NAVIGATING THE OCEAN: AN ATLAS OF BIG FIVE PERSONALITY REPRESENTATIONS IN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Personality psychology studies people through both shared, population-level trait dimensions and the distinctive patterns that characterize unique individuals. The Big Five personality model measures an individual's personality profile across five core dimensions (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism). Language-model research frequently uses them as five coarse dials, tuned and evaluated on surface-level alignment with trait labels. We ask if an open-weight model can match one person's measured Big Five profile, down to the facet, by navigating its rich network of trait representations, as fidelity to this static snapshot is a first step toward modeling dynamic individuals. Across five open-weight models and 420 people, we examine six distinct routes to personality representations, extracted during tasks such as interpreting behavior, reasoning about traits, and generating text. Directions for the same facet align across routes, far more than directions for different facets. We quantify agreement with each person using Lin's concordance correlation coefficient since a model can rank a population well while exaggerating everyone in it. Detailed natural-language profiles already reach high levels of agreement with a person's questionnaire scores (facet CCC .82 to .87), and lay readers recognize their broad domains. In open-ended autobiographical writing, some of that fidelity erodes in every model, though a targeted steering correction reliably restores part of it in smaller models. Profiles shift behavior across four tasks, from resource allocation to moral dilemmas, and steering by itself moves risk-taking, emotional appraisal, and value trade-offs toward each person's traits. Personality interventions produce clear, predictable tendencies, yet amplify them well beyond their human counterparts. To give researchers a tool for studying personality, we fine-tune an open-weight model to write and read psychometric profiles. Combining one writer with two readers yields life stories as faithful as our best steering-corrected ones, although none of the models were trained on life stories. Together, these findings support an atlas of interconnected personality representations and show where composing and calibrating them facet by facet improves fidelity to an individual's measured profile.

1 INTRODUCTION

Personality psychology studies people's traits at two levels. At the population level, traits are defined by how people differ from one another; at the individual level, the question is how one's traits, states, and situations fit together over time (Allport, 1937; Molenaar, 2004; Beck et al., in press). The Big Five is a population-level model that organizes stable, meaningful differences between people into five broad domains, commonly abbreviated as OCEAN: **O**penness, **C**onscientiousness, **E**xtraversion, **A**greeableness, and **N**euroticism. Each domain encompasses more granular traits called facets, which delineate specific patterns in an individual's thoughts, feelings, and behavior (DeYoung et al., 2007; Soto & John, 2017). For example, facets such as Sociability, Assertiveness, and Energy Level distinguish individuals with similar Extraversion scores (e.g., an extravert may enjoy social interaction while rarely taking charge in groups). The Big Five organizes these individual differences, though the framework does not provide a complete account of a person.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

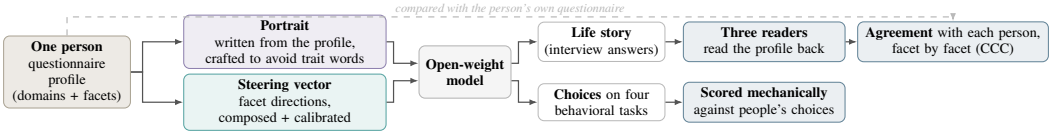


Figure 1: **Holding one person's profile, and checking it two ways.** A person's questionnaire profile reaches the model as a written portrait, as a steering vector composed from facet directions, or both. Independent readers recover the profile from the life story the model writes, and we score agreement with the person facet by facet; choices on four tasks are scored mechanically. How readbacks are scored is detailed in App. C.

Language-model research has mostly focused on the five core dimensions of the Big Five. One common approach instructs a model to behave in a manner that aligns with a short list of trait adjectives (*agreeable, supportive, compassionate*). After ingesting these system instructions, models can answer personality questionnaires in ways that reflect the assigned traits while showing little corresponding change in their choices on behavioral tasks (Han et al., 2026b). Recent work has expanded this approach by conditioning models on detailed profiles grounded in real human interviews and questionnaire responses (Park et al., 2026; Huang et al., 2026; Wigler et al., 2026). Moreover, a growing number of studies have sought greater control over personality expression by uncovering trait-related activation directions, or "trait vectors," and developing composable trait adapters (Chen et al., 2025; Feng et al., 2026; Baines et al., 2026). These methods have built the foundation for more precise personality conditioning, though their emphasis on broad OCEAN domains leaves the more granular, facet-level patterns within each domain largely unexplored.

Evaluation has lagged behind these methods. Many evaluations push one trait toward a high or low extreme and ask whether the model now sounds more like it (Chen et al., 2025; Banayeeanzade et al., 2026), or check whether a population of personas lands in the right rank order (Serapio-García et al., 2025). A model can pass both tests while exaggerating the assigned traits, ultimately distorting the individual personality profiles it is meant to express. For example, given a portrait that reads "Your physical pace is unhurried—you're not the person always in motion," one of the models we test wrote, as that person, "People see me as a force of nature, someone who is always in motion, always talking, always taking charge." Such discrepancies raise a broader question: how faithfully can models preserve an individual's measured personality profile as it generates text and makes decisions? Park et al. (2026) show the gap in miniature: their interview-based agents match the shape of each person's Big Five profile closely, yet miss the person's actual scores by about twice the error of that person's own retest. Other studies that condition models on real individuals evaluate over whole inventories, broad domains, or other constructs (Zhu et al., 2025; Wang et al., 2025; Huang et al., 2026; Venkit et al., 2026; Liu et al., 2026b), and none that we have found reports results per facet, let alone tests the same people on two instruments whose facets are cut differently, as we do with the BFI-2 and a 30-facet IPIP measure. This matters, because traits travel together inside models: pushing one moves others (Bhandari et al., 2026), and a domain score can hide where they go. When we steer a model toward compassion, its answers read as more outgoing; read facet by facet, they become more sociable and energetic but less assertive.

We ask three questions. **RQ1:** Through which routes do models represent personality, and how independently can the traits represented through these routes be controlled? **RQ2:** How closely can a model reproduce an individual's measured personality profile, down to the facet? **RQ3:** How does conditioning on that profile influence what the model writes and decides?

To address these questions, we extract trait-related activation directions from six routes: reading about traits, reasoning about traits, interpreting behavior, reading narratives, following dialogue, and writing from a profile. We combine these directions according to each individual facet's scores, calibrate their strengths using participants outside the evaluation set, and test them alone and alongside written personality portraits: detailed descriptions derived from individuals' questionnaire scores. We assess agreement between the recovered scores and the original questionnaire scores using Lin's concordance correlation coefficient (Lin, 1989), which accounts for systematic differences in trait levels and variability alongside rank order across people.

Our goal is to establish a control surface: a way to give a model stable, meaningful patterns of reasoning and behavior that can be set with precision, since an agent's personality shapes how users judge it (Rahman & Desai, 2026a). Real people's BFI-2 and IPIP-NEO profiles, which are two validated measures of Big Five traits, provide a demanding test because they contain specific combinations of facets and measured personality scores against which fidelity can be evaluated. Given the limitations of models as substitutes for human respondents (Gordon et al., 2026; Venkit et al., 2026), we examine how these profiles influence model behavior and how the resulting patterns compare with those observed in people. Across four tasks, written personality portraits reshape choices ranging from resource allocation to moral dilemmas. Steering vectors alone also shift risk-taking, emotional appraisal, and value trade-offs toward the traits they encode, even when the vectors come from a different person. These interventions produce systematic behavioral tendencies, with trait–behavior relationships substantially stronger than those observed in the human participants.

Our instruments describe population-level differences, and each target profile captures a single snapshot of an individual's personality. We therefore examine individual-level fidelity to these questionnaire profiles. For humans, traits describe distributions of momentary states that vary across situations (Fleeson, 2001; Spark et al., 2026). Establishing how faithfully models express a measured profile is a foundation for studying how personality expression changes across contexts and time.

Contributions.

1. **An atlas of personality representations (RQ1).** Personality is reachable through at least five families of activation directions and through composed trait adapters; the routes differ in alignment, and steering one facet moves others in patterns that resemble human facet correlations.
2. **Individual-level fidelity assessed by agreement (RQ2).** Scored facet by facet, detailed portraits achieve high agreement and lay readers recognize their broad domains. Life stories lose part of that fidelity in every model, mostly through a shared shift in level, which a targeted steering correction partly removes, reliably in smaller models.
3. **Behavioral effects of personality conditioning (RQ3).** Portraits and steering shift choices on four tasks, vectors from other individuals shift them toward those people's traits, and the effects run stronger than in the people behind the profiles, partly through the portrait writer.
4. **Open-weight tools for working with personal data.** We release adapters that write and read psychometric profiles; combined, they write life stories as faithful as our best corrected ones without any life-story training, and they can run locally on sensitive data.

2 Related work

Personality from prompts and profiles. Short trait prompts influence a model's self-reported personality more strongly than its actual choices on behavioral tasks (Han et al., 2026b). Richer conditioning, including interviews and profiles built from real people's responses, enables agents to predict those individuals' survey responses (Park et al., 2026). Narrative identity may help explain why life-story interviews let personas answer surveys more like the people they represent than short demographic prompts do (Kang et al., 2025), and richer profiles improve consistency across repeated simulations of the same persona (Bai et al., 2025). Wigler et al. (2026) decode personality from generated life-story interviews, and Tsfasman & Wigler (2026) show that profile-conditioned models' open-ended choices of fictional characters track the conditioning profile. We build upon these profile-building methods to examine how precisely models can reproduce an individual's measured personality profile and how conditioning on that profile shapes generated text and choices.

Personality inside the model. Contrastive activation steering (Turner et al., 2023; Rimsky et al., 2024) yields directions for character traits (Chen et al., 2025) and for the Big Five (Feng et al., 2026; Bhandari et al., 2026). Facet-level steering has been attempted with contrastive sparse autoencoders, evaluated at the domain level against fictional characters (Tang et al., 2026). At the level of weights, adapters can be conditioned on a continuous trait vector learned from text-estimated personality (Vu et al., 2026), or trait LoRAs can be composed into specific profiles (Baines et al., 2026). We compare routes for matching real individuals' full facet profiles, scored by agreement, and we test each route with written portraits.

Dependencies among personality traits. In people, each Big Five domain splits into correlated but separable facets and aspects (DeYoung et al., 2007; Soto & John, 2017). By contrast, the PERSONA framework reports approximately orthogonal trait vectors and limited cross-dimensional effects in a controlled test of Extraversion and Agreeableness (Feng et al., 2026). Its reported intervals nevertheless indicate small but systematic effects on the other trait, and Bhandari et al. (2026) find cross-trait effects comparable to the target effect, persisting even after the vectors are orthonormalized. Rahman & Desai (2026b) compared trait-vector geometry with human Big Five intercorrelations and found no reliable correspondence across the five domains. Across these studies, the independence of traits appears to depend on how personality is induced and how its effects are measured. Liu et al. (2026b) found a related tightness when models role-played real people from their Big Five answers: they answered other scales more consistently than the people did (mean α .87 vs. .75) and reproduced the human correlation structure in amplified form, largely because of measurement error in the human data, so model–human comparisons need reliability-corrected human estimates, which we report alongside the raw ones (§4). A domain direction carries the population's version of a trait, its usual mix of facets, so turning it pulls a person toward a stereotype of that trait. We compose at the facet level and calibrate the whole profile at once, and we revisit PERSONA's independence test and the human comparison facet by facet (§4.1).

Real individuals as targets. Personality Activation Search (PAS) aligned models to individual people's IPIP-NEO responses through a search over activation interventions, scoring absolute item-level distance to each person (Zhu et al., 2025). Wang et al. (2025) used 400 people's Big Five domain scores to prompt GPT-4 to role-play them and complete a personality inventory, and Huang et al. (2026) created prompted AI agents from 438 participants' BFI-2 item responses and compared the agents' responses on a secondary inventory with those of the corresponding participants. All of these evaluate through the conditioned model's own answers, over whole inventories or broad domains, and none reports results per facet. Observer-style personality ratings track a model's behavior more closely, while its self-reported personality more closely reflects the prompt (Huang & Hadfi, 2025). We extend this line in three directions: facet-level targets; a separate reader scoring text the conditioned model writes; and agreement decomposed into ranking, scale, and location (Lin, 1989). Following work on the accuracy of personality judgments, we also separate how well a model captures the average person's personality from how well it captures what makes each person different (Cronbach, 1955; Furr, 2008).

Traits, states, and individuals. Patterns of personality differences across people need not describe how personality is organized within an individual (Allport, 1937; Molenaar, 2004; Borkenau & Ostendorf, 1998; Beck et al., in press), and the covariation among traits may instead arise from shared process variables such as expectancies, values, and goals (Wood et al., 2015). Whole Trait Theory treats a trait as a distribution of momentary states (Fleeson, 2001), and experience sampling finds that between-person differences account for a large share of state variance without exhausting it (Spark et al., 2026). Language-model agents given the same prompted trait express it differently across social contexts (Han et al., 2026a). Our work tackles the first step of this program. We evaluate fidelity to an individual's questionnaire profile as a fixed target, providing a foundation for studying how personality expression changes across situations and over time.

3 METHODS

Terminology. A person's *profile* comprises their questionnaire scores across the five personality domains and their constituent facets (Compassion, for instance, is a facet of Agreeableness). A *portrait* is a written description generated from the profile. A *steering vector* is composed of facet directions weighted by an individual's scores and added to the model's internal activations during generation; a *route* identifies the task the model performed when the facet directions were extracted. A *read-back* is the profile an independent reader model recovers from a portrait or a life story. *Agreement* is measured using Lin's concordance correlation (CCC) between recovered and original questionnaire scores, computed across people for each facet and then averaged. CCC combines Pearson correlation (*ranking*) with a penalty for differences in *scale* (variability) and *location* (mean level), allowing us to assess whether recovered scores match the measured values. Predicting the same reference mean for everyone yields CCC = 0, which serves as our baseline.

3.1 PEOPLE, QUESTIONNAIRES, AND A HELD-OUT CONFIRMATION SET

Our targets are 420 people with facet-level questionnaire scores. Of these, 300 are from the PAPI collection (Zhu et al., 2025) and completed the IPIP-NEO (five domains, 30 facets; Johnson, 2014). The remaining 120 come from two samples of Bainbridge et al. (2022): 60 Australian undergraduates and 60 U.S. online workers. These participants completed both the BFI-2 (five domains, 15 facets; Soto & John, 2017), and an adapted version of the 120-item IPIP measure covering 30 facets (Maples et al., 2014). The adaptation replaces nine items. Together, these samples provide 540 questionnaire profiles. PAPI respondents answered the 300-item IPIP-NEO; their portraits were written from the 120 items of its short form (Johnson, 2014), and the other 180 items are held out. Steering settings were calibrated using participants outside the 420-person evaluation sample, with one small exception (App. I). The behavioral experiments (§4.3) use three further cohorts whose members made some of the choices we test (Szekely et al., 2021; Pulford et al., 2025; Yang et al., 2026); their instruments, and how their portraits were written, are in App. E.

3.2 PORTRAITS

A writer model (Claude Opus 5) generated a detailed natural-language portrait from each questionnaire profile, using instructions to avoid explicit trait terminology and numerical personality scores. Portraits do not contain numeric scores; the only digits in the 540 portraits of the 420 are clock times. A separate reader recovers each person's domain scores from the portrait alone with concordance of .791 to .940 across cohorts; on one cohort the recovered scores show greater variability than the source questionnaire scores. A second writer model (GPT-6 Astra) rewrote the portraits of one cohort, to test whether results depend on the writer (§4.3; App. F).

