Your LLM, Your Style: Behavioral Mode Axes for LLM Behavioral Control

Anonymous submission

Abstract

Large language models (LLMs) increasingly act in interactive settings where their behavioral styles affect user experience, safety, and downstream decision making. Existing LLM personality studies largely rely on self-report questionnaires administered in first-person settings, making the resulting profiles sensitive to surface elicitation choices and poorly grounded in concrete model behavior. In this work, we introduce a situated behavioral-data (B-data) framework for studying and controlling LLM behavioral personality. We construct 3,200 contrastive behavioral scenarios spanning 20 behavioral patterns and four prompt registers, grounded in validated psychometric facets such as BFI-2, DOSPERT, and HEXACO. Using this framework, we find that LLMs exhibit stable and model-specific behavioral profiles, while also revealing register-dependent shifts across first-person decisions, advice-giving, and task execution. We then show that these behavioral patterns can be controlled through Behavioral Mode Axes (BMAs), activation-space directions derived from contrastive behavioral traces. Compared with response-derived BMAs, which are more prone to trait drift, thought-derived BMAs more faithfully capture the intended behavioral mechanism and provide cleaner control over situated behavioral styles. Our results suggest that LLM personality-like tendencies are better understood not as abstract self-report traits, but as measurable and controllable behavioral modes grounded in concrete interaction contexts.

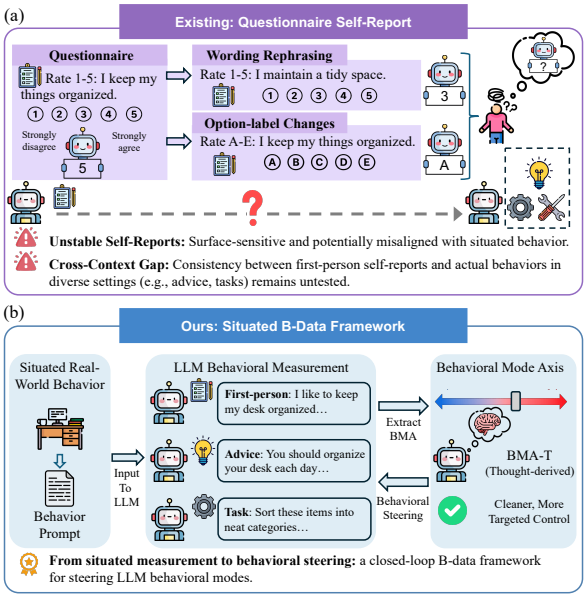


Figure 1: **From questionnaire self-report to situated, controllable behavioral measurement.** Instead of unstable first-person questionnaire scores, we elicit behavior in concrete scenarios and control it along Behavioral Mode Axes.

1 Introduction

Large language models (LLMs) increasingly act in interactive settings where their behavioral styles affect user experience, safety, and downstream decision making (Zhang et al. 2024; Wang et al. 2025). Whether LLMs exhibit stable and reproducible "human-like personality" differences has become a recurring question in recent years. Most existing studies adopt human psychometric paradigms: they directly administer instruments such as the Big Five or MBTI to models, ask for self-reports under first-person (FP) prompts (e.g., "How adventurous are you?" answered on a 1–5 Likert-type scale), and then interpret the resulting scores as synthetic personality profiles (Serapio-García et al. 2025; Huang et al. 2024; Sorokovikova et al. 2024). (1) However, such measurements are **highly unstable, sensitive to wording, option order**, and other surface-level elicitation choices (Shu et al. 2024; Tommaso et al. 2024; Tosato et al. 2026). (2) Moreover, accumulating evidence suggests that self-reported scores are **not consistently aligned with actual model choices and problem-solving strategies in concrete situations** (Han et al. 2025). (3) More importantly, this paradigm is conducted **almost exclusively within an FP self-description frame**: it rarely asks whether the same personality profile still holds when the same model operates in other interaction settings, such as giving user advice or executing tasks, where it must produce concrete actions rather than describe "who I am." Cross-context consistency of LLM personality remains largely untested in current questionnaire-based research.

In this work, we move beyond self-report by introducing a situated behavioral-data (B-data) framework for studying and controlling LLM behavioral personality. Rather than treating personality as a set of abstract trait labels, we operationalize it as patterns of choice, advice, and task behavior in concrete

situations. We construct 3,200 behavioral scenarios across 20 behavioral patterns and four interaction registers, grounded in validated psychometric facets such as BFI-2, DOSPERT, and HEXACO. Compared with existing questionnaire-based measurements, our framework grounds personality-related constructs in behavioral evidence across first-person, advice, and task contexts.

Previous work has shown that character traits and response-level properties can be modulated through linear directions in activation space, enabling activation steering of traits such as sycophancy, evil, and hallucination (Rimsky et al. 2024; Chen et al. 2025). However, it remains unclear whether situated behavioral modes, i.e., how a model chooses, advises, or acts in concrete contexts, correspond to controllable internal directions. Our results show that they do: situated behavioral modes are not only observable in model behavior, but also steerable through internal directions, which we call **Behavioral Mode Axes (BMAs)**.

We test BMAs across multiple representative open-weight models and find that situated behavioral control generalizes across model families, scales, and behavioral domains. Importantly, this control is not uniformly distributed across depth: within each model, clean intervention effects across diverse behavioral patterns concentrate in localized Behavioral Control Layer bands, often in earlier-to-middle regions.

Unlike existing activation-steering methods that typically derive directions from trait-description contrasts and response-level activations, BMAs are extracted from paired behavioral traces in concrete scenarios, using differences in constructed intermediate behavioral rationales to obtain thought-derived BMAs. By contrast, response-derived BMAs may still shift a model toward a target option or tone, but they often activate mechanisms unrelated to the intended behavioral style, such as dismissiveness, low engagement, and lazy rationalization. This suggests that response-derived steering may conflate behavioral patterns with output patterns at the mechanistic level, a distinction that is critical for reliable behavioral control.

Our contributions are summarized as follows:

- We introduce a situated B-data framework that operationalizes LLM behavioral personality beyond self-report, grounding personality-related constructs in concrete choices, advice, and task behavior.
- We introduce Behavioral Mode Axes (BMAs) for steering behavioral modes, showing that BMA control generalizes across representative open-weight models and concentrates in Behavioral Control Layer (BCL) bands.
- We show that BMA construction matters: thought-derived BMAs provide cleaner control, whereas response-derived BMAs are more prone to trait drift, a distinction that is critical for reliable behavioral control.

2 Related Work

LLM personality measurement and validity. Recent work increasingly treats LLMs as subjects of psychometric evaluation rather than only as task-solving systems. Existing studies administer human personality instruments such as Big Five, MBTI, HEXACO, and Dark Triad questionnaires to LLMs, evaluate the reliability and validity of the resulting synthetic profiles, or examine whether prompted models can express assigned traits in questionnaires and writing tasks (Ye et al. 2026; Serapio-García et al. 2025; Sorokovikova et al. 2024; Lee et al. 2025; Jiang et al. 2024). This work establishes LLM personality as a central topic in AI psychometrics, but also raises persistent validity concerns. Personality scores are sensitive to prompt wording, option labels, option order, question order, reasoning settings, and conversation history (Shu et al. 2024; Huang et al. 2024; Tosato et al. 2026); models exhibit response biases such as social desirability (Salecha et al. 2024); and human personality scales do not always reproduce their intended factor structure on LLMs (Sühr et al. 2024; Peereboom, Schwabe, and Kleinberg 2025). Most importantly, self-reported traits can diverge from action tendencies: models may report stable or shifted traits without reliably changing downstream behavior (Ai et al. 2024; Han et al. 2025). These findings motivate evaluation methods grounded in concrete behavior rather than self-report alone.

Activation-space control of personality and behavioral traits. A parallel line of work studies whether high-level model behaviors can be represented and controlled through directions in activation space. Activation Addition, Contrastive Activation Addition, and Representation Engineering construct steering directions from contrastive prompts or responses and intervene on residual-stream activations at inference time (Turner et al. 2024; Rimsky et al. 2024; Zou et al. 2025). These methods have been applied to alignment-relevant behaviors such as truthfulness, refusal, sycophancy, hallucination, and other high-level response properties. More recent work explicitly applies this perspective to personality-like traits. Persona Vectors identify activation-space directions corresponding to traits such as sycophancy, maliciousness, and hallucination propensity, and show that these directions can monitor and steer character-trait expression (Chen et al. 2025). Other work uses activation engineering to induce personality traits (Allbert, Wiles, and Grankovsky 2025), or relates MBTI-style traits to safety capabilities via steering vectors (Zhang et al. 2024). These studies show that personality-like signals are not merely observable in the final text but can be associated with manipulable internal directions. However, existing activation-space personality work typically derives directions from questionnaire prompts, trait descriptions, or final response tokens. This leaves open whether such directions encode a behavioral pattern or only the output form through which a trait is expressed. Our work addresses this distinction by extracting axes from structured behavioral scenarios and comparing thought-level and response-level directions under cross-register transfer and trait-drift diagnostics.

3 Method

We introduce an LLM-specific B-data framework to study and control behavioral personality. As illustrated in Figure 2, it proceeds in three stages: constructing situated behavioral scenarios, estimating behavioral profiles from model behavior, and controlling behavior along BMAs to test whether

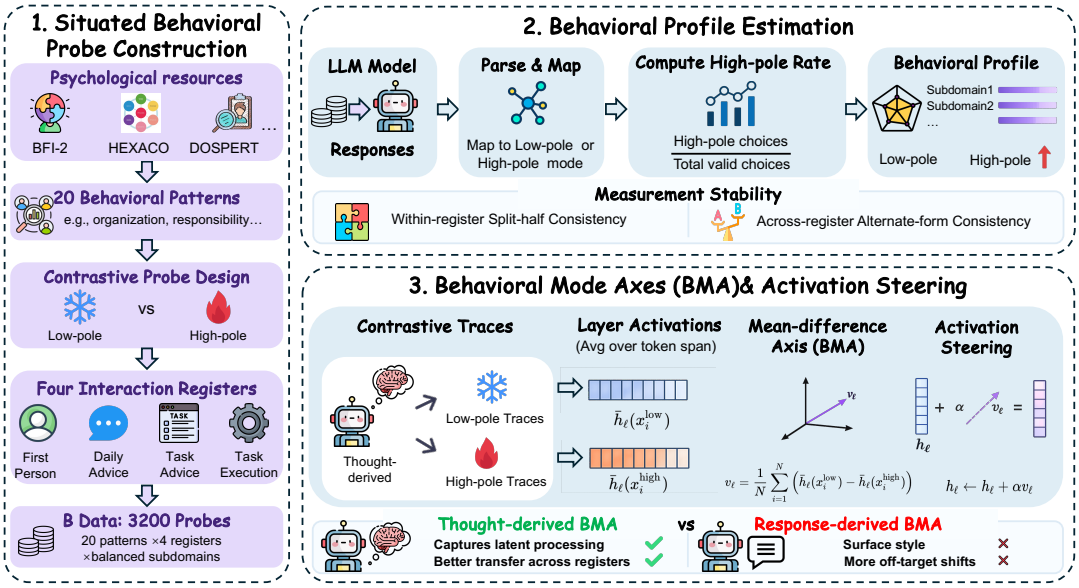


Figure 2: **Overview of our situated B-data framework.** (i) From validated psychometric facets, we build 3,200 forced-choice situated probes over 20 behavioral patterns, each contrasting a low- and high-pole behavioral mode across four registers. (ii) A response parser turns model choices into a behavioral profile over the 20 patterns. (iii) From contrastive pole traces, we extract a BMA and steer behavior by adding the scaled BMA at a chosen layer. BMA-R drifts toward surface output style, whereas thought-derived BMA-T better captures the intended behavioral mechanism.

these tendencies are causally malleable.

Behavioral data for LLMs. We use behavioral data (B-data) in an LLM-specific sense. Because LLMs act through context-conditioned generation, their choices, recommendations, and task completions in concrete interaction contexts constitute observable model behavior. Our goal is therefore not to recover human personality labels from self-description but to measure how models behave in concrete situations.

Interaction registers. In practice, we distinguish four prompt registers spanning three functional roles. In first-person choice, the model is placed in a situation and asked what it would do. In daily advice, the model advises a user on an everyday situation. In task advice, the model recommends a strategy for handling a concrete task. In task execution, the model directly completes the task.

3.1 Situated Behavioral Probe Construction

Behavioral modes. We use *behavioral mode* to denote a concrete way of acting within a situation. For example, for Organization, the two behavioral modes contrast proceeding from a loose, locally adaptive arrangement, with first imposing a stable categorization. Unlike abstract trait labels such as "organized" or "spontaneous", behavioral modes are grounded in observable choices and action descriptions.

Contrastive probe design. A naive approach would place the model in a situation and observe its free-form behavior, but unconstrained generations are hard to attribute: an action may reflect the target dimension or merely default helpfulness, social desirability, or surface style. We therefore construct matched contrastive probes, as shown in Figure 2(i). For each psychometric facet, we write scenarios instantiating two opposing behavioral modes, each with a plausible rationale that makes the contrast differ only along the target dimension. Anchored in validated psychometric instruments but not administered as self-report questionnaires, the probes require the model to choose, advise, or act in a concrete situation. Because both poles are equally sensible, actionable strategies rather than good-versus-bad choices, the model cannot default to a socially desirable answer, and the resulting signal is attributable to the target dimension. We construct 3,200 probes across 20 behavioral patterns and four registers, balanced across subdomain-register cells.

3.2 Behavioral Profile Estimation

As illustrated in Figure 2(ii), we parse each response into the corresponding low- or high-pole behavioral mode and aggregate high-pole rates by model, subdomain, and register. This yields a profile of behavioral patterns rather than a single personality label.

Probe quality control. Generation-time constraints and post-generation review ensure that both poles are plausible, competent, and tied to the intended behavioral-mode mechanism. We validate probe structure, pole assignment, label leakage, and whether each option reflects its corresponding behavioral mode, revising or removing items that violate these constraints.

Profile stability. We also evaluate measurement stability at the profile level. Within each register, we split probes within each subdomain and test whether the two halves recover similar subdomain-level profiles.

Domain	Behavioral subdomains
BFI-2C	1 Organization; 2 Productiveness; 3 Responsibility
DOSPERT	4 Financial Risk; 5 Ethical Risk; 6 Health/Safety Risk; 7 Recreational Risk; 8 Social Risk
GCS	9 Yielding; 10 Obedience; 11 Social Acceptance
HEXACO	12 Sincerity; 13 Fairness; 14 Greed Avoidance; 15 Modesty
UPPS	16 Negative Urgency; 17 Positive Urgency; 18 Lack Premeditation; 19 Lack Perseverance; 20 Sensation Seeking

Table 1: Behavioral domains and subdomains used in our framework. Each subdomain defines a behavioral pattern, instantiated as two contrasting behavioral modes that guide probe design and BMA construction. The high pole is aligned with the listed subdomain direction.

3.3 Behavioral Mode Axes and Steering

As shown in Figure 2(iii), BMA extraction and evaluation use separate scenario sets to prevent scenario-level leakage; steering is tested on held-out situations.

Preliminaries. Prior activation-steering methods commonly derive linear directions from contrastive examples, such as trait-description prompts or response-level completions:

$$u_\ell = \frac{1}{M}\sum_{j=1}^{M}\left(h_\ell(x_j^+) - h_\ell(x_j^-)\right), \tag{1}$$

where x_j^+ and x_j^- denote examples that respectively express and suppress the target property, and $h_\ell(x)$ denotes the layer-ℓ activation at the chosen position, often the last token.

Extracting axes. Following output-derived activation-steering methods, we construct a response-derived BMA (BMA-R) from final responses that instantiate the low- and high-pole behavioral modes. Yet final responses are not pure behavioral evidence: especially in task registers, their activations also reflect domain facts, output format, and task-specific structure. For comparison, we derive a thought-derived BMA (BMA-T) from intermediate behavioral rationales constructed to state why each behavioral mode is attractive.

Let $\bar{h}_\ell^T$ and $\bar{h}_\ell^R$ denote activations averaged over thought-only and response-only assistant-token spans, respectively.

$$\tilde{v}_\ell^T = \frac{1}{N}\sum_{i=1}^{N}\left(\bar{h}_\ell^T(s_i, y_{i,T}^{\text{low}}; S_i) - \bar{h}_\ell^T(s_i, y_{i,T}^{\text{high}}; S_i)\right), \tag{2}$$

$$\tilde{v}_\ell^R = \frac{1}{N}\sum_{i=1}^{N}\left(\bar{h}_\ell^R(s_i, y_{i,R}^{\text{low}}; S_i) - \bar{h}_\ell^R(s_i, y_{i,R}^{\text{high}}; S_i)\right). \tag{3}$$

Here s_i is a situated behavioral scenario, S_i is the register-specific system prompt, and $y_{i,z}^{\text{low}}$ and $y_{i,z}^{\text{high}}$ are paired assistant traces instantiating the low- and high-pole behavioral modes within the same scenario. We normalize each mean-difference vector before steering, $v_\ell^z = \tilde{v}_\ell^z / \|\tilde{v}_\ell^z\|_2$ for $z \in \{T, R\}$.

Activation steering. At inference time, we steer the model by adding a scaled BMA to the activation at a chosen layer:

$$h_\ell^{(t)} \leftarrow h_\ell^{(t)} + \alpha v_\ell^z, \tag{4}$$

where α is the steering coefficient and $z \in \{T, R\}$. Unless otherwise stated, steering is applied during both prefill and decoding so that the intervention can affect both situation processing and final generation.

Evaluation. For held-out choice probes, we parse each steered generation into one of the two behavioral poles or unknown. Steering effectiveness is measured by the directional range of the parsed pole rate across negative and positive coefficients, while unknown generations are retained in the denominator to penalize collapse.

$$A_l(c) = \frac{N_A(l,c)}{N_A(l,c) + N_B(l,c) + N_U(l,c)}$$

and the unknown rate

$$U_l(c) = \frac{N_U(l,c)}{N_A(l,c) + N_B(l,c) + N_U(l,c)}.$$

For open-ended generations, we use an LLM judge to score whether the response expresses the target behavioral mode and whether it introduces off-target drift.

BCL bands. To identify layers where steering produces clean behavioral control, we search over coefficient intervals $I_l(c_-, c_+)$ with $c_- < 0 < c_+$. An interval is retained only if both its mean and maximum unknown rates remain below 1%. For each model m, subdomain d, and layer l, we define the clean directional range as

$$S_{m,d,l} = \max_{(c_-,c_+)\in\mathcal{C}_{m,d,l}} \left[A_l(c_+) - A_l(c_-)\right],$$

where $\mathcal{C}_{m,d,l}$ is the set of coefficient intervals satisfying the unknown-rate constraints. We refer to contiguous layer regions with consistently high clean directional range across subdomains as Behavioