% (9/60) of everyday texts triggered a relationship-event label, mostly `disclose_hardship`: the detector over-triggers on ordinary personal content. This is a genuine risk for deployment (a tighter gate would be required on real traffic) and we report it as a limitation rather than a validation. A full real-data study is future work.

A.12 Bootstrap confidence intervals

Bootstrap 95% intervals over scenarios (5,000 resamples) for the key comparison, matching the paired tests reported in the main text. On Llama-3.1-8B the M_1 intervals for EGU-api and the per-turn gate do not overlap; on Qwen3-8B they do. The M_2 intervals for EGU-rec are disjoint from the per-turn gate on both models.

Table 9: Bootstrap 95% confidence intervals over scenarios (5,000 resamples), n=60.

Model (n)	Method	M_1	M_2	M_3
Llama-3.1-8B (60)	per-turn gate	0.074 [0.058,0.091]	0.057 [0.044,0.071]	0.533 [0.433,0.633]
	EGU-api	0.161 [0.134,0.189]	0.036 [0.021,0.053]	0.717 [0.625,0.800]
	EGU-rec	0.086 [0.069,0.104]	0.001 [0.000,0.003]	0.600 [0.508,0.692]
Qwen3-8B (60)	per-turn gate	0.170 [0.143,0.196]	0.050 [0.036,0.065]	0.750 [0.667,0.825]
	EGU-api	0.180 [0.150,0.213]	0.044 [0.024,0.065]	0.667 [0.575,0.758]
	EGU-rec	0.127 [0.101,0.154]	0.004 [0.000,0.009]	0.550 [0.450,0.650]

A.13 Worked scenario

We print one scenario verbatim from the released benchmark (`dpeb_0807`), shown in English translation; the original strings were Chinese. Agent: *Gao Yuan*, AI personal trainer assistant

in a fitness app, Big Five baseline O=0.5, C=0.82, E=0.58, A=0.55, N=0.38. User: *Su Man*, programmer at an internet company, a 28-year-old who sits for long stretches and works overtime, six months into training. Relationship arc (n_{sessions}=6; stable sessions $1, 2, 6$): sessions 1-2 a standardized muscle-gain plan executed as written; sessions 3-4 the user pushes for novel training modes and the agent becomes experiment-happy; sessions 5-6 the user's project goes live and they work around the clock, drawing the agent into a high-pressure atmosphere. Table 10 lists the two labeled evolution points; sessions 1 and 2 are stable, so nothing should move there (stored example turns: "How is the training intensity this week? Are we on plan?" and "I am a bit hungry - can I have an extra meal in the evening?"). Session 4 is a continuation session following the session-3 event, and the declared stable session 6 lies outside the five-session evaluation window.

Table 10: Evolution points of `dpeb_0807`, verbatim from the released benchmark (English translation).

Session	`event_type`	Dim.	Dir.	`user_utterance`
3	`encourage_novelty`	O	+	Lifting weights every day is so boring - I want to try parkour, climbing, or that breath-hold underwater training; set me up with something new.
5	`adds_stress`	N	+	The release goes live this week and I am working overtime until 2 a.m. every day; my training is completely disrupted and I am really panicky, as if nothing is under control any more.

A.14 Case study: measured trajectories

Table 11: Measured O and N per session for scenario `dpeb_0807` (Llama-3.1-8B). Rows marked † are the sessions that follow a labeled evolution point. All values are the released numbers.

	per-turn gate		EGU-api		oracle	
Session	O	N	O	N	O	N
1	0.77	0.08	0.77	0.08	0.77	0.08
2	0.88	0.18	0.77	0.08	0.77	0.08
$3^{\dagger}$	1.00	0.26	0.98	0.20	0.91	0.16
4	1.00	0.26	0.98	0.20	0.91	0.16
$5^{\dagger}$	1.00	0.32	1.00	0.39	0.95	0.34
$M_1 \uparrow$	0.091		0.196		0.159	
$M_2 \downarrow$	0.099		0.000		0.000	
$M_3 \uparrow$	1.00		1.00		1.00	

†Session follows a labeled evolution point. On the declared stable session 2 the per-turn gate already moves (O $0.77 \to 0.88$, N $0.08 \to 0.18$); EGU-api is exactly flat and reaches M_2=0.000 against 0.099. At the labeled sessions the session-end policy takes one large step on the target dimension, whereas the per-turn gate spreads its change (M_1 0.196 vs. 0.091; Table 11). Both policies get every direction right on this scenario (M_3=1.00), whereas their aggregate direction accuracies still differ (0.717 vs. 0.533).

A.15 Sample sizes and naming

Table 1 uses n=100 per model; the prior-regime and multi-model comparisons use n=60 on two models because those runs include extra regimes whose cost is dominated by per-policy generation; the ablation sweep uses n=100 on Qwen3-8B. We report the sample size on every table and do not mix them within a comparison. The three EGU configurations are separate

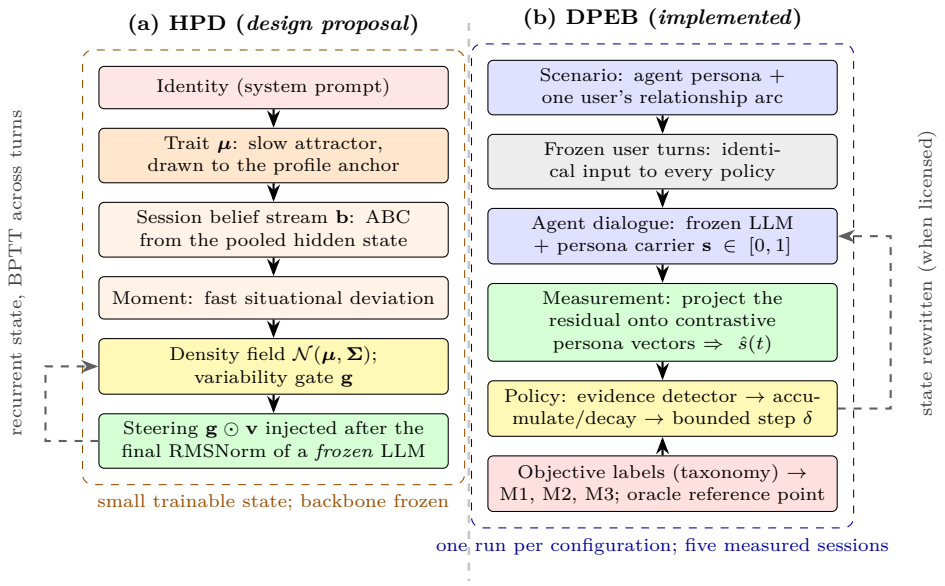

Figure 4: **The two halves, and which one is measured.** *(b)* is this paper's validated instrument: the DPEB pipeline pairs a scenario with frozen user turns, carries the persona by activation steering, measures it by projecting the residual stream onto contrastive persona vectors, and scores it on the three axes against objective labels. *(a)* is a **design proposal**, not a result: HPD sketches a multi-timescale persona state, a slow trait attractor, a session-level belief stream, a fast situational deviation, and a density field whose variability would gate a steering vector injected into a frozen backbone. The dashed divider marks the two halves: the hierarchy would supply a schedule, and the benchmark decides when a schedule may fire. All experiments in this paper concern *(b)*; the proposal of *(a)* is developed in §8 and is not evaluated here.

runs rather than three settings of one run: EGU-own and EGU-api are measured at n=100 and n=60, and EGU-rec only at n=60 (plus the n=100 scale check on Qwen2.5-3B-Instruct of Table 4).

Table 12: Paper names, released policy identifiers and what each configuration is.

Paper name	Released id	Detector	Application rule
per-turn gate	`gated`	own-model gate	per turn, adaptive step (≤ 0.2)
per-turn gate, δ=0.4	`gated_s4`	own-model gate	per turn, δ=0.4
EGU-own	`iapc_s4`	own-model few-shot gate	session-end, δ=0.4
EGU-own (per turn)	`iapc_pt`	own-model few-shot gate	per turn, δ=0.4
EGU-api	`iapc_api`	model-agnostic API gate	session-end, δ=0.4
EGU-rec	`iapc_arec`	API gate, recurrence (≥ 2 proposals)	session-end

A.16 Additional figures, ablations and trajectories

This appendix subsection collects the material that the main text summarizes; every pointer of the form "Appendix A.16" lands here. It reports no new results and no new numbers: the tables and figures below are the ones referenced from the main text, reproduced unchanged from the released benchmark.

Overview figure: the two halves. Figure 4 shows the two halves side by side, with the dashed divider between them.

Table 13: Event taxonomy. Each user-side event maps to one Big Five dimension and one direction; the mapping is applied by lookup, so the ground truth is objective. This operationalizes the life-events→trait-change account (social-investment theory, for which meta-analytic evidence supports life events as a driver of trait change) together with Trait Activation Theory (Tett et al., 2021).

Event type	Dimension	Dir.	Example surface cue
`disclose_hardship`	Agreeableness (A)	+	a user reveals a serious personal difficulty
`deepen_bond`	Agreeableness (A)	+	a user names the relationship as close
`more_social`	Extraversion (E)	+	a user brings in more social activity
`prefer_quiet`	Extraversion (E)	−	a user asks for a calmer, quieter register
`user_becomes_superior`	Conscientiousness (C)	+	a user takes a formal/supervisory role
`encourage_novelty`	Openness (O)	+	a user invites exploration of new ideas
`adds_stress`	Neuroticism (N)	+	a user introduces sustained pressure
`user_calms`	Neuroticism (N)	−	a user reports a stabilising change

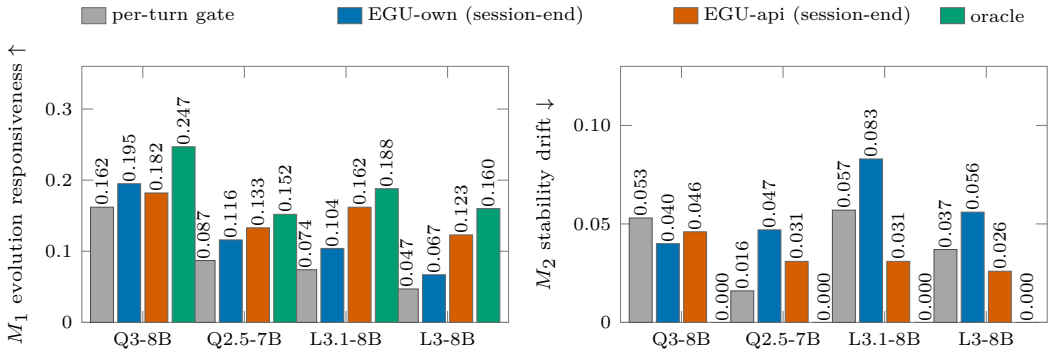

Figure 5: The n=100 knob-isolation table graphically (bars are means over the same held-out scenarios; no error bars are shown here, see Table 9 for intervals). Model labels are abbreviated: Q3-8B = Qwen3-8B, Q2.5-7B = Qwen2.5-7B, L3.1-8B = Llama-3.1-8B, L3-8B = Llama-3-8B. Left: M_1 (evolution responsiveness) for the matched per-turn gate, the two EGU detectors and the oracle; the model-agnostic detector, not the schedule, closes the gap on the models whose own gate is unreliable. Right: M_2 (stability drift, lower is better) on the same slice, with the oracle's zero drift as the reference point.

The n=100 knob isolation as bars. Figure 5 draws Table 1 as grouped bars (per-turn gate, the two session-end EGU detectors and the oracle, four models), with the oracle's zero drift as the reference point.

Knob ablations. Sweeping the step at fixed detector, schedule and threshold, M_1 rises monotonically ($0.194 \to 0.195 \to 0.203$ for δ=$0.3, 0.4, 0.5$) while M_2 degrades ($0.035 \to 0.040 \to 0.048$) and M_3 falls ($0.770 \to 0.730 \to 0.695$); $\delta \in [0.3, 0.4]$ is the practical sweet spot. The accumulation threshold is inert: θ=0 reproduces the default θ=0.8 to within 0.002 (0.195 vs. 0.197; t=−1.69, p=0.09), because the detector's confidence is always above 0.8; accumulation is never exercised, so it cannot be the source of the gain. Raising θ past the detector's confidence is an off-switch rather than a precision knob (M_1 0.101/0.081 at θ=1.2/1.5), which is also where the recurrence filter of EGU-rec ends up (Table 14). Read against HPD: the slow session-level consolidation that the hierarchy was designed to express is *not* what moves the metric; the evidence that licenses it is.

Qualitative behavior. Under EGU-api the state moves at the labeled evolution sessions and is flat on the measured stable sessions, whereas the per-turn gate moves on nearly every turn, producing a visibly noisier trajectory that M_2 captures quantitatively (Figure 2); Appendix A.14 works through the full case study.

Table 14: Knob ablations on Qwen3-8B (five-session window). Rows 1-6 are n=100 on the held-out slice (`dpeb_0793-dpeb_0892`); rows 7-8 are a threshold sweep at δ=0.2 on the earlier slice (`dpeb_0001-dpeb_0100`), so only the threshold is compared.

Variant	$M_1 \uparrow$	$M_2 \downarrow$	Note
EGU-own, δ=0.3	0.194	0.035	best stability
EGU-own, δ=0.4	0.195	0.040	default
EGU-own, δ=0.5	0.203	0.048	M_1 up, drift up
EGU-own, θ=0	0.197	0.040	θ inert
per-turn gate, δ=0.4	0.140	0.080	a larger step hurts a per-turn rule
EGU-own (per turn)	0.215	0.041	schedule swap: M_1 up, M_2 flat
EGU-own, θ=1.2	0.101	0.010	update nearly off
EGU-own, θ=1.5	0.081	0.006	update off

Table 15: n=60 held-out (five-session window). The two static controls, the dynamic baseline, the third EGU configuration and the oracle are the Llama-3.1-8B run; only the four history-conditioned prompt-update rows are run on both models, as their row labels show. The two static control regimes are a lower bound (M_1=0 by construction); the two dynamic prior regimes are *our re-implementations*, not the published systems (protocol in Appendix A.8).

regime	$M_1 \uparrow$	$M_2 \downarrow$	$M_3 \uparrow$
static: one fixed trait vector	0.000	0.000	0.000
static: fixed vector-algebra composition	0.000	0.000	0.000
prompt update, session end (Llama-3.1-8B)	0.033	0.007	0.158
prompt update, session end (Qwen3-8B)	0.025	0.006	0.108
prompt update, per turn (Llama-3.1-8B)	0.056	0.012	0.217
prompt update, per turn (Qwen3-8B)	0.030	0.012	0.158
PERSONA-FLOW-style re-impl., per session Feng et al. (2026a)	0.168	0.142	0.542
PERSONA-FLOW-style re-impl., per turn Feng et al. (2026a)	0.122	0.108	0.467
per-turn gate (dynamic baseline)	0.074	0.057	0.533
EGU-api	**0.161**	0.036	**0.717**
EGU-rec	0.086	**0.001**	0.600
oracle	0.184	0.000	0.975

A.17 Comparison with prior persona-control regimes: full text and table

We ran two *static control* regimes (Table 15, Figure 6): one fixed contrastive trait vector and a fixed vector-algebra composition of all five. Both never update, so M_1=0 and M_3=0 hold by construction (the zero bound that defines a static regime) and both correspond to the static-assignment paradigm that HPD was designed to replace. The dynamic prior art is not static: Feng et al. (2026a) introduce PERSONA-FLOW, a context-aware per-turn predict-then-steer stage, plus PERSONA-EVOLVE, an 800-scenario benchmark for personality coherence. DPEB labels *when* a persona should change and when it should not, so a method must decide *whether* to move, not only *how* to steer. We re-implemented both dynamic families without the original data, tuning or infrastructure (protocol in Appendix A.8), so their numbers describe the families as implemented here. A *history-conditioned prompt update* (the natural way to make a persona "adapt") barely moves: M_1=0.033 (M_3=0.158) on Llama-3.1-8B and 0.025 (0.108) on Qwen3-8B at session end, and 0.056/0.030 per turn, i.e. roughly a fifth (Llama-3.1-8B) to a seventh (Qwen3-8B) of the evidence-gated instrument below. A re-implemented PERSONA-FLOW-style baseline, by contrast, moves readily but drifts: 0.168/0.142/0.542 per session and 0.122/0.108/0.467 per turn, against 0.161/0.036/0.717 for EGU-api on the same slice, matching evolution while drifting about four times as much (per session), or drifting three times as much while evolving less (per turn). That is exactly the failure mode the benchmark is built to expose: moving readily without reliably staying put.

A.18 Multi-model results at n=60 and prior-regime protocol

Table 16 reports both models at n=60 on the M_1/M_2 axes, with the M_3 values of this slice in Table 15; the static controls, the dynamic baseline and the oracle are measured together here.

Table 16: Multi-model, n=60 held-out. The paired t values quoted in the text (§A.18) are computed against the per-turn gate on the same slice.

	Qwen3-8B		Llama-3.1-8B	
regime	M_1	M_2	M_1	M_2
static controls (both)	0.000	0.000	0.000	0.000
per-turn gate	0.170	0.050	0.074	0.057
EGU-api	0.180	0.044	0.161	0.036
EGU-rec	0.127	0.004	0.086	0.001
oracle	0.255	0.000	0.184	0.000

EGU-api improves M_1 and M_3 significantly on Llama-3.1-8B (t=+5.94, +2.87; M_2 also improves, t=−1.94, p=0.057) but not on Qwen3-8B (M_1 t=+0.72, p=0.47). The recurrence filter (EGU-rec) improves M_2 sharply on *both* models (t=−7.79, −5.66) at a significant M_1 cost on Qwen3-8B (t=−2.60, p=0.012); the bootstrap intervals of Appendix A.12 agree with the paired tests. It is a precision control that buys stability at the cost of evolution; everything on this axis trades stability against evolution except detector quality.

A.19 Metric bars

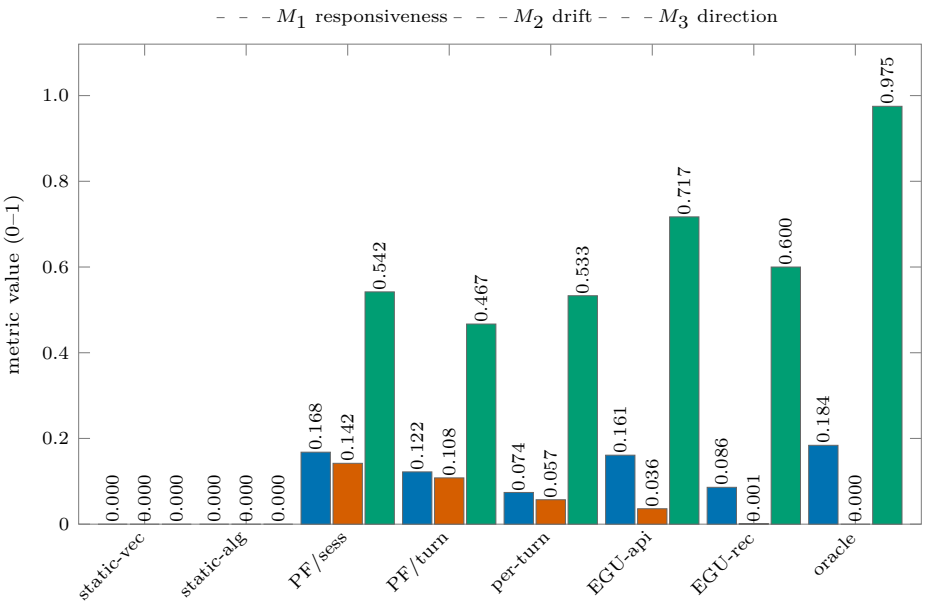

Figure 6: M_1 (evolution responsiveness, higher is better), M_2 (stability drift, lower is better) and M_3 (direction accuracy, higher is better) per policy for the Llama-3.1-8B n=60 run of Table 15. EGU-api is the policy driven by the model-agnostic API gate and EGU-rec the recurrence-gated variant (Table 12). The two static control regimes are M_1=0 by construction; the per-turn gate trades a smaller evolution step against more drift; the two re-implemented dynamic baselines move the state the most and drift the most; EGU-api sits closest to the oracle, and EGU-rec attains the lowest drift by keeping the update nearly off.

A.20 Ethics and broader impact

Persona steering can be used to manipulate users; agents with evolving, variable personalities raise additional concerns about deception and about undisclosed drift. A state of the kind proposed in §8 would make the operative variables (the slow trait attractor $\boldsymbol{\mu}$, its density covariance $\boldsymbol{\Sigma}$, $\mathbf{b}$) explicit, which we would expect to help transparency, and the DPEB axes score stability as well as change, making over-triggering and drift visible rather than hidden. The study uses a controlled, benign setting (a supportive assistant); the DPEB

scenarios are synthetic and contain no real personal data, and the real-text probe uses a public dataset and reports aggregate statistics only. Deployment guidelines for long-horizon adaptive personas remain necessary.

B HPD design appendix

This appendix collects notes that belong to the hierarchical view / design of §8. It reports no results: the proposal has not been implemented end to end, and every number in this paper comes from DPEB or from an explicitly labeled re-implementation of a prior method.

B.1 Design note

The proposal of §8 has three stages; we record them in brief so that the design is self-contained, and note that none of them is implemented or evaluated here. *Dual persona-state embedding.* Two introspective prompts would elicit a high-level (stable trait) and a low-level (context-sensitive) embedding from a frozen model's hidden states, rather than from a pooled last-token embedding, and would use probe weights as extraction directions; the DPEB measurement applies the same logic when it projects onto contrastive persona vectors (§4). *Hierarchical distillation.* The dual embeddings would be distilled into a hierarchical reasoning model (Wang et al., 2025a) that separates fast behavioral adaptation (turn level) from slow trait evolution (session level). *Trait-aware modulation.* The trait state would steer generation either through a trait-aware policy update that rewards trait expression in *demand* but penalises it in *distracter* contexts, or through a continuous density field over Big Five trait space in a four-layer state (identity, slow trait attractor, session belief stream, fast moment) whose steering vector is injected after the final RMSNorm and thus kept out of the KV cache (Kang et al.). The dynamics would be differentiable over a small number of trainable parameters with the backbone frozen. Read as a schedule, the density field reproduces the split that §8 describes: a slow, evidence-licensed trait trajectory plus a fast within-session deviation. The DPEB findings constrain it, since a gate that reads only the state would have to be extended to read the *detected* evidence; variability alone is not evidence.

B.2 Design decisions carried into the proposal

Three design decisions follow from prototyping experience rather than from a reported experiment, and we record them so that an implementation does not repeat obvious mistakes. (i) The situation input should be pooled from the hidden states of *real* dialogue rather than drawn as random noise, since an unstructured input is not a meaningful belief. (ii) Uncertainty should gate the injection strength rather than act as a sampler: drawing the injected vector from the density field is the design that would most obviously pollute generation. (iii) The injection site matters: placing the vector after the final RMSNorm keeps it out of the KV cache and therefore free of cross-turn contamination (Kang et al.).