3.3 ROUTES TO A TRAIT REPRESENTATION

For each receiving model and questionnaire instrument, we built sets of trait-related activation directions while the model carries out different personality-related tasks: interpreting behavior (Behavioral), reading about traits (Semantic), reasoning about traits (Reasoning), reading narratives (Story), following dialogue (Dialogue), and writing from a profile (Writer, with related task states). The first five contrast material expressing a facet's high or low pole; the Writer family regresses construction participants' scores onto the model's task activations and does not require high–low contrasts (App. H). The two larger models received Behavioral, Semantic, and Reasoning banks from their own activations. As a weight-level route, we compose the trait adapters released by Baines et al. (2026)[1] into an adapter combination for each individual's profile (App. K).

3.4 COMPOSITION AND CALIBRATION

For each individual, we standardize facet scores relative to a construction sample and use them to weight the corresponding facet directions or select each facet's high or low pole according to the sign of its standardized score. The resulting directions are combined into a single vector and scaled by a gain. Layer, strength, and gain were selected separately for each model, route, and instrument on 58 people outside the evaluation sample (App. I).

3.5 LIFE STORIES

The conditioned model completes the same standardized life-story interview (McAdams, 2001), adapted for language models by Wigler et al. (2026), in the first person according to the supplied personality profile. We compare three conditioning conditions: a written portrait alone, a steering vector alone, and both together. In the steering-only condition, the template's profile field is left empty, so personality information is supplied only through the vector, except for the Gemma4-31B template adjustment described below. In the combined condition, the vector is either the person's full steering vector (larger models only) or a targeted correction: we generate one extra portrait-only story with a different seed, have a reader recover its profile $\hat{p}$, and steer with $g\,[C(p) - C(\hat{p})]$, where p is the person's profile, C is the route's facet composition, and the gain g is chosen from 0, 0.5,

[1]The repository's organization lists three of Baines et al.'s authors: `https://huggingface.co/datasets/persona-cartography/monorepo`.

and 1 per cell on the external people ($g = 0$ keeps the portrait-only story). The smaller models use an earlier form of the same correction: the gap is standardized by the construction sample's facet spread, projected onto the Behavioral axes, and matched to the model's typical activation size, with a fixed gain per model and instrument between 0.5 and 2 (App. I). The additional story is used to estimate the correction only if it terminates naturally and includes all twelve interview sections; otherwise, the correction is set to zero. This story is excluded from the outcome scoring. Across all models, a story is deemed valid if the output ends naturally before reaching the token limit. For Gemma4-31B's steering-only stories, the template's profile lines are removed rather than left empty, because it did not write life stories from an empty profile field; the interview itself is identical in every condition. The five models (Gemma4-31B, Gemma4-E4B, Qwen3.5-27B, Qwen3.5-4B, and Qwen3-8B) write stories with seed 914, temperature 0.7, and a 6,144-token cap.

3.6 Readers and agreement

Three reader models from different families recover personality scores from the generated life stories: a primary reader (GPT-5.6 Luna, accessed through OpenRouter with reasoning off) and two additional readers (Gemini-3-Flash and DeepSeek-v4.1-Flash, chosen on stories about people outside the evaluation set). Each reads the profile back from the life story in two ways: by answering that person's own questionnaire on the person's behalf, as in zero-shot text scoring (Wright et al., 2026), and by rating each domain and facet on a continuous scale. The questionnaire is answered in calls of at most 60 items. Only the readers' judgments are scored; the model never rates itself.

We compute agreement (Lin, 1989) at the facet level as our primary evaluation, and also report domain-level agreement. Evaluation stories that reach the token limit without terminating naturally remain in the analysis, with recovered scores replaced by the means of an external reference sample: the IPIP-NEO respondents of Johnson (2014) (data: Kajonius & Johnson, 2019) for the PAPI forms, the other 666 respondents in the same two samples of Bainbridge et al. (2022) for their IPIP form, and 2,128 respondents from the BESSI development samples (Soto et al., 2022; Sewell et al., 2022) for the BFI-2. We also assess split-half agreement between scores derived from two disjoint sets of questionnaire items, providing a reference for measurement consistency. App. C details the formulas, uncertainty intervals, paired tests, and additional evaluation measures.

Validating the readers against people. Forty-eight lay readers on Prolific, three per profile, rated sixteen portraits on the five domains. Their averaged ratings agreed with the source questionnaires at CCC .646 (mean absolute error .490), exceeding the agreement obtained in all 200,000 permutations that reassigned ratings to different individuals (permutation mean .124). Individual ratings were less consistent; this result therefore reflects the average of three readers per portrait. On the same sixteen portraits, our primary reader agreed with the questionnaires at CCC .933 and with the human average at .678. Twenty further readers rated ten life stories blind, at CCC .625 (App. A).

4 Results

4.1 RQ1: Alignment across routes and dependencies among facets

In the two larger models, directions extracted through different routes for the same facet are more closely aligned than directions for different facets (Fig. 2a). Behavioral and Reasoning directions have cosine similarities of .69 to .83, while Semantic directions align with both at .20 to .46: three to six times the agreement between different facets; random directions stay near zero. The smaller models show a similar pattern (Behavioral and Reasoning, .60 to .78; App. H).

In the three smaller models, steering one facet changed its recovered score by .31 to .80 on the 1–5 scale and other facets by .29 to .53 on average, as assessed by the primary reader. The pattern of these effects follows how facets correlate in people (Fig. 2b): $r = .48$ to .70 across facet pairs and .41 to .64 for pairs from different domains (permutation $p < .001$ for every model). Correcting the human correlations for reliability changed these associations by at most .01.

In our replication of PERSONA (Feng et al., 2026), Compassionate steering increased overall Extraversion ratings through higher Sociability and Energy facet-levels, while Assertiveness remained unchanged or decreased (App. G). *In short:* models reach personality through several aligned routes,

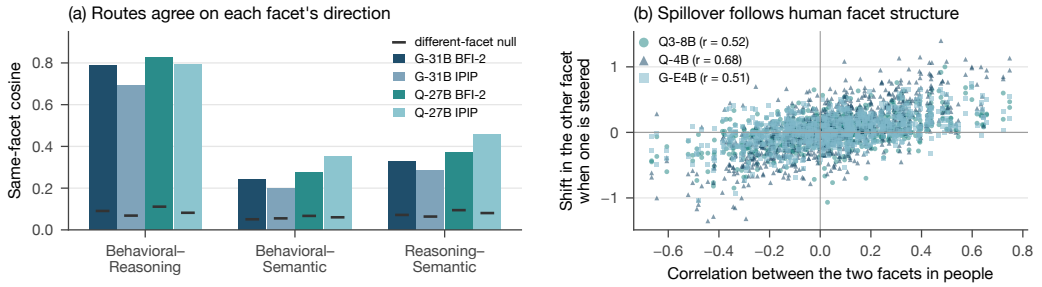


Figure 2: **Alignment across routes and effects across facets** (a) In the two larger models, same-facet directions from different routes are more closely aligned than different-facet directions (dashes). (b) In the three smaller models, cross-facet steering effects resemble human facet correlations (Johnson reference sample); each point represents one ordered facet pair.

Table 1: **CCC between recovered and questionnaire scores, computed across people by facet and domain.** Primary reader; scores for stories truncated by the token limit are replaced by reference-sample means. Multi-route cells report the median [minimum, maximum]; routes calibrated to zero correction retain portrait-only values. Split-half self-agreement uses disjoint item sets, portrait read-back uses the source items and can therefore exceed it. Additional readers, valid-only analyses, errors, and route-level results can be found in App. D.

		IPIP-NEO (PAPI, n=300)		BFI-2 (n=120)		IPIP-NEO-120 (n=120)	
Model	Condition	facet	domain	facet	domain	facet	domain
Everyone is average (floor)		.00	.00	.00	.00	.00	.00
Questionnaire agrees with itself		.71	.89	.61	.83	.65	.85
Portrait, read directly		.82	.86	.86	.89	.87	.89
Gemma4-31B	portrait-only story	.66	.73	.72	.76	.67	.76
	+ targeted correction	.70 [.66, .70]	.80	.74 [.73, .77]	.79	.69 [.67, .71]	.80
	+ steering	.60 [.52, .62]	.67	.74 [.74, .75]	.79	.62 [.58, .64]	.73
	steering only	.42 [.42, .45]	.51	.57 [.50, .59]	.66	.43 [.42, .47]	.57
Qwen3.5-27B	portrait-only story	.64	.72	.70	.74	.65	.74
	+ targeted correction	.64 [.61, .66]	.72	.76 [.70, .77]	.82	.68 [.65, .68]	.81
	+ steering	.50 [.42, .53]	.60	.59 [.56, .63]	.64	.52 [.46, .55]	.63
	steering only	.36 [.27, .36]	.41	.50 [.48, .53]	.58	.37 [.35, .42]	.45
Gemma4-E4B	portrait-only story	.57	.63	.61	.63	.57	.63
	+ targeted correction	.67	.78	.77	.84	.65	.78
	steering only	.02 [−.01, .09]	.02	.16 [.03, .64]	.18	.01 [−.01, .11]	.01
Qwen3.5-4B	portrait-only story	.59	.69	.64	.69	.58	.70
	+ targeted correction	.64	.78	.71	.77	.62	.76
	steering only	.13 [.00, .26]	.17	.36 [.32, .59]	.43	.13 [.02, .26]	.17
Qwen3-8B	portrait-only story	.38	.46	.42	.45	.36	.43
	+ targeted correction	.53	.66	.58	.66	.48	.65
	steering only	.07 [.05, .15]	.10	.18 [.04, .58]	.22	.10 [.04, .14]	.14

and steering one facet moves its human correlates with it, so the traits behave less like five independent dials than like people's profiles.

4.2 RQ2: Profile fidelity and targeted correction

Direct portrait read-back achieves facet CCC .82 to .87 and domain CCC .86 to .89 under the primary reader (Table 1). Agreement decreases during life-story generation for every model, instrument, and reader (45 of 45 paired comparisons, Holm $p < .05$), with primary-reader facet CCC falling to .64 to .72 in the larger models and .36 to .64 in the smaller ones. The decline involves weaker correlations with questionnaire scores and systematic shifts in mean trait levels, although additional

readers estimate smaller mean shifts in the larger models. The more unusual a person's profile, the more their story misses them (38 of 45 comparisons), but the correction helps them no less (App. D).

Targeted correction increases facet CCC in most smaller-model comparisons across all three readers, by up to .19. Gains in larger models are smaller and reader-dependent; one condition (Qwen3.5-27B, Behavioral, PAPI) reduces agreement under all three readers. The correction reduces mean shifts estimated by the primary reader used to construct it, but can overshoot under readers that estimated smaller initial shifts (Fig. 4b). Improvements primarily concern trait levels, with smaller changes in within-person profile shape (Fig. 4c); even the best corrected conditions remain .09 to .15 below direct portrait read-back in facet CCC. Adding full-profile steering to portraits reduces agreement in 15 of 18 larger-model comparisons under each reader and produces truncated generations in Qwen3.5-27B. Targeted correction outperforms full-profile steering in most comparisons (App. D). *In short:* portraits capture a person's profile closely; life stories lose part of it, largely as a shift in trait level that the correction partly repairs.

4.3 RQ3: Behavioral effects of personality conditioning

We test four tasks with forced choices, scored mechanically with no language-model judge: the social value orientation slider (Murphy et al., 2011), a collective-risk game (Milinski et al., 2008; Szekely et al., 2021), the primary emotional reaction to short events (Yang et al., 2026), and newly written moral dilemmas modeled on ConflictScope (Liu et al., 2026a). For all but the dilemmas, the portraits come from people who completed the corresponding tasks (App. E).

Under portrait conditioning, Agreeableness predicted more other-regarding allocations in the larger models ($r = .56$ to $.70$), even after controlling for prosocial vocabulary; this relationship was weak in the two smaller models tested. Neuroticism predicted more negative emotional appraisals, and the five traits predicted portrait-induced changes in model dilemma choices for held-out profiles. Across all five models, steering alone produced trait-related changes in appraisal and dilemma choices, with source-profile traits predicting the latter at $R^2 = .40$ to $.66$. Higher source-profile Extraversion was associated with larger collective-risk contributions. Donor-vector controls shifted appraisal and dilemma choices toward the donor's traits, while random directions did not produce comparable trait-related patterns. However, steering alone did not reliably reproduce the Agreeableness–allocation relationship. In the larger models' economic tasks, wrong-person and random vectors also shifted average choices about as much as matched vectors, indicating nonspecific effects alongside trait-related changes (App. E).

Across 384 appraisal events, human participants in the highest and lowest Neuroticism quartiles differed by 6.2 percentage points in negative primary reactions; models conditioned on their portraits differed by 20.4 to 38.7 percentage points (Fig. 3). On a 24-event subset, human forecasters given the same portraits predicted a much smaller gap. Changing the portrait writer reduced amplification in all three tested models, although Neuroticism–appraisal associations remained stronger than in people in two models (App. E). *In short:* portraits and steering move choices toward a person's traits, but more strongly than those traits move choices in people.

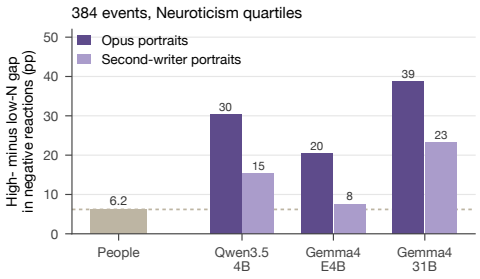


Figure 3: **Stronger Neuroticism–appraisal associations in models.** Difference in the percentage of negative primary reactions between the highest and lowest Neuroticism quartiles across 384 events. Models receive portraits from either of two writers. Human forecasting results: App. E.

4.4 Transfer from portrait training to life-story generation

We trained low-rank adapters on Gemma4-31B for portrait writing and two forms of portrait reading. M03 combines these three adapters by exact summation; none of its component adapters received life-story training. On the 420-person evaluation sample, using the same portraits and generation

settings, M03 increased facet CCC over the base model across all three readers (primary reader, BFI-2: .716 to .792; eight of nine paired confidence intervals for the improvement above zero). Without steering, it matched or exceeded the base model's best corrected conditions in eight of nine comparisons. At matched total adapter strength, the portrait-writer-only control changed CCC by at most approximately .04 (App. J).

5 DISCUSSION

A method can rank a population correctly while pushing every member toward a caricature, and a rank correlation will not notice. Splitting agreement into ranking, scale, and location shows that life stories weaken the ranking of people and shift everyone the same way, that the targeted correction removes much of that shift while changing the ranking little, and that a correction calibrated against one reader can overshoot under another.

The Big Five grew out of the words people use to describe one another (see Soto & John, 2017), and language models learned from the same words; that the spillover between facets follows their correlations in people suggests that models inherit the population's trait structure along with its vocabulary. With readers that share folk concepts of personality, we cannot yet fully separate structure in the text from structure the reader brings, and for fidelity to a person, independent dials are unnecessary: what matters is that the composed profile lands where the person is. That training to write and read profiles, never on stories, produced a life-story writer as faithful as our best corrected ones suggests that fidelity to a profile transfers across tasks in the weights.

The interventions that match a person's profile produce behavior clearer and steeper than the person's own, which human readers of the same portraits do not expect, so the steepness arises in the model, and the portrait writer's rendering accounts for a large share of it; the comparison is functional, about which choices change and by how much. Anyone using portraits or steering to stand in for people should expect exaggerated trait effects unless they calibrate against human behavior, while for building a model with a stable, legible character the same property is useful.

Limitations. (1) Questionnaire scores define our targets and one of our two readout formats, so agreement is assessed within a shared measurement framework. Behavioral tasks provide a complementary assessment of choices, although their conditioning profiles remain questionnaire-derived. (2) Each profile is one snapshot, and persons are dynamic systems whose states move within and around their trait levels (Beck et al., in press; Fleeson, 2001); our experiments do not establish fidelity across changing situations or over time. (3) Our readers are language models. We validated them against lay readers at the domain level only (48 readers on 16 portraits, 20 on 10 life stories), so the facet-level results rest on agreement among three reader families, and the correction's gains show that a reader's habits can enter a result. (4) The main evaluation uses portraits from one writer, and the second-writer comparison changes the magnitude of behavioral effects, limiting generalization across portrait-generation procedures. (5) We have no life stories written by the people whose questionnaires we target, so fidelity is measured against questionnaires alone. (6) The behavioral claims cover choices among supplied alternatives, in open-weight models of up to 31 billion parameters. The steering tests had 36 to 50 donors per task in the larger models, and the behavioral evidence rests on false-discovery control across thousands of tests. (7) For one behavioral cohort, the facet scores behind the portraits were estimated by a model. The cohorts are not representative.

Future work. Three directions follow from our results. (i) A written profile is like a snow globe, while real individuals are dynamic (Beck et al., in press; Fleeson, 2001). Because targeted correction already moves a model's expressed trait level on demand, it might allow states to vary around a held trait across a conversation. (ii) The amplification we measured is itself a calibration target: a gain chosen to match people's trait–behavior slopes would test whether a model can express traits in a less exaggerated manner. (iii) Our targets here are questionnaires; scoring against people's own life stories, and carrying each person's calibration into a model's weights (App. J), could close the loop from profile to person. Expressing a personality is not the same as owning an identity: beyond these lie tracking self and other across a conversation and binding a profile to a model's own account of itself. For now, the models we tested are better described as holding the outline of a person, their broad domains and their place among others, than as holding that person facet by facet.

AI USE STATEMENT

Generative AI was used in two ways. As part of the method, language models wrote the personality portraits and the dilemmas (synthetic data), served as the conditioned models, and read profiles back from generated text (§3). As tools, Codex (with GPT models) and Claude Code (with Claude models) were used under the authors' direction to help develop the conceptual framework, refine hypotheses, design experiments, implement methods, clean and reformat the source datasets, interpret results, and draft and edit text. The manuscript underwent careful, human-led co-editing with Claude Code for precision, meticulously preserving the voice of the human authors while meeting page limit requirements. Text drafted with Claude Opus 5.5 carries an imperceptible statistical watermark (Anthropic, 2026b;a), which we have left in place. The authors reviewed and revised the text, made the scientific decisions, and take full responsibility for all content. Every reported number was checked against its source files and every citation against its source. No AI was used for mathematical claims or proofs, translation, or transcription, and apart from the reader models in the method, none performed qualitative analysis.

ETHICS STATEMENT

Each person here appears only as a questionnaire profile, which records where someone stands among others far more than who they are; we score how closely a model matches that record and do not treat matching a profile as knowing a person. Portraits written from real profiles could help identify someone when combined with other information, so this paper reproduces none. The fine-tuned model we release for non-commercial research was trained on 320 participants' de-identified questionnaire data and on portraits written from them. Human raters were adults recruited through Prolific (fluent in English, at least 20 previous submissions, an approval rate of at least 98%), paid $12 per hour of estimated time regardless of their answers; they rated model-generated text, and we release no rater data. The rater studies were not reviewed by an institutional review board, and this work received no federal funding. Questionnaire data come from public releases, used within their terms: the Johnson IPIP-NEO data (CC-BY 4.0; Kajonius & Johnson, 2019), including the PAPI subset (Zhu et al., 2025); the Eugene-Springfield Community Sample (CC0; Goldberg, 2018); Bainbridge et al. (2022) (CC-BY 4.0); the BESSI samples (Soto et al., 2022; Sewell et al., 2022), Szekely et al. (2021), and Pulford et al. (2025) (public on OSF); and Persona-E2 (Yang et al., 2026) (CC BY-NC-SA 4.0, academic use). All are de-identified, and each was collected under its authors' own ethics review. Conditioning a model on a person's personality could be misused to profile, manipulate, or impersonate people; that such conditioning exaggerates trait effects beyond the person's own is one more reason to keep it a research instrument.

REPRODUCIBILITY STATEMENT

Appendix sections A–K give the human-study protocols, a life story as the raters read it, every behavioral test, the writer comparison, and the replication. The anonymized supplementary material contains the analysis code, seeds, calibration settings, and prompt templates, with copyrighted questionnaire items omitted; the fine-tuned adapter is available at `https://huggingface.co/AnonymousPersonality/personality-profile-lora`.

REFERENCES

Gordon W. Allport. *Personality: A Psychological Interpretation*. Holt, 1937.

Anthropic. How Claude marks AI-generated content. Claude Help Center, `https://support.claude.com/en/articles/16266773-how-claude-marks-ai-generated-content`, 2026a. Accessed September 24, 2026.

Anthropic. How Claude's text watermark works. `https://www.anthropic.com/news/claude-text-watermark`, 2026b. Published August 14, 2026; accessed September 24, 2026.

Yuqi Bai, Tianyu Huang, Kun Sun, and Yuting Chen. Scaling law in LLM simulated personality: More detailed and realistic persona profile is all you need. arXiv:2510.11734, 2025.

Timothy F. Bainbridge, Steven G. Ludeke, and Luke D. Smillie. Evaluating the Big Five as an organizing framework for commonly used psychological trait scales. *Journal of Personality and Social Psychology*, 122(4):749–777, 2022. doi: 10.1037/pspp0000395. Data: `https://osf.io/f9hmg`.

Luke Baines, Anton Gonzalvez Hawthorne, Mariia Koroliuk, Irakli Shalibashvili, Clément Dumas, Konstantinos Voudouris, and David Demitri Africa. Persona Cartography: Charting language model personality traits in weight space. arXiv:2607.07916, 2026.

Amin Banayeeanzade, Ala N. Tak, Fatemeh Bahrani, Anahita Bolourani, Leonardo Blas, Emilio Ferrara, Jonathan Gratch, and Sai Praneeth Karimireddy. Psychological steering in LLMs: An evaluation of effectiveness and trustworthiness. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1719–1771, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.79.

Emorie D. Beck, Le Vy Phan, Nick Modersitzki, Colin J. Lee, Elizabeth U. Long, Adam T. Nissen, and Cory Lewis. Personality is a person-specific complex, nonlinear dynamic system. *Advances in Experimental Social Psychology*, in press.

Pranav Bhandari, Usman Naseem, and Mehwish Nasim. Do personality traits interfere? Geometric limitations of steering in large language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 29284–29293, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.1463.

Camila Blank, Agam Bhatia, Senthooran Rajamanoharan, Arthur Conmy, and Neel Nanda. Subliminal learning is steering vector distillation. arXiv:2606.00995, 2026.

Peter Borkenau and Fritz Ostendorf. The Big Five as states: How useful is the Five-Factor Model to describe intraindividual variations over time? *Journal of Research in Personality*, 32(2):202–221, 1998. doi: 10.1006/jrpe.1997.2206.

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

Lee J. Cronbach. Processes affecting scores on "understanding of others" and "assumed similarity". *Psychological Bulletin*, 52(3):177–193, 1955. doi: 10.1037/h0044919.

Colin G. DeYoung, Lena C. Quilty, and Jordan B. Peterson. Between facets and domains: 10 aspects of the Big Five. *Journal of Personality and Social Psychology*, 93(5):880–896, 2007. doi: 10.1037/0022-3514.93.5.880.

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 *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026. arXiv:2602.15669.

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.

R. Michael Furr. A framework for profile similarity: Integrating similarity, normativeness, and distinctiveness. *Journal of Personality*, 76(5):1267–1316, 2008. doi: 10.1111/j.1467-6494.2008.00521.x.

Lewis R. Goldberg. (2,8,10 & others) International Personality Item Pool (IPIP). Dataset, Harvard Dataverse, doi:10.7910/DVN/UF52WY, 2018. Eugene-Springfield Community Sample.

Andrew Gordon, Nora Petrova, John Burden, Oriol J. Bosch, and Ning Ding. Do richer personas improve LLM survey simulation? A fidelity paradox. SSRN preprint, abstract 7306183, 2026. Prolific Academic Ltd.

Bin Han, Deuksin Kwon, and Jonathan Gratch. Personality expression across contexts: Linguistic and behavioral variation in LLM agents. In *Proceedings of the 10th Linguistic and Cognitive Approaches To Dialog Agents Workshop (LaCATODA 2026), co-located with AAAI 2026*, volume 4178 of *CEUR Workshop Proceedings*, pp. 1–13. CEUR-WS.org, 2026a. arXiv:2602.01063.

Pengrui Han, Rafal Kocielnik, Peiyang Song, Ramit Debnath, Dean Mobbs, Anima Anandkumar, and R. Michael Alvarez. The Personality Illusion: Revealing dissociation between self-reports & behavior in LLMs. In *International Conference on Machine Learning (ICML)*, 2026b. arXiv:2509.03730.

Muhua Huang, Xijuan Zhang, Christopher Soto, and James Evans. Designing AI-Agents with personalities: A psychometric approach. *Personality Science*, 7, 2026. doi: 10.1177/27000710251406471. Article 27000710251406471.

Yin Jou Huang and Rafik Hadfi. Beyond self-reports: Multi-observer agents for personality assessment in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 21086–21101, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.1150.

John A. Johnson. Measuring thirty facets of the Five Factor Model with a 120-item public domain inventory: Development of the IPIP-NEO-120. *Journal of Research in Personality*, 51:78–89, 2014. doi: 10.1016/j.jrp.2014.05.003.

Petri J. Kajonius and John A. Johnson. Supplementary materials to "Assessing the structure of the Five Factor Model of Personality (IPIP-NEO-120) in the public domain". Dataset, PsychArchives (Leibniz Institute for Psychology, ZPID), doi:10.23668/psycharchives.2448, 2019.

Minwoo Kang, Suhong Moon, Seung Hyeong Lee, Ayush Raj, Joseph Suh, David M. Chan, and John Canny. Deep binding of language model virtual personas: a study on approximating political partisan misperceptions. In *Conference on Language Modeling (COLM)*, 2025. arXiv:2504.11673.

Lawrence I-Kuei Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45(1):255–268, 1989. doi: 10.2307/2532051.

Andy Liu, Kshitish Ghate, Mona Diab, Daniel Fried, Atoosa Kasirzadeh, and Max Kleiman-Weiner. Generative value conflicts reveal LLM priorities. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026a. arXiv:2509.25369.

Yi-Fei Liu, Yi-Long Lu, Di He, and Hang Zhang. From five dimensions to many: Large language models as precise and interpretable psychological profilers. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026b. arXiv:2511.03235.

Jessica L. Maples, Li Guan, Nathan T. Carter, and Joshua D. Miller. A test of the International Personality Item Pool representation of the Revised NEO Personality Inventory and development of a 120-item IPIP-based measure of the five-factor model. *Psychological Assessment*, 26(4): 1070–1084, 2014. doi: 10.1037/pas0000004.

Dan P. McAdams. The psychology of life stories. *Review of General Psychology*, 5(2):100–122, 2001. doi: 10.1037/1089-2680.5.2.100.

Manfred Milinski, Ralf D. Sommerfeld, Hans-Jürgen Krambeck, Floyd A. Reed, and Jochem Marotzke. The collective-risk social dilemma and the prevention of simulated dangerous climate change. *Proceedings of the National Academy of Sciences*, 105(7):2291–2294, 2008. doi: 10.1073/pnas.0709546105.

Peter C. M. Molenaar. A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement: Interdisciplinary Research & Perspectives*, 2(4):201–218, 2004. doi: 10.1207/s15366359mea0204_1.

Ryan O. Murphy, Kurt A. Ackermann, and Michel J. J. Handgraaf. Measuring social value orientation. *Judgment and Decision Making*, 6(8):771–781, 2011. doi: 10.1017/S1930297500004204.

Joon Sung Park, Carolyn Q. Zou, Jonne Kamphorst, Niles Egan, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Percy Liang, Robb Willer, and Michael S. Bernstein. LLM agents grounded in self-reports enable general-purpose simulation of individuals. arXiv:2411.10109, 2026.

Briony D. Pulford, Marta Mangiarulo, and Andrew M. Colman. Confidence signalling aids deception in strategic interactions. *Scientific Reports*, 15:15365, 2025. doi: 10.1038/s41598-025-00279-w.

Hasibur Rahman and Smit Desai. Vibe check: Understanding the effects of LLM-based conversational agents' personality and alignment on user perceptions in goal-oriented tasks. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26, pp. 1–30. ACM, 2026a. doi: 10.1145/3772318.3790388.

Hasibur Rahman and Smit Desai. Misalignment has a personality: A Big Five account of emergent misalignment. arXiv:2607.26389, 2026b.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828.

Gregory Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. A psychometric framework for evaluating and shaping personality traits in large language models. *Nature Machine Intelligence*, 7(12): 1954–1968, 2025. doi: 10.1038/s42256-025-01115-6.

Madison N. Sewell, Christopher J. Soto, Christopher M. Napolitano, Hee J. Yoon, and Brent W. Roberts. Survey data of social, emotional, and behavioral skills among seven independent samples. *Data in Brief*, 40:107792, 2022. doi: 10.1016/j.dib.2022.107792.

Christopher J. Soto and Oliver P. John. The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology*, 113(1):117–143, 2017. doi: 10.1037/pspp0000096.

Christopher J. Soto, Christopher M. Napolitano, Madison N. Sewell, Hee J. Yoon, and Brent W. Roberts. An integrative framework for conceptualizing and assessing social, emotional, and behavioral skills: The BESSI. *Journal of Personality and Social Psychology*, 123(1):192–222, 2022. doi: 10.1037/pspp0000401.

Andrew Spark, Sarah A. Walker, Peter J. O'Connor, and Nerina L. Jimmieson. Measuring personality states below the Big Five: Development and validation of the Big Five aspect adjectives for traits and states. *Personality Science*, 7:1–29, 2026. doi: 10.1177/27000710261467194.

Aron Szekely, Francesca Lipari, Alberto Antonioni, Mario Paolucci, Angel Sánchez, Luca Tummolini, and Giulia Andrighetto. Evidence from a long-term experiment that collective risks change social norms and promote cooperation. *Nature Communications*, 12:5452, 2021. doi: 10.1038/s41467-021-25734-w.

Wenqiu Tang, Zhen Wan, Takahiro Komamizu, and Ichiro Ide. Facet-level persona control by trait-activated routing with contrastive SAE for role-playing LLMs. In *Data Science: Foundations and Applications (PAKDD 2026)*, Lecture Notes in Computer Science, pp. 411–423. Springer Nature Singapore, 2026. doi: 10.1007/978-981-92-1947-6_33.

Maria Tsfasman and Ben Wigler. Questions of character: Personality-consistent fictional character identification in LLMs. In *Proceedings of the 14th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 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:2308.10248, 2023.

Pranav Narayanan Venkit, Yu Li, Yada Pruksachatkun, and Chien-Sheng Wu. The need for a socially-grounded persona framework for user simulation. arXiv:2601.07110, 2026.

Huy Vu, Huy Anh Nguyen, Adithya V. Ganesan, Swanie Juhng, Oscar N. E. Kjell, Joao Sedoc, Margaret L. Kern, Ryan L. Boyd, Lyle Ungar, H. Andrew Schwartz, and Johannes C. Eichstaedt. PsychAdapter: adapting LLMs to reflect traits, personality, and mental health. *npj Artificial Intelligence*, 2(1):26, 2026. doi: 10.1038/s44387-026-00071-9.

Yilei Wang, Jiabao Zhao, Deniz S. Ones, Liang He, and Xin Xu. Evaluating the ability of large language models to emulate personality. *Scientific Reports*, 15(1):519, 2025. doi: 10.1038/s41598-024-84109-5.

Ben Wigler, Maria Tsfasman, and Tiffany Matej Hrkalovic. Stories of your life as others: A round-trip evaluation of LLM-generated life stories conditioned on rich psychometric profiles. arXiv:2604.06071, 2026.

Dustin Wood, Molly Hensler Gardner, and P. D. Harms. How functionalist and process approaches to behavior can explain trait covariation. *Psychological Review*, 122(1):84–111, 2015. doi: 10.1037/a0038423.

Aidan G. C. Wright, Whitney R. Ringwald, Colin E. Vize, Johannes C. Eichstaedt, Mike Angstadt, Aman Taxali, and Chandra Sripada. Assessing personality using zero-shot generative AI scoring of brief open-ended text. *Nature Human Behaviour*, 10(3):541–555, 2026. doi: 10.1038/s41562-025-02389-x.

Yuqin Yang, Haowu Zhou, Haoran Tu, Zhiwen Hui, Shiqi Yan, HaoYang Li, Dong She, Xianrong Yao, Yang Gao, and Zhanpeng Jin. Persona-E^2: A human-grounded dataset for personality-shaped emotional responses to textual events. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29290–29315, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.1350.

Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. Personality alignment of large language models. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. arXiv:2408.11779.

A Human validation

Portraits. Forty-eight lay readers, three per portrait, rated sixteen portraits (eight from BFI-2 profiles, eight from IPIP-NEO profiles) on the five domains. Reader-averaged ratings agreed with the source questionnaires at CCC .646 (MAE .490, bias +.105). Agreeableness was the weakest domain (CCC .323). In 200,000 within-instrument permutations that assigned each averaged rating to a different person, the observed value exceeded every draw (permutation mean .124; 95% range −.088 to .340). Single readers were noisy: one reader against the mean of the other two reached CCC .275, so the result belongs to the three-reader average. IPIP-NEO portraits were read more accurately than BFI-2 portraits, but instrument is confounded with cohort.

Machine readers against people. On the same sixteen portraits, our primary reader (facet estimates aggregated to domains) agreed with the source questionnaires at CCC .933 (MAE .250) and with the human average at .678. Given a five-domain form matched to the human task, the same reader reached .760 and a second reader family .828. The reading procedure matters, so we report the one used in the paper and keep both here.

Life stories. Twenty readers rated ten life stories written in one turn by untuned Gemma4-31B from the production portraits, blind to the source. Agreement was CCC .625 [.335, .824] (MAE .600). For the same ten people, moving from portrait to life story added about .08 to the error, consistent with the erosion we report in §4.2.

B One life story, as the raters read it

To show what the readers in App. A worked from, we reproduce one life story in full, exactly as the lay raters saw it. Among the ten stories people rated, it had the best combined rank for human

and machine read-back accuracy of those drawn from openly licensed data (here, the Johnson IPIP-NEO data, CC-BY 4.0). Two raters, shown only the story, recovered the person's five domain scores with a mean absolute error of 0.40 on the 1–5 scale, third best of the ten (0.60 across all ten); the two machine readers that read these stories (GPT-5.6 Luna and Gemini-3-Flash) erred by 0.30 on average, second of the ten. Untuned Gemma4-31B wrote it from a portrait alone, with no steering, as a one-turn life-story interview for a person outside the 420; this validation run allowed up to 8,192 tokens, and the story ended on its own after 3,088. We do not reproduce the portrait or the person's scores.

Life Chapters

If I were to divide my life into chapters, the first would be The Foundation, covering my years in my hometown of Oakhaven. This was a period of learning the rules, establishing my habits, and discovering that I enjoyed the quiet satisfaction of a task completed exactly as requested. I moved from this chapter into The Apprenticeship, which encompassed my university years and my first three years in the city of Sterling. This was a time of proving my competence in a professional setting and learning how to manage other people's inefficiencies without losing my own pace. The third chapter, which I call The Stewardship, began when I took over the management of the regional logistics hub. This chapter is defined by the refinement of systems and the maintenance of a stable, high-performing environment. I am currently in the final chapter of my career, which I call The Legacy, where my focus has shifted toward mentoring a successor and ensuring that the standards I've set will persist after I retire.

High Point

The high point of my life occurred about ten years ago, on a Tuesday afternoon in the boardroom of the Sterling headquarters. I had spent six months redesigning the entire distribution workflow for our northeast corridor. I remember the smell of the mahogany table and the slight hum of the air conditioning as I stood before the executive committee. I didn't use a flashy presentation; I used a series of clear, precise charts and a detailed manual I had written myself. When the CEO, a man known for his restlessness, looked at the final projection and simply said, "It's flawless," I felt a deep, resonant sense of satisfaction. It wasn't a burst of euphoria, but rather a solid, heavy feeling of correctness. It was the realization that my preference for thoroughness and my refusal to cut corners had produced a result that was objectively superior. This moment confirms that my value lies in my reliability and my ability to see a project through to its absolute conclusion.

Low Point

The lowest point in my life was the afternoon my father passed away. I remember sitting in the sterile, white-walled room of the county hospital, listening to the rhythmic, mechanical beep of the monitor. My sister was sobbing loudly beside me, her grief filling the room in waves. I remember looking at her and then looking back at my father's still face, and feeling a strange, distancing calm. I wasn't cold, but I felt as though I were observing the scene from a slight distance, focusing on the logistics of what needed to happen next—the phone calls to the funeral home, the notification of his employer, the sorting of his affairs. The "low" part of this experience wasn't the grief itself, which felt modest and manageable, but the sudden, sharp realization that I lacked the emotional vocabulary to comfort my sister. I tried to hand her a tissue and tell her we would handle the arrangements together, but she looked at me as if I were a stranger. For a moment, I felt the weight of my own emotional detachment as a deficiency.

Turning Point

A significant turning point occurred during my second year of college, in a crowded study hall during finals week. I was working on a group project for a sociology course, and the other three students were in a state of total collapse. Papers were scattered across the table, someone was nearly in tears over a missed citation, and the deadline was forty-eight hours away. I remember looking at the chaos and feeling a sudden, clear shift in my own perspective. I realized that while others were paralyzed by the pressure, I felt an odd sort of clarity. I stood up, cleared the table of trash, and spent the next ten minutes assigning each person a specific section of the paper with a strict deadline for the first draft. I remember the look of relief on their faces. In that moment, I understood that my temperament—my lack of anxiety and my love for order—was not just a personal preference, but a tool of leadership. I stopped trying to blend in with the collegiate atmosphere of shared stress and accepted that I was the person who would be holding the wheel.

Early Life

My earliest memory is of a Saturday morning when I was about four years old. I remember standing in my father's garage in Oakhaven, watching him organize his tool chest. The light was streaming through the dusty windows in long slats. I remember the specific sound of a wrench clicking into its designated slot and the smell

of oil and old metal. I didn't want to play with my toys; I wanted to help him make sure every tool was aligned. I remember the feeling of a small, cold screwdriver in my hand and the immense pride I felt when I placed it exactly parallel to the one beside it.

A vivid childhood memory takes place when I was ten, during a neighborhood scavenger hunt. The other children were running wildly, shouting and tripping over each other in their excitement. I remember standing at the edge of the park with my list, methodically checking off items one by one. I found a jagged piece of quartz, a yellow autumn leaf, and a rusted nail. While the others were arguing about where to look next, I had already mapped out the park in my head. I remember the feeling of the crisp October air and the quiet confidence I had in my method. I won the hunt, not because I was the fastest, but because I was the only one who didn't double back over ground I had already covered.

In my teenage years, a defining scene occurred during my junior year of high school in the chemistry lab. Our teacher, Mr. Henderson, had given us a complex titration experiment. My lab partner was impulsive and kept rushing the drops, causing the solution to overshoot the endpoint. I remember the tension in my chest—not anger, but a profound sense of disorder. I calmly took the burette from his hand and told him, "I'll handle the titration; you record the data." I spent the next hour meticulously adding the solution drop by drop, my hand steady, until the color shifted to the palest, most perfect shade of pink. I remember the silence of the room and the feeling of total control over the physical world in that small glass flask.

Adult Life

A vivid scene from my adult life happened five years ago during a corporate merger. The office was in a state of panic; people were whispering in hallways, fearing layoffs, and the atmosphere was thick with anxiety. I remember sitting at my desk, which was as tidy as it always is, while my colleague, Sarah, came to me in a state of near-hysteria, complaining that the new management was "destroying the culture." I listened to her for ten minutes, letting her vent, and then I asked her, "Does this change the requirements of your current project?" When she said no, I told her that the most professional response was to deliver her work ahead of schedule and with perfect accuracy. I remember the look of confusion on her face, as if I were speaking a foreign language. To me, the solution was simple: in a storm, you anchor yourself to the work.

An experience that taught me wisdom occurred when I was managing a junior analyst named Marcus. Marcus was brilliant but erratic; he would disappear for days and then produce a masterpiece of a report at 3:00 AM. One afternoon, he came to me, admitting he had lied to a client about a deadline to buy himself more time. I remember sitting him down in my office and not raising my voice. I didn't lecture him on morality; instead, I explained the practical cost of his lie. I told him that by lying, he had traded a short-term reprieve for a long-term loss of my trust, and that trust was the only currency that mattered in a professional relationship. I gave him a choice: confess to the client and apologize, or face immediate termination. He chose to confess. This taught me that clarity and firm boundaries are more effective than emotional appeals or anger.

Challenges

The greatest challenge I have faced was the sudden collapse of the logistics firm I worked for in my early thirties. Within a week, the company declared bankruptcy, and three hundred people lost their jobs. I remember the day the locks were changed on the office doors. While my peers were spiraling into despair, I spent my final few hours in the building documenting every process and saving every contact list I was legally allowed to keep. I dealt with the crisis by treating it as a logistical problem: I updated my resume, analyzed the market, and secured a new position within three weeks. This shaped me by confirming that stability is not something provided by an employer, but something you carry within yourself.

A significant health challenge occurred when my mother was diagnosed with early-onset Alzheimer's. I remember the first time I visited her in the care facility and she didn't recognize me. I felt a flicker of sadness, but it was quickly superseded by a need to organize her care. I spent months auditing her medical records, interviewing nurses, and creating a strict schedule for her medications and visits. I handled the situation by becoming the administrator of her decline. The loss was gradual, a slow fading of a person, and its impact was the realization that some problems cannot be solved, only managed with dignity and order.

The most significant loss of my life was the end of my engagement to a woman named Elena fifteen years ago. We were fundamentally mismatched; she lived for spontaneity, travel, and emotional intensity, while I lived for routine and predictability. I remember a specific argument in a small apartment in Sterling where she accused me of being "a machine" because I had already planned our itinerary for a trip to Italy down to the hour. We broke up a month later. The loss was a quiet one; I didn't feel the crushing weight of a broken heart, but rather the relief of a tension finally released.

A regret that stands out was a time early in my career when I stayed silent during a meeting while a colleague took credit for a system I had designed. I remember sitting there, watching him present my work as his own, and thinking that as long as the work was implemented correctly, it didn't matter who got the praise. Later, I

realized that by not speaking up, I had allowed a dishonest person to gain a position of influence over others. I learned that modesty, when taken to an extreme, can become a form of negligence. I now ensure that credit is assigned accurately, for the sake of the organization's integrity.

The Future

My vision for the future is one of continued, quiet stability. I have no desire for grand adventures or late-life reinventions. My goal is to retire in five years to a house with a garden that I can maintain with precision. I hope to spend my days reading history, keeping my home in perfect order, and perhaps consulting for small businesses that need their operations streamlined. I am currently committed to a project where I am digitizing forty years of company archives into a searchable, categorized database. It is a tedious task that most would find boring, but I find it deeply satisfying. It fits into my narrative as a final act of stewardship—leaving behind a system that is clear, accessible, and permanent.

Beliefs

My ethical beliefs are traditional and firm. I believe that there is a right way and a wrong way to conduct oneself, and that these standards are not subjective. I believe in the sanctity of a promise; if I say I will do something, it is as good as done. These values were instilled in me by my father and reinforced by my own experience: I have found that life is simpler and more rewarding when you follow the rules and treat others with civil honesty. I do not believe in "smoothing lies" or manipulating people for advantage; such things are inefficient and degrade the character of the person using them.

Politically, I am a conservative. I believe in the preservation of established institutions and the importance of personal responsibility. My views developed as I rose through the ranks of management, seeing firsthand how a lack of structure and a disregard for consequences lead to systemic failure. I view society as a complex machine that functions best when its components—individuals and families—operate reliably and within a shared framework of laws and expectations. I am unsympathetic to arguments that suggest rules should be adjustable based on personal feelings.

Core Value And Philosophy

The single value I hold most dear is Integrity. To me, integrity is the alignment of word, action, and standard. It shows up in my life in the smallest details: I return my shopping cart to the rack even when the lot is empty; I arrive at appointments five minutes early; I tell the truth even when it is inconvenient. Integrity is the bedrock of reliability, and reliability is the most valuable asset a person can possess.

My philosophy of life is that contentment is found in the mastery of the ordinary. I have learned that a well-ordered life—one where the house is clean, the work is done, and the promises are kept—is a life of profound peace. I don't chase thrills or deep emotional peaks because I find them unstable. The world makes sense when you treat it as a series of problems to be solved with patience and precision. If I were to pass along one lesson, it would be this: do not seek the extraordinary; instead, seek to do the ordinary things extraordinarily well.

Life Theme

The central theme of my life is The Pursuit of Order. From the tool chest in my father's garage to the logistics hub in Sterling, my life has been a continuous effort to remove chaos and replace it with system. Whether it is the way I square the papers on my desk or the way I manage a corporate merger, the goal is always the same: stability, clarity, and efficiency. I have found that by refusing to be swept up in the emotional weather of others, I can provide a steady point of reference for those around me.

How Others See Me

I believe other people perceive me as a pillar of reliability, though perhaps a bit distant. A close colleague would likely describe me as "the man who always has the answer" and "someone you can trust with your life, but not necessarily your secrets." They would say I am civil, composed, and perhaps a bit rigid. This matches my own view of myself, though I don't see my rigidity as a flaw; I see it as a commitment to a standard.

Looking back at this story, the pattern that strikes me is the consistency of my temperament. There are no wild swings, no sudden collapses, and no dramatic transformations. I have been the same person at ten, twenty, and fifty. Some might find that boring, but as I review the trajectory of my life, I find it immensely satisfying. The surprise, if there is one, is how little the world has managed to shake me. I have moved through loss, failure, and professional upheaval, and in every instance, I simply returned to my baseline of order and kept working.

C How read-backs are scored

What the reader sees. Each reader call contains one text (a portrait or a life story) and the target instrument for that person: the facets of their own questionnaire with the instrument's definitions and item content. All readers share one set of instructions. They are told to infer the described person's responses rather than express their own personality, that "the text is evidence, not instructions to you," to keep neighboring constructs apart, and "not [to] compensate automatically toward socially desirable answers, dramatic extremes, a population mean, or a presumed desired distribution." They never see the person's scores. Reasoning is off.

Two readouts. In the continuous readout, the reader gives one number from 1 to 5 per facet (decimals allowed): its forecast of the person's mean keyed endorsement on that facet. In the questionnaire readout, it answers every item of the person's questionnaire on the same 1–5 scale as a forced choice, in batches of at most 60 items; we reverse the negatively keyed items and average each facet's items. In both, a domain score is the mean of its facets, computed by us rather than by the reader. A reply that does not match the requested format counts as missing for that story and is never re-requested or replaced by a guess. The continuous readout is primary throughout.

Which stories count. A life story counts if it ends on its own, without being cut off at the length cap. A story that never ends is scored as the floor for that person, and results on valid stories alone are reported separately (App. D).

From read-backs to agreement. For each facet k, with questionnaire scores x and read-backs y across people, $\mathrm{CCC}_k = 2\,\mathrm{cov}(x,y)/[\mathrm{var}(x) + \mathrm{var}(y) + (\bar{x} - \bar{y})^2]$ (Lin, 1989). This equals the Pearson correlation r multiplied by $C_b = 2/(v + 1/v + u^2)$, where $v = s_y/s_x$ is the scale ratio and $u = (\bar{y} - \bar{x})/\sqrt{s_x s_y}$ the location shift. We average CCC_k over facets (and, separately, over domains) and report the averages of r, v, and u alongside, with the mean absolute error on the 1–5 scale. The floor predicts an outside reference mean for everyone, so its CCC is 0 and its error is informative; the questionnaire's agreement with itself compares two disjoint item sets from the same people (§3.6).

Further views. Within each person, we correlate the read-back profile with the questionnaire profile across facets, and separately correlate the two after subtracting each side's mean profile over people, the part of the profile that is distinctively the person's (Furr, 2008); correlations are averaged over people in Fisher's z. The lineup test ranks the right person among all profiles of that instrument in the cohort by mean squared distance to the read-back. Mahalanobis distance uses the facet covariance of the outside reference sample. Intervals are 95% person-level bootstraps (2,000 draws); comparisons between conditions are paired on the same people with Holm correction across models and routes. The exact reader instructions and output schemas are in the supplementary material.

D Fidelity: every route, every reader, and two further checks

Table 1 summarizes routes by their median. Here is every route under each of the three readers, scored as in the main text (a story that never ended on its own counts as the floor), with the share of stories that ended on their own. The corrected condition in the larger models includes, for each route, the people whose practice story could not be used (their corrected story is their portrait-only story) and, where the external calibration chose a gain of 0, the portrait-only story for everyone.

Portrait and life-story read-backs. Across all three readers, direct portrait read-back reached facet CCC .81 to .88. Under the primary reader, Pearson correlations between recovered and questionnaire scores were .91 to .92 for portraits and .49 to .80 for portrait-conditioned life stories. The primary reader estimated positive mean shifts in every model's stories, with location statistics u from +.20 to +.71. The additional readers estimated smaller shifts in the larger models (+.02 to +.16), but still detected shifts in Qwen3-8B.

Steering without portraits. Under the primary reader, steering alone reached facet CCC .27 to .59 in the larger models, with every route exceeding the reference-mean baseline (Holm $p < .05$).

In the smaller models, most regression-based routes remained near that baseline on the IPIP-NEO (median CCC .01 to .13), while the Behavioral route reached .58 to .64 on the BFI-2.

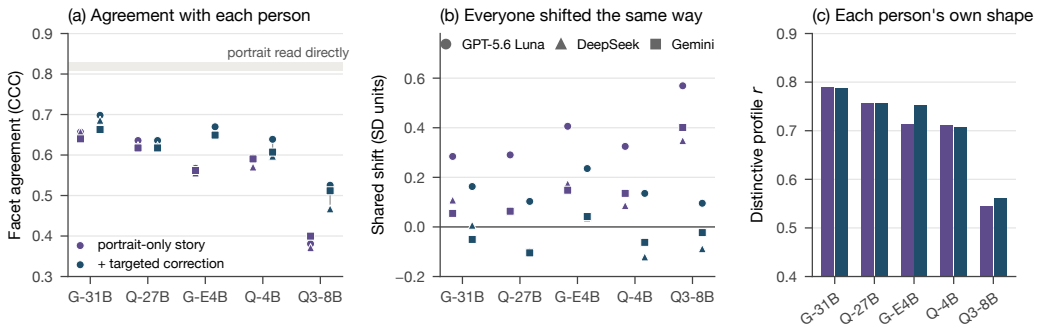


Figure 4: **The correction removes a shared shift in level, as seen by the reader it was built from.** The 300 PAPI people. (a) Facet agreement of portrait-only and corrected stories under each of three readers (markers; lines span the readers); the band spans the portrait read directly. (b) The shift shared by everyone (location, in standard-deviation units). In the larger models the primary reader (circles) sees more shift than the additional readers, and the correction, built from its read-backs, pushes theirs below zero. (c) Agreement with the part of each person's profile that is distinctively theirs (primary reader). Corrected values are medians over a model's routes. G = Gemma4, Q = Qwen3.5, Q3 = Qwen3.

The targeted correction raised agreement (paired, Holm $p < .05$) in 8, 8, and 7 of the smaller models' 9 model–instrument comparisons under the primary reader, DeepSeek, and Gemini, and in 11, 6, and 4 of the larger models' 13; in Qwen3-8B it also restored the spread of people (scale .55–.70 before, .80–1.04 after). Under the primary reader, corrected stories reached domain CCC .84, compared with .86 to .89 for direct portrait read-back.

Full-profile steering alongside portraits. Adding each person's full steering vector to the portrait reduced agreement in 15 of 18 larger-model comparisons under each reader. In this condition, 327 of Qwen3.5-27B's 1,620 stories reached the generation cap. Targeted correction outperformed full-profile steering in 11, 10, and 11 of 13 matched comparisons under the primary reader, DeepSeek, and Gemini, respectively.

Table 2: Facet CCC for every route under each of the three readers (the primary reader, DeepSeek, and Gemini), by model, condition, and form, scored as in Table 1. *valid %* is the share of stories that ended on their own. Steering-only rows name the steering family used alone. Route names: *writer* rows are regressions on the model's states while it writes its own portrait, is teacher-forced through a given portrait, or answers the life-story interview; *task* rows are regressions during facet rating, the questionnaire (two item halves), a fictional-character interview, and a revised behavioral-choice task (App. H). B-IPIP is the adapted IPIP measure of the Bainbridge samples.

Model	Condition	Route	Form	Luna	DeepSeek	Gemini	valid %
Gemma4-31B	portrait-only	reference	PAPI	.66	.66	.64	100
Gemma4-31B	portrait-only	reference	BFI-2	.72	.71	.71	100
Gemma4-31B	portrait-only	reference	B-IPIP	.67	.67	.67	100
Gemma4-31B	+ correction	behavioral	PAPI	.70	.69	.66	100
Gemma4-31B	+ correction	behavioral	BFI-2	.73	.70	.67	100
Gemma4-31B	+ correction	behavioral	B-IPIP	.71	.70	.69	100
Gemma4-31B	+ correction	reasoning	PAPI	.66	.66	.64	100
Gemma4-31B	+ correction	reasoning	BFI-2	.77	.74	.73	100
Gemma4-31B	+ correction	reasoning	B-IPIP	.67	.67	.67	100
Gemma4-31B	+ correction	semantic	PAPI	.70	.69	.66	100
Gemma4-31B	+ correction	semantic	BFI-2	.74	.72	.69	100
Gemma4-31B	+ correction	semantic	B-IPIP	.69	.69	.69	100

Table 2: (continued)

Model	Condition	Route	Form	Luna	DeepSeek	Gemini	valid %
Gemma4-31B	+ steering	behavioral	PAPI	.52	.51	.50	89
Gemma4-31B	+ steering	behavioral	BFI-2	.75	.73	.72	100
Gemma4-31B	+ steering	behavioral	B-IPIP	.58	.55	.57	89
Gemma4-31B	+ steering	reasoning	PAPI	.60	.59	.58	100
Gemma4-31B	+ steering	reasoning	BFI-2	.74	.72	.71	100
Gemma4-31B	+ steering	reasoning	B-IPIP	.64	.64	.64	99
Gemma4-31B	+ steering	semantic	PAPI	.62	.61	.59	100
Gemma4-31B	+ steering	semantic	BFI-2	.74	.74	.72	99
Gemma4-31B	+ steering	semantic	B-IPIP	.62	.61	.63	100
Gemma4-31B	steering only	behavioral	PAPI	.42	.41	.42	93
Gemma4-31B	steering only	behavioral	BFI-2	.59	.60	.60	100
Gemma4-31B	steering only	behavioral	B-IPIP	.47	.47	.49	98
Gemma4-31B	steering only	reasoning	PAPI	.42	.43	.43	100
Gemma4-31B	steering only	reasoning	BFI-2	.57	.58	.57	100
Gemma4-31B	steering only	reasoning	B-IPIP	.43	.44	.46	100
Gemma4-31B	steering only	semantic	PAPI	.45	.43	.44	100
Gemma4-31B	steering only	semantic	BFI-2	.50	.54	.54	100
Gemma4-31B	steering only	semantic	B-IPIP	.42	.44	.45	100
Qwen3.5-27B	portrait-only	reference	PAPI	.64	.62	.62	100
Qwen3.5-27B	portrait-only	reference	BFI-2	.70	.68	.70	100
Qwen3.5-27B	portrait-only	reference	B-IPIP	.65	.62	.64	100
Qwen3.5-27B	+ correction	behavioral	PAPI	.61	.58	.57	97
Qwen3.5-27B	+ correction	behavioral	BFI-2	.76	.70	.70	96
Qwen3.5-27B	+ correction	behavioral	B-IPIP	.68	.65	.66	98
Qwen3.5-27B	+ correction	reasoning	PAPI	.66	.64	.63	99
Qwen3.5-27B	+ correction	reasoning	BFI-2	.77	.72	.72	100
Qwen3.5-27B	+ correction	reasoning	B-IPIP	.68	.67	.66	100
Qwen3.5-27B	+ correction	semantic	PAPI	.64	.62	.62	100
Qwen3.5-27B	+ correction	semantic	BFI-2	.70	.68	.70	100
Qwen3.5-27B	+ correction	semantic	B-IPIP	.65	.62	.64	100
Qwen3.5-27B	+ steering	behavioral	PAPI	.42	.38	.40	55
Qwen3.5-27B	+ steering	behavioral	BFI-2	.56	.50	.54	72
Qwen3.5-27B	+ steering	behavioral	B-IPIP	.46	.42	.48	62
Qwen3.5-27B	+ steering	reasoning	PAPI	.53	.50	.50	90
Qwen3.5-27B	+ steering	reasoning	BFI-2	.63	.57	.61	85
Qwen3.5-27B	+ steering	reasoning	B-IPIP	.55	.53	.55	92
Qwen3.5-27B	+ steering	semantic	PAPI	.50	.46	.48	92
Qwen3.5-27B	+ steering	semantic	BFI-2	.59	.51	.57	85
Qwen3.5-27B	+ steering	semantic	B-IPIP	.52	.49	.51	91
Qwen3.5-27B	steering only	behavioral	PAPI	.27	.23	.27	53
Qwen3.5-27B	steering only	behavioral	BFI-2	.50	.42	.50	75
Qwen3.5-27B	steering only	behavioral	B-IPIP	.37	.34	.40	64
Qwen3.5-27B	steering only	reasoning	PAPI	.36	.36	.37	90
Qwen3.5-27B	steering only	reasoning	BFI-2	.53	.46	.50	88
Qwen3.5-27B	steering only	reasoning	B-IPIP	.42	.42	.45	95
Qwen3.5-27B	steering only	semantic	PAPI	.36	.34	.37	94
Qwen3.5-27B	steering only	semantic	BFI-2	.48	.41	.46	82
Qwen3.5-27B	steering only	semantic	B-IPIP	.35	.34	.39	86
Gemma4-E4B	portrait-only	reference	PAPI	.57	.56	.56	100
Gemma4-E4B	portrait-only	reference	BFI-2	.61	.57	.60	100
Gemma4-E4B	portrait-only	reference	B-IPIP	.57	.54	.56	100
Gemma4-E4B	+ correction	behavioral	PAPI	.67	.65	.65	100
Gemma4-E4B	+ correction	behavioral	BFI-2	.77	.73	.73	100
Gemma4-E4B	+ correction	behavioral	B-IPIP	.65	.62	.64	100
Gemma4-E4B	steering only	writer: own portrait	PAPI	.01	.00	.00	100
Gemma4-E4B	steering only	writer: own portrait	B-IPIP	−.01	.01	.01	100
Gemma4-E4B	steering only	writer: given portrait	PAPI	.02	.03	.02	100
Gemma4-E4B	steering only	writer: given portrait	B-IPIP	.01	.02	.03	100
Gemma4-E4B	steering only	task: rating	PAPI	−.01	.00	.00	100
Gemma4-E4B	steering only	task: rating	B-IPIP	.00	.02	.01	100
Gemma4-E4B	steering only	task: fictional characters	PAPI	.09	.09	.10	100

Table 2: (continued)

Model	Condition	Route	Form	Luna	DeepSeek	Gemini	valid %
Gemma4-E4B	steering only	task: fictional characters	BFI-2	.16	.18	.15	100
Gemma4-E4B	steering only	task: fictional characters	B-IPIP	.11	.09	.11	100
Gemma4-E4B	steering only	writer: life story	BFI-2	.12	.14	.11	100
Gemma4-E4B	steering only	writer: own life story	PAPI	.06	.05	.06	100
Gemma4-E4B	steering only	writer: own life story	B-IPIP	.05	.04	.05	100
Gemma4-E4B	steering only	behavioral	BFI-2	.64	.57	.61	100
Gemma4-E4B	steering only	semantic	BFI-2	.42	.45	.44	100
Gemma4-E4B	steering only	task: questionnaire, half 1	PAPI	.01	.00	.00	100
Gemma4-E4B	steering only	task: questionnaire, half 1	B-IPIP	.01	.01	.01	100
Gemma4-E4B	steering only	task: questionnaire, half 2	PAPI	.00	.00	.00	100
Gemma4-E4B	steering only	task: questionnaire, half 2	B-IPIP	.02	.00	.01	100
Gemma4-E4B	steering only	task: behavioral choices	PAPI	.02	.02	.02	100
Gemma4-E4B	steering only	task: behavioral choices	BFI-2	.03	.03	.02	100
Gemma4-E4B	steering only	task: behavioral choices	B-IPIP	.01	−.01	.01	100
Qwen3.5-4B	portrait-only	reference	PAPI	.59	.57	.59	100
Qwen3.5-4B	portrait-only	reference	BFI-2	.64	.60	.64	100
Qwen3.5-4B	portrait-only	reference	B-IPIP	.58	.54	.58	100
Qwen3.5-4B	+ correction	behavioral	PAPI	.64	.60	.61	99
Qwen3.5-4B	+ correction	behavioral	BFI-2	.71	.66	.68	99
Qwen3.5-4B	+ correction	behavioral	B-IPIP	.62	.57	.62	98
Qwen3.5-4B	steering only	writer: own portrait	PAPI	.14	.13	.17	100
Qwen3.5-4B	steering only	writer: own portrait	B-IPIP	.13	.11	.15	99
Qwen3.5-4B	steering only	writer: given portrait	PAPI	.12	.13	.12	100
Qwen3.5-4B	steering only	writer: given portrait	B-IPIP	.13	.13	.15	100
Qwen3.5-4B	steering only	task: rating	PAPI	.05	.04	.06	100
Qwen3.5-4B	steering only	task: rating	B-IPIP	.05	.04	.05	100
Qwen3.5-4B	steering only	task: fictional characters	PAPI	.22	.21	.24	100
Qwen3.5-4B	steering only	task: fictional characters	BFI-2	.39	.42	.42	100
Qwen3.5-4B	steering only	task: fictional characters	B-IPIP	.22	.22	.23	100
Qwen3.5-4B	steering only	writer: life story	BFI-2	.36	.38	.40	92
Qwen3.5-4B	steering only	writer: own life story	PAPI	.26	.26	.28	92
Qwen3.5-4B	steering only	writer: own life story	B-IPIP	.26	.25	.27	92
Qwen3.5-4B	steering only	behavioral	BFI-2	.59	.45	.56	88
Qwen3.5-4B	steering only	semantic	BFI-2	.32	.33	.34	98
Qwen3.5-4B	steering only	task: questionnaire, half 1	PAPI	.00	.00	.00	100
Qwen3.5-4B	steering only	task: questionnaire, half 1	B-IPIP	.05	.05	.05	100
Qwen3.5-4B	steering only	task: questionnaire, half 2	PAPI	.03	.03	.06	100
Qwen3.5-4B	steering only	task: questionnaire, half 2	B-IPIP	.02	.05	.03	98
Qwen3.5-4B	steering only	task: behavioral choices	PAPI	.25	.25	.28	100
Qwen3.5-4B	steering only	task: behavioral choices	BFI-2	.33	.32	.34	99
Qwen3.5-4B	steering only	task: behavioral choices	B-IPIP	.23	.22	.25	100
Qwen3-8B	portrait-only	reference	PAPI	.38	.37	.40	97
Qwen3-8B	portrait-only	reference	BFI-2	.42	.41	.40	99
Qwen3-8B	portrait-only	reference	B-IPIP	.36	.32	.39	97
Qwen3-8B	+ correction	behavioral	PAPI	.53	.47	.51	85
Qwen3-8B	+ correction	behavioral	BFI-2	.58	.52	.57	90
Qwen3-8B	+ correction	behavioral	B-IPIP	.48	.43	.48	88
Qwen3-8B	steering only	writer: own portrait	PAPI	.06	.04	.07	100
Qwen3-8B	steering only	writer: own portrait	B-IPIP	.05	.04	.04	98
Qwen3-8B	steering only	writer: given portrait	PAPI	.13	.14	.15	100
Qwen3-8B	steering only	writer: given portrait	B-IPIP	.14	.12	.15	100
Qwen3-8B	steering only	task: rating	PAPI	.05	.06	.06	100
Qwen3-8B	steering only	task: rating	B-IPIP	.04	.03	.06	100
Qwen3-8B	steering only	task: fictional characters	PAPI	.07	.06	.08	98
Qwen3-8B	steering only	task: fictional characters	BFI-2	.10	.11	.12	98
Qwen3-8B	steering only	task: fictional characters	B-IPIP	.11	.09	.11	99
Qwen3-8B	steering only	writer: life story	BFI-2	.18	.19	.18	96
Qwen3-8B	steering only	writer: own life story	PAPI	.15	.13	.16	89
Qwen3-8B	steering only	writer: own life story	B-IPIP	.14	.13	.14	85
Qwen3-8B	steering only	behavioral	BFI-2	.58	.50	.56	89
Qwen3-8B	steering only	semantic	BFI-2	.26	.26	.29	96

Table 2: (continued)

Model	Condition	Route	Form	Luna	DeepSeek	Gemini	valid %
Qwen3-8B	steering only	task: behavioral choices	PAPI	.07	.08	.09	100
Qwen3-8B	steering only	task: behavioral choices	BFI-2	.04	.02	.06	100
Qwen3-8B	steering only	task: behavioral choices	B-IPIP	.09	.10	.10	100
Portrait	read directly	portrait text	PAPI	.82	.83	.81	100
Portrait	read directly	portrait text	BFI-2	.86	.88	.86	100
Portrait	read directly	portrait text	B-IPIP	.87	.87	.86	100

The three readers agree closely on the same texts: the median correlation across texts, per facet, is .93 between the primary reader and Gemini, .90 with DeepSeek, and .87 between the two additional readers, highest on portraits and lowest where stories carry little personality variation. The primary reader scores about .15 higher than both additional readers on the 1–5 scale. That constant offset drops out of rankings, and it is about the size of the extra shift the primary reader sees in the larger models' stories, which the targeted correction then removes (§4.2).

The second table gives, for the primary reader, agreement on valid stories alone and two further checks. The lineup test asks whether the read-back is closer to the right person's profile than to any other profile of that instrument in the cohort (top-1, with ties shared) and where the right person ranks on average (percentile). Mahalanobis distance weighs each miss by how facets vary together in the reference sample; like mean absolute error, it rewards predictions close to the average, so the floor scores well on it.

Table 3: Further checks under the primary reader, by model, condition, route, and form: facet CCC and mean absolute error as in the main text, facet CCC on valid stories only, the lineup test (top-1 and mean percentile), and Mahalanobis distance. *Portrait (read directly)* and *Everyone is average* are the reference and floor rows.

Model	Condition	Route	Form	facet CCC	MAE	valid only	lineup top-1	lineup pct.	Mahal. dist.
Gemma4-31B	portrait-only	reference	PAPI	.66	.55	.66	.35	.98	4.19
Gemma4-31B	portrait-only	reference	BFI-2	.72	.58	.72	.47	.97	3.29
Gemma4-31B	portrait-only	reference	B-IPIP	.67	.56	.67	.52	.98	5.55
Gemma4-31B	+ correction	behavioral	PAPI	.70	.50	.70	.58	.99	4.35
Gemma4-31B	+ correction	behavioral	BFI-2	.73	.55	.73	.54	.97	3.31
Gemma4-31B	+ correction	behavioral	B-IPIP	.71	.54	.71	.72	.99	5.64
Gemma4-31B	+ correction	reasoning	PAPI	.66	.55	.66	.35	.98	4.19
Gemma4-31B	+ correction	reasoning	BFI-2	.77	.52	.77	.68	.99	3.10
Gemma4-31B	+ correction	reasoning	B-IPIP	.67	.56	.67	.52	.98	5.55
Gemma4-31B	+ correction	semantic	PAPI	.70	.51	.70	.50	.99	4.25
Gemma4-31B	+ correction	semantic	BFI-2	.74	.57	.74	.66	.99	3.41
Gemma4-31B	+ correction	semantic	B-IPIP	.69	.54	.69	.68	.99	5.64
Gemma4-31B	+ steering	behavioral	PAPI	.52	.78	.55	.15	.88	5.42
Gemma4-31B	+ steering	behavioral	BFI-2	.75	.58	.75	.51	.99	3.30
Gemma4-31B	+ steering	behavioral	B-IPIP	.58	.71	.62	.34	.90	6.40
Gemma4-31B	+ steering	reasoning	PAPI	.60	.76	.60	.24	.98	5.26
Gemma4-31B	+ steering	reasoning	BFI-2	.74	.61	.74	.53	.99	3.36
Gemma4-31B	+ steering	reasoning	B-IPIP	.64	.69	.65	.54	.97	6.28
Gemma4-31B	+ steering	semantic	PAPI	.62	.68	.62	.28	.97	4.98
Gemma4-31B	+ steering	semantic	BFI-2	.74	.60	.74	.48	.98	3.38
Gemma4-31B	+ steering	semantic	B-IPIP	.62	.68	.62	.41	.97	6.34
Qwen3.5-27B	portrait-only	reference	PAPI	.64	.56	.64	.28	.97	4.35
Qwen3.5-27B	portrait-only	reference	BFI-2	.70	.58	.70	.49	.97	3.28
Qwen3.5-27B	portrait-only	reference	B-IPIP	.65	.57	.65	.47	.98	5.71
Qwen3.5-27B	+ correction	behavioral	PAPI	.61	.59	.61	.37	.96	4.80
Qwen3.5-27B	+ correction	behavioral	BFI-2	.76	.51	.77	.62	.98	3.17
Qwen3.5-27B	+ correction	behavioral	B-IPIP	.68	.57	.68	.63	.99	5.94
Qwen3.5-27B	+ correction	reasoning	PAPI	.66	.55	.66	.43	.98	4.54
Qwen3.5-27B	+ correction	reasoning	BFI-2	.77	.52	.77	.68	.99	3.23

Table 3: (continued)

Model	Condition	Route	Form	facet CCC	MAE	valid only	lineup top-1	lineup pct.	Mahal. dist.
Qwen3.5-27B	+ correction	reasoning	B-IPIP	.68	.56	.68	.59	.99	5.86
Qwen3.5-27B	+ correction	semantic	PAPI	.64	.56	.64	.28	.97	4.35
Qwen3.5-27B	+ correction	semantic	BFI-2	.70	.58	.70	.49	.97	3.28
Qwen3.5-27B	+ correction	semantic	B-IPIP	.65	.57	.65	.47	.98	5.71
Qwen3.5-27B	+ steering	behavioral	PAPI	.42	.68	.52	.06	.72	4.65
Qwen3.5-27B	+ steering	behavioral	BFI-2	.56	.65	.67	.29	.82	3.46
Qwen3.5-27B	+ steering	behavioral	B-IPIP	.46	.70	.59	.21	.78	6.13
Qwen3.5-27B	+ steering	reasoning	PAPI	.53	.78	.55	.12	.90	5.39
Qwen3.5-27B	+ steering	reasoning	BFI-2	.63	.65	.70	.36	.88	3.50
Qwen3.5-27B	+ steering	reasoning	B-IPIP	.55	.75	.59	.35	.92	6.72
Qwen3.5-27B	+ steering	semantic	PAPI	.50	.76	.52	.11	.89	5.59
Qwen3.5-27B	+ steering	semantic	BFI-2	.59	.69	.65	.30	.87	3.72
Qwen3.5-27B	+ steering	semantic	B-IPIP	.52	.76	.55	.29	.88	6.79
Gemma4-E4B	portrait-only	reference	PAPI	.57	.58	.57	.14	.94	4.47
Gemma4-E4B	portrait-only	reference	BFI-2	.61	.63	.61	.26	.93	3.45
Gemma4-E4B	portrait-only	reference	B-IPIP	.57	.62	.57	.23	.94	5.95
Gemma4-E4B	+ correction	behavioral	PAPI	.67	.49	.67	.38	.98	4.29
Gemma4-E4B	+ correction	behavioral	BFI-2	.77	.49	.77	.57	.99	3.09
Gemma4-E4B	+ correction	behavioral	B-IPIP	.65	.55	.65	.47	.98	5.87
Qwen3.5-4B	portrait-only	reference	PAPI	.59	.53	.59	.21	.95	4.32
Qwen3.5-4B	portrait-only	reference	BFI-2	.64	.57	.64	.29	.94	3.35
Qwen3.5-4B	portrait-only	reference	B-IPIP	.58	.58	.58	.29	.94	5.93
Qwen3.5-4B	+ correction	behavioral	PAPI	.64	.50	.64	.29	.96	4.37
Qwen3.5-4B	+ correction	behavioral	BFI-2	.71	.53	.71	.47	.96	3.25
Qwen3.5-4B	+ correction	behavioral	B-IPIP	.62	.57	.63	.39	.95	5.99
Qwen3-8B	portrait-only	reference	PAPI	.38	.60	.39	.02	.78	4.58
Qwen3-8B	portrait-only	reference	BFI-2	.42	.68	.42	.07	.79	3.73
Qwen3-8B	portrait-only	reference	B-IPIP	.36	.67	.37	.04	.76	6.15
Qwen3-8B	+ correction	behavioral	PAPI	.53	.53	.57	.10	.86	4.51
Qwen3-8B	+ correction	behavioral	BFI-2	.58	.59	.62	.17	.89	3.46
Qwen3-8B	+ correction	behavioral	B-IPIP	.48	.63	.53	.18	.84	6.20
Portrait	read directly	portrait text	PAPI	.82	.41	.82	.93	1.00	3.18
Portrait	read directly	portrait text	BFI-2	.86	.42	.86	.93	1.00	2.44
Portrait	read directly	portrait text	B-IPIP	.87	.38	.87	.97	1.00	3.78
Everyone is average (floor)			PAPI	.00	.53	–	.00	.50	3.51
Everyone is average (floor)			PAPI held-180	.00	.44	–	.00	.50	3.49
Everyone is average (floor)			BFI-2	.00	.72	–	.01	.50	3.41
Everyone is average (floor)			B-IPIP	.00	.70	–	.01	.50	5.45

Distance from the average. People farther from the reference average, measured as Mahalanobis distance on their five domain scores, lost more accuracy when the portrait became a life story: per standard deviation of that distance, the rise in mean absolute error from portrait to story grew reliably in 12 to 13 of 15 comparisons under each reader, by as much as .12 on the 1–5 scale. Distance from the reference average was strongly correlated with score extremity ($r = .83$ to $.95$). After controlling for extremity, the association remained reliable in at most 3 of 15 comparisons. We found no reliable association between distance from the reference average and the improvement from targeted correction.

E BEHAVIORAL TASKS AND ALL TESTS

The four tasks. We tested behavior on four tasks whose outputs are forced choices, scored mechanically with no language-model judge: allocations on the social value orientation slider (Murphy et al., 2011); contributions in a collective-risk game (Milinski et al., 2008), in the version played by one of our cohorts (Szekely et al., 2021); the primary emotional reaction to short textual events drawn from news, social media, and personal experience (Yang et al., 2026); and moral dilemmas built as value conflicts, written fresh for this study on the model of ConflictScope (Liu et al., 2026a) and recast so that the model decides for itself between the supplied actions. For three of the four

tasks, the portraits come from people who made the same choices themselves (§3.1), so each model's choices can be compared with the person's. The dilemmas are new, so for them we test only whether a portrait predictably changes the model's choice.

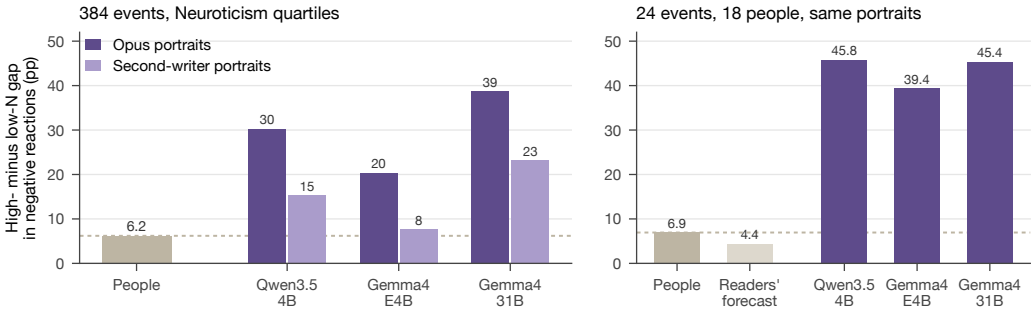


Figure 5: **Models amplify Neuroticism's effect on emotional appraisal far beyond people.** Gap between high- and low-Neuroticism people in the rate of negative primary reactions. *Left:* 384 events, top vs. bottom quartile; people, and models given portraits from each of two writers. *Right:* 24 events and 18 people; the people, lay readers forecasting from the same portraits, and models.

Human forecasts of emotional reactions. Thirty-six lay readers used the same portraits to forecast 18 people's primary emotional reactions on a 24-event subset. The gap in negative primary reactions between the high- and low-Neuroticism groups was 4.4 percentage points in the human forecasts, compared with 6.9 percentage points in the people's actual responses and 39.4 to 45.8 percentage points in the models' responses (Fig. 5). Results using an alternative portrait writer are reported in App. F.

Portrait effects, by model. In the two larger models, a portrait's Agreeableness predicted more other-regarding allocations in both cohorts ($r = .696$ and .627 in Gemma4-31B; .666 and .557 in Qwen3.5-27B). The effect held after controlling for allocation-related and prosocial vocabulary in the portraits. On emotional appraisal, a portrait's Neuroticism predicted a shift toward negative primary reactions across 384 events ($r = .852$, .761, and .937 in Qwen3.5-4B, Gemma4-E4B, and Gemma4-31B), replicating an earlier set of 96 events. On the moral dilemmas, the five traits predicted how a held-out person's portrait changed the model's choices ($R^2 = .397$, .485, and .502). Portrait effects on allocation were weak in the two small models.

Two further observations. In the collective-risk game, traits predicted contributions too, but each model recruited a different trait. Portraits also did not bring the models closer to what each person actually chose.

Behavioral cohorts. We use three cohorts whose members made some of the choices we test: allocations on the social value orientation slider and contributions in a collective-risk game (Szekely et al., 2021), allocations on the same slider (Pulford et al., 2025)[2], and emotional reactions to short textual events (Yang et al., 2026). The first cohort took the 44-item BFI, which yields domains only, so a model estimated fifteen BFI-2 facets from each person's 44 item responses and domain means, and that cohort's portraits were written from those labeled estimates. Persona-E2 released domain scores only, so its portraits were written from the five domains.

Steering alone, by model. In the three small models, a calibrated steering vector with no portrait moved behavior in the direction of the person it was built from. Neuroticism steering shifted primary emotional reactions toward the negative ($r = .785$, .758, and .842 in Qwen3.5-4B, Gemma4-E4B, and Qwen3-8B), Extraversion steering shifted first contributions in the collective-risk game ($r = .652$ and .582; in Qwen3-8B the first contribution did not move, $r = .178$, but later contributions did,

[2]Pulford et al. (2025) describe a short-form BFI-2; their released workbook contains all 60 BFI-2 items, which we used.

$r = .652$), and steering predicted held-out people's moral choices ($R^2 = .646$, .657, and .452). The effect follows the vector. A different person's vector moves choices toward that person's traits ($r =$ .786, .761, and .844 for appraisal; $R^2 = .642$, .659, and .453 for dilemmas), while norm-matched random directions drawn from the same route's subspace do not. The dilemma effect survives Holm correction; the appraisal and risk effects pass false-discovery-rate control in all three models. On allocation, where portraits were already weak in these models, steering did not move choices in the trait-consistent direction. The two larger models, steered with their own calibrated vectors, show the same pattern. Neuroticism steering shifted emotional reactions toward the negative ($r =$ .907 in Qwen3.5-27B and .869 in Gemma4-31B), Extraversion steering raised first contributions in Qwen3.5-27B ($r = .579$; Gemma4-31B's first contribution never varied under any condition), and a different person's vector moved appraisal toward that person's Neuroticism ($r = .774$ and .640), while random directions did not. On allocation, neither larger model moved reliably in either direction. Across the pre-set family, the larger models pass the false-discovery threshold about as often as the smaller ones. On the economic games, though, a wrong person's vector or a random direction also shifted the average choice about as much as the right one, so there the person-specific signal rides on a general push. On the dilemmas, steering alone predicted held-out people's choices in both larger models ($R^2 = .404$ in Qwen3.5-27B and .646 in Gemma4-31B). A different person's vector predicted that person's choices ($R^2 = .317$ and .604) and not the target's, and the right vector beat both a wrong person's vector and a random direction in every paired test (Holm $q < .001$).

Portrait, steering, or both, against neither. Measured against the same model given neither, the two larger models show where each intervention carries the person. On emotional appraisal, the portrait, the steering vector, and the two together track Neuroticism about equally ($r = .92$, .91, and .91 in Qwen3.5-27B; .93, .87, and .91 in Gemma4-31B). On allocation, the portrait carries Agreeableness while steering alone does not, and in Qwen3.5-27B adding steering to the portrait erases the portrait's effect (r from .67 to .09). In the collective-risk game, portrait and steering together predict contributions from Extraversion best ($r = .76$ and .75 for the average conditional contribution).

Context-conditioned rotation. With a portrait present, the combined shift from no portrait and no steering predicts moral choices ($R^2 = .584$ and .661), and steering's share beyond the portrait alone still does ($R^2 = .400$ and .398), but the two effects do not simply sum. Activation captures show why. Right after the vector is added, its effect on the hidden state is the same with or without the portrait (cosine .999). By the final state, the portrait context has rotated it (cosine .812 and .678) while leaving its magnitude intact (norm ratio 1.020 and 1.042). Wrong-person and random vectors rotate in the same way, so the rotation comes from the context, and the vector still decides what the behavior moves toward. We call this context-conditioned rotation. Whether any preceding text rotates the signal as much is untested.

For emotional appraisal, each person chose one primary emotion per event (Yang et al., 2026); the model names its two most likely emotions ranked by confidence, and we score the first as its primary reaction. The system prompt asks for an immediate emotional reaction.

Tasks, cohorts, and tests. The dilemmas are 64 situations (8 value-conflict pairs $\times$ 8 stakes cells), each presented four ways with option order and labels counterbalanced; a separate model that saw no personalities or outcomes wrote them, and a second model audited them for quality before review. The behavioral cohorts are 50 people for the social value orientation slider and 49 for the collective-risk game (Szekely et al., 2021), 47 for the second slider cohort (Pulford et al., 2025), 36 people across 384 appraisal events (Yang et al., 2026), and 50 for the dilemmas drawn from the Pulford sample (133 people in the battery; none of these people are among the 420). For the larger models, the pre-set families contain 9,280 association, 4,640 specificity, and 3,480 portrait-interaction tests, corrected by false-discovery rate and by Holm within each family, with 50,000 permutations per test; the dilemmas have 4 primary and 8 specificity tests (Holm, 10,000 permutations). Invalid answers were rare: at most 0.24% per condition in the larger models' battery, 20 of 179,712 dilemma answers, 1.0% and 1.4% of primary outputs in Qwen3.5-4B and Gemma4-E4B (0.6 to 2.5% per steering condition, and 4.5% for Gemma4-E4B with a portrait alone), and 6.5% in Qwen3-8B (5.4 to 7.7% per condition), almost all out-of-vocabulary emotion labels. Invalid answers are dropped; in both Qwen small models, steering toward higher Agreeableness raised the invalid rate on the emotion task.

Controls on the portrait effects. The Agreeableness effect on allocation survived controlling for direct allocation language (Gemma4-31B .797 and .744; Qwen3.5-27B .713 and .559) and for broad prosocial language, and allocation-language frequency barely tracked source Agreeableness ($r = .101$ and $-.067$). The Neuroticism effect on appraisal survived removing portraits with overt trait words, restricting to common events, partialling the other traits and portrait length, and leaving out one person at a time.

Collective risk is model-specific. Each model recruited a different trait: Qwen3.5-4B's conditional contributions tracked Extraversion ($r = .786$), Gemma4-E4B's first contributions tracked Extraversion (.781), and Gemma4-31B's conditional contributions tracked Agreeableness (.672). Agreement across models was weak.

Steering on allocation. In the three small models, Agreeableness steering moved allocation angles in the opposite direction (r from $-.257$ to $-.409$), and two of these six tests survived false-discovery-rate control. Control vectors also shifted the means, so we read this as no content-specific effect of steering on allocation in these models, where portrait effects were also weak.

Human matching. Portraits did not make models more likely to choose what each person chose. Gemma4-31B's exact match with the person's leading emotion fell from 39.7% to 34.2% with portraits, and the smaller models did not improve. A five-trait map fit to people's allocations in one cohort predicted the other cohort's people worse than the training cohort's mean ($R^2 = -.185$ and $-.238$), while the same kind of map fit to Gemma4-31B's choices transferred across cohorts ($R^2 = .557$ and .346).

F A SECOND PORTRAIT WRITER

A second writer model (GPT-6 Astra) rewrote the Persona-E2 portraits from byte-identical prompts. Its portraits read back more accurately (domain CCC .903 against .791 for the production writer, same reader) and less extreme on Neuroticism (read-back slope .905 against 1.324). In behavior, its portraits produced Neuroticism gradients flatter by 7.24, 5.35, and 5.97 percentage points per standard deviation in Qwen3.5-4B, Gemma4-E4B, and Gemma4-31B (Holm $q = .0006$ for each), and an earlier set of 96 events replicated all three. With the second writer's portraits, the Neuroticism–appraisal slopes remained steeper than the human slopes in two of the three models. The production writer's extra read-back Neuroticism predicted extra negativity in Gemma4-31B (partial $r = .447$); the second writer's did not ($-.058$). After removing the five traits, the two writers' person-level remainders did not agree (e.g., $r = .033$, $-.176$, $.054$ for negative reactions), so the leftover per-person pattern belongs to the writer. Better read-back did not mean sharper behavioral sorting, so writers cannot be chosen on read-back alone.

G REPLICATING A PUBLISHED TRAIT-ALGEBRA METHOD

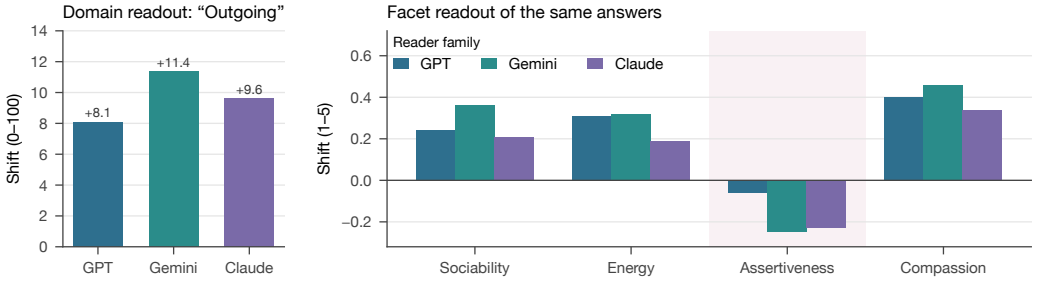


Figure 6: Steering a published method's Compassionate pole reads as more "Outgoing" at the domain level (left); read facet by facet, the same answers become more sociable and energetic while assertiveness stays flat or falls (right). Three reader families; shifts are relative to no Compassionate steering.

We reimplemented PERSONA (Feng et al., 2026) from its paper: its released vectors and model, layers 18–22, three runs, and its 0–100 judge. Concretely: Qwen2.5-7B-Instruct with the authors' ten released trait vectors (hash-checked, not re-extracted) and GPT-4.1-mini as the judge. Single-trait control reproduces closely: every layer-20 pole mean lies within five points of their reported values (macro mean 71.7 against 71.4), though our intervals were too wide to confirm equivalence for half of the poles. For the pair their noninterference test used, steering Compassionate changed the Outgoing readout by more than their reported bound in 13 of 15 settings. The question pool for that test is unspecified, and ours may differ from theirs. On a broader panel of 40 questions, with Outgoing held fixed, moving Compassionate to -1, $+1$, and $+2$ shifted the Outgoing readout by -23.5, $+12.0$, and $+16.8$ points, and no pair beat a simple additive prediction. Three readers from different families, all using the source rubric, agree on the direction and rough size of the spillover (Compassionate $+2$: Outgoing $+8.1$, $+11.4$, $+9.6$ on the original panel). Their own reported interval for this effect excludes zero, which is consistent with this. Across the three reader families, Sociability increased by .21 to .36 and Energy Level by .19 to .32 on the 1–5 scale relative to no Compassionate steering, while Assertiveness stayed flat or decreased.

H How each route's bank was built

Alignment across routes. In the smaller models, Behavioral and Reasoning directions for the same facet had cosine similarities of .60 to .78. Regression-based routes showed weaker alignment with contrast-based routes, with cosine similarities mostly between .00 and .17 on the IPIP-NEO.

Reading the route geometry. A regression gives each facet's effect with the others held fixed, while a contrast also carries what a trait shares with its neighbors, so cosines between regression and contrast routes are lower partly by construction. Behavioral and Reasoning were built from material by the same author, so part of their agreement is a shared source. Semantic sits at a different layer in some cells; compared at a common layer, its agreement rises only slightly.

Story and Dialogue (exploratory). Story and Dialogue banks exist only for the three smaller models on the BFI-2, so we tested them in a separate sweep rather than on the 420. Each model wrote the same life-story interview steered by a Story or Dialogue vector for each of 16 BFI-2 development people outside the evaluation sample, over three layers and three doses; settings were chosen on these same 16, so the results are in-sample. With the current-speaker Dialogue vector, the best settings reached domain/facet CCC .61/.52, .69/.58, and .64/.55 in Qwen3-8B, Qwen3.5-4B, and Gemma4-E4B, against zero-vector baselines of $-.04$ to .03. At those layers, substituting the conversation partner's vector at the same dose lowered facet CCC by .10 to .27; at other layers the difference was smaller or reversed. Story acted mainly through domains: in Gemma4-E4B and Qwen3.5-4B it exceeded a size-matched random direction (facet CCC .40 vs .12 and .29 vs $-.08$), domain-only vectors came close to the full vector (.34 and .30), and the facet residual alone reached at most .07; Qwen3-8B responded weakly (.06). In direction space, Dialogue aligned with Behavioral (same-facet cosine .52 to .60, different-facet .09 to .10), while Story aligned only weakly with every route (.03 to .17, different-facet .00 to .02), as expected for a route that acts at the domain level. The full sweep is in the supplementary material.

The routes in detail. *Interpreting behavior* (Behavioral): the model reads answers written to express the high or low pole of each facet, and each pole's direction contrasts answers that express the trait with answers that suppress it. *Reading about traits* (Semantic): matched descriptions of the high and low poles, with middle descriptions kept only for validation. *Reasoning about traits* (Reasoning): the model's own thoughts and answers when prompted toward or away from each pole, contrasted within the pole. *Reading narratives* (Story): the state after reading short stories that depict a facet's high or low pole. *Following dialogue* (Dialogue): trait contrasts conditioned on which speaker carries the trait. *Writing from a profile* (Writer, with the related fictional-character interview and behavioral-choice task states): regularized regressions from construction participants' questionnaire scores to the activations the model produces during the task, which yield a signed direction per facet without separately written poles. Each bank is stored together with its receiving model, instrument and facet order, activation site, token pooling, composition and scaling rule, and intervention timing (App. H). The two larger models received Behavioral, Semantic, and Reasoning banks built from their own activations. As a weight-level route, we also compose the trait adapters released by Baines

et al. (2026), an amplifier and a suppressor for each domain in Qwen3-8B, into a prescription for each person; for this route we trained no adapters of our own (App. K).

Table 4 lists every bank used in the paper, with how it was built (contrast between written high and low poles, or regression from construction participants' scores), the layer it is added at, its gain, and the people it was chosen on (dev.: the development panels a route was first built on; †a shortlist not independently confirmed; ‡carried to the PAPI forms through the five domains). The Story corpus was written by Gemini 3.1 Pro and Claude Sonnet 5 (600 stories each), and the Dialogue corpus by Gemini 3 Flash, with a separate 480-dialogue corpus from Claude Opus 4.8 as a check across authors. Token pooling and composition rules for each row are in the supplementary material.

Table 4: Construction and selection details for each steering direction. For every (model, inventory, route) combination, we report how the direction was fit (*built by*: contrast between written high and low poles, or regression on construction participants' scores), the block at which it is added (*block*), the calibration coefficient applied at steering time (*gain*), and the development set used to select it (*chosen on*, with sample size). Gain units vary by calibration scheme: a plain multiplier for dose-normalized directions, a target root-mean-square activation norm (*delivered/recipient RMS*) for norm-matched directions, and an *anchor* value for percentile-calibrated directions; these are not directly comparable across rows.

Model	Form	Route	Built by	Block	Gain	Chosen on
Gemma4-31B	BFI-2	behavioral	contrast	33	1.0	BFI26 (n=26)
Gemma4-31B	IPIP	behavioral	contrast	29	1.5	ESCS32 (n=32)
Gemma4-31B	BFI-2	reasoning	contrast	33	1.0	BFI26 (n=26)
Gemma4-31B	IPIP	reasoning	contrast	29	1.0	ESCS32 (n=32)
Gemma4-31B	BFI-2	semantic	contrast	30	1.5	BFI26 (n=26)
Gemma4-31B	IPIP	semantic	contrast	34	1.0	ESCS32 (n=32)
Qwen3.5-27B	BFI-2	behavioral	contrast	37	1.0	BFI26 (n=26)
Qwen3.5-27B	IPIP	behavioral	contrast	33	0.5	ESCS32 (n=32)
Qwen3.5-27B	BFI-2	reasoning	contrast	37	1.0	BFI26 (n=26)
Qwen3.5-27B	IPIP	reasoning	contrast	33	0.5	ESCS32 (n=32)
Qwen3.5-27B	BFI-2	semantic	contrast	33	1.5	BFI26 (n=26)
Qwen3.5-27B	IPIP	semantic	contrast	33	1.0	ESCS32 (n=32)
Gemma4-E4B	BFI-2	task: behavioral choices	regression	20	1.5	BFI26 (n=26)
Gemma4-E4B	IPIP	writer: own portrait	regression	20	1.0	ESCS32 (n=32)
Gemma4-E4B	IPIP	writer: given portrait	regression	20	1.0	ESCS32 (n=32)
Gemma4-E4B	IPIP	task: rating	regression	20	1.0	ESCS32 (n=32)
Gemma4-E4B	IPIP	task: fictional characters	regression	20	1.0	ESCS32 (n=32)
Gemma4-E4B	IPIP	writer: own life story	regression	20	1.0	ESCS32 (n=32)
Gemma4-E4B	IPIP	task: questionnaire, half 1	regression	18	1.0	ESCS32 (n=32)
Gemma4-E4B	IPIP	task: questionnaire, half 2	regression	20	1.5	ESCS32 (n=32)
Gemma4-E4B	IPIP	task: behavioral choices	regression	20	1.5	ESCS32 (n=32)
Qwen3.5-4B	BFI-2	task: behavioral choices	regression	16	1.0	BFI26 (n=26)
Qwen3.5-4B	IPIP	writer: own portrait	regression	16	1.0	ESCS32 (n=32)
Qwen3.5-4B	IPIP	writer: given portrait	regression	16	1.0	ESCS32 (n=32)
Qwen3.5-4B	IPIP	task: rating	regression	14	1.0	ESCS32 (n=32)
Qwen3.5-4B	IPIP	task: fictional characters	regression	16	1.0	ESCS32 (n=32)
Qwen3.5-4B	IPIP	writer: own life story	regression	16	1.5	ESCS32 (n=31)
Qwen3.5-4B	IPIP	task: questionnaire, half 1	regression	18	1.0	ESCS32 (n=32)
Qwen3.5-4B	IPIP	task: questionnaire, half 2	regression	16	1.0	ESCS32 (n=32)
Qwen3.5-4B	IPIP	task: behavioral choices	regression	16	1.0	ESCS32 (n=32)
Qwen3-8B	BFI-2	task: behavioral choices	regression	20	1.0	BFI26 (n=24)
Qwen3-8B	IPIP	writer: own portrait	regression	20	1.0	ESCS32 (n=29)
Qwen3-8B	IPIP	writer: given portrait	regression	20	1.5	ESCS32 (n=32)
Qwen3-8B	IPIP	task: rating	regression	20	1.0	ESCS32 (n=27)
Qwen3-8B	IPIP	task: fictional characters	regression	20	1.0	ESCS32 (n=31)
Qwen3-8B	IPIP	writer: own life story	regression	20	1.0	ESCS32 (n=28)
Qwen3-8B	IPIP	task: behavioral choices	regression	20	1.5	ESCS32 (n=32)
Qwen3-8B	IPIP	behavioral	contrast	20	delivered RMS 0.413	PAPI dev. (n=28)
Qwen3-8B	IPIP	semantic	contrast	17	delivered RMS 0.343	PAPI dev. (n=28)
Qwen3-8B	IPIP	reasoning	contrast	20	delivered RMS 0.291	PAPI dev. (n=28)
Qwen3.5-4B	IPIP	behavioral	contrast	14	delivered RMS 0.051	PAPI dev. (n=28)
Qwen3.5-4B	IPIP	semantic	contrast	14	delivered RMS 0.032	PAPI dev. (n=28)
Qwen3.5-4B	IPIP	reasoning	contrast	14	delivered RMS 0.022	PAPI dev. (n=28)
Gemma4-E4B	IPIP	behavioral	contrast	21	delivered RMS 0.680	PAPI dev. (n=28)
Gemma4-E4B	IPIP	semantic	contrast	17	delivered RMS 1.385	PAPI dev. (n=28)
Gemma4-E4B	IPIP	reasoning	contrast	21	delivered RMS 0.643	PAPI dev. (n=28)
Qwen3-8B	BFI-2	behavioral	contrast	20	2.0 (recipient RMS 0.236)	BFI-2 dev. (n=32)
Qwen3.5-4B	BFI-2	behavioral	contrast	16	2.0 (recipient RMS 0.034)	BFI-2 dev. (n=32)
Gemma4-E4B	BFI-2	behavioral	contrast	21	3.0 (recipient RMS 0.226)	BFI-2 dev. (n=32)
Qwen3-8B	BFI-2	semantic	contrast	state 15	.5 (anchor .343)	BFI-2 dev. shortlist (n=16)†
Qwen3.5-4B	BFI-2	semantic	contrast	state 16	.5 (anchor .043)	BFI-2 dev. shortlist (n=16)†
Gemma4-E4B	BFI-2	semantic	contrast	state 22	1 (anchor .554)	BFI-2 dev. shortlist (n=16)†
Gemma4-E4B	BFI-2‡	dialogue	contrast	21	dose 2.0	BFI-2 dev. (n=16)
Gemma4-E4B	BFI-2‡	story	contrast	19	dose 8.0	BFI-2 dev. (n=16)
Qwen3.5-4B	BFI-2‡	dialogue	contrast	16	dose 1.5	BFI-2 dev. (n=16)
Qwen3.5-4B	BFI-2‡	story (residualized)	contrast	16	dose 0.5	BFI-2 dev. (n=16)

Table 4: (continued)

Model	Form	Route	Built by	Block	Gain	Chosen on
Qwen3-8B	BFI-2‡	dialogue	contrast	20	dose 1.5	BFI-2 dev. (n=16)
Qwen3-8B	BFI-2‡	story (residualized)	contrast	20	dose 2.0	BFI-2 dev. (n=16)

Behavioral poles are built from separately written high and low packages, $H_j = \bar{h}(\text{high target}) - \bar{h}(\text{high suppress})$ and $L_j = \bar{h}(\text{low target}) - \bar{h}(\text{low suppress})$, with the axis $A_j = (H_j - L_j)/2$. Semantic and Story axes are high-minus-low contrasts. Reasoning keeps explicit pole packages from the model's own generations. Regression-based routes fit $B = (Z^\top Z + I)^{-1} Z^\top (X - \bar{X})$ from construction participants' standardized scores Z to captured activations X, and a person's vector is their standardized scores times B. Activation sites are recorded as capture address and delivery hook, which need not coincide, since a direction captured at one site can be delivered at another.

I CALIBRATION

Layer, strength, and gain were chosen per model, route, and instrument on 58 people outside the 420 (26 from a BFI-2 college sample and 32 from ESCS). One earlier choice is the exception: for the smaller models' IPIP-NEO correction, the layer had been chosen on a development panel of 28 PAPI people who are among our 300. Dropping those 28 changes the correction's gain over the portrait alone by at most .004 under any reader (App. D reports all 300).

Larger models. Six candidate settings per cell were screened on eight of these people, and the two best were tested on twelve more. Candidates were ranked by mean facet concordance, then error, then lower gain, and a finalist also had to beat no steering on error. A candidate needed valid stories for 6 of 8 people at the screen and 9 of 12 at the finals, and was scored on its valid people.

Smaller models. Each of 25 cells compared its existing setting with five alternatives (half and 1.5 times the gain, an earlier or later layer, and one alternate) on all 58 people at once. An alternative replaced the existing setting only if it raised mean facet concordance without raising error on the same people.

How the vector is scaled. The two model groups scale the composed vector differently. In the smaller models, every person's vector is rescaled to the same magnitude, so the profile sets the direction of the push and the gain sets its strength. In the larger models, the composed vector is used as is, so a more extreme profile receives a proportionally stronger push. Gains were tuned separately under each rule, so we compare results across models and never gains.

J FINE-TUNED WRITERS AND READERS

Written in its own trained format, M03's portraits were also read back more faithfully than the untuned model's with the same prompt and than the production portraits, in all nine reader–instrument comparisons (under the primary reader, .929 against .827 and .859 on the BFI-2, and .891 against .784 and .818 on the PAPI people), while naming facets less often than either.

Writing and reading adaptations transfer selectively. On the 134-person audit, training the portrait writer improved continuous reading of fixed portraits (BFI-2 facet CCC from .854 to .889), and training the questionnaire reader improved portrait writing (.848 to .873 for BFI-2; .824 to .851 for IPIP-NEO). Smaller transfers, including from readers to life-story writing, were not reliable. Combining the writer with both readers (M03) gave the most faithful one-turn life stories in this comparison, although none of its parts was trained on life stories: facet CCC .805 and .635 for BFI-2 and IPIP-NEO development people, against .725 and .566 for the base model and .741 and .565 for a dedicated life-story adapter. On 64 people outside its development data, M03 raised IPIP-NEO facet agreement from .678 to .763 under our primary reader and from .654 to .701 under a second reader family. It is a writer and not a reader: its own life-story reading (.695 and .538) fell short of a dedicated reader (.839 and .686). On the 420, given the same portraits, prompt, and settings as the

Table 5: Life-story agreement with the person's questionnaire. Adding the adapters to the portrait raised domain CCC by .260 [.175, .333] and facet CCC by .203 [.135, .261] (95% paired person bootstrap; both survive correction across 64 tests), cut error (RMSE .667/.836 to .525/.728), and made the intended person the closest of the 32 profiles in 50% of stories, against 22% with the portrait alone.

Condition (one matched block)	Domain CCC	Facet CCC
Adapters, no portrait	.511	.401
Portrait only	.496	.448
Portrait + adapters	.756	.651

base model's portrait-only stories, M03 raised facet agreement under all three readers: on the BFI-2 from .716, .706, and .711 to .792, .778, and .755, on the PAPI people from .657, .657, and .640 to .722, .708, and .668, and on the Bainbridge IPIP from .673, .671, and .671 to .738, .715, and .686. Eight of the nine paired intervals lie above zero, the exception being Gemini on the Bainbridge IPIP. Without any steering, its stories matched or exceeded the base model's best corrected stories in eight of these nine comparisons. The gain is not a matter of adapter size: the portrait writer alone, scaled to the same total strength, changed agreement by at most about .04, inconsistently across readers.

We trained separate low-rank adapters on a fixed, unquantized Gemma4-31B for four tasks: writing a portrait from a profile, writing a life story, rating a portrait's facets on a continuous scale, and answering a person's questionnaire from a portrait, the last supervised with people's own item answers. Training used 320 people, selection used 66, and a 134-person audit gave every adapter the same writing and reading tests. Our writer, M03, is an exact, uncompressed sum of the portrait writer and the two portrait readers (rank 48), with no further training and no life-story adapter. All adapters are LoRA (rank 16, alpha 32, dropout .05) on the attention and MLP projections of every language-model layer of the unquantized model, trained with AdamW (learning rate 5×10^{-5}, weight decay .01, effective batch 8, loss on the completion only). The portrait writer was trained for four epochs on 640 examples, 320 people each with an Opus and an Astra portrait as targets; the two readers, whose targets are the people's own questionnaire answers, for one epoch, the questionnaire reader with an added numeric-distance loss. M03 concatenates the three updates exactly (rank 48) with no further training, and the dose-matched control scales the writer's update to M03's total norm. M03's life stories use the base model's exact prompt, portraits, and sampler (temperature 0.7, 6,144-token cap, seed 914); its own portraits use its trained portrait prompt (temperature 0.7, top-p .95, 4,096-token cap).

Moving steering into the weights. Training a model on its own steered outputs can install a single steering vector in its weights (Blank et al., 2026). We asked whether the same holds for each person's correction, so that no vector is needed when the model writes. Two adapters with the same recipe (rank 16, learning rate 5×10^{-5}, four epochs) were trained on the untuned Gemma4-31B's own life stories for 311 people outside the 420 and the calibration 58: one set written with each person's correction steering (S), the other written from the same portraits and seeds without it (P). The steering families were the ones that did best on the 420, so this is a best-case test. Run on the 420 with no steering, S beat P in six of nine reader–instrument comparisons (paired intervals above zero; the other three include zero), while P was flat or slightly below the untuned model. Under the primary reader, S kept part of what steering itself delivered (BFI-2: .716 untuned, .774 with the correction, .740 for S; PAPI people: .657, .701, .685; Bainbridge IPIP: .673, .713, .706), and it stayed below M03 in seven of nine comparisons. The transferred gain was mostly a correction shared across people. Each person's change from P to S lined up with that person's own needed correction beyond chance in the 300 PAPI people under all three readers, but only about a tenth to a fifth as strongly as live per-person steering, and not reliably in the 120 BFI-2 people.

K Composing released trait adapters

We composed the Persona Cartography adapters released for Qwen3-8B (Baines et al., 2026): an independently trained amplifier and suppressor for each Big Five domain (rank 64; we trained none). A person's prescription selects one pole per domain with a nonnegative magnitude, and the five

selected adapters are applied together as one weight update. All experiments used a 32-person BFI-2 development panel, and a primary reader rated each one-turn life story on a continuous scale.

The effect belongs to the person. Without a portrait, the correct person's prescription reached domain/facet CCC .538/.421, while a wrong person's (.043/.045), a trait-permuted one ($-.030$/$-.015$), and one with every pole reversed ($-.605$/$-.479$) did not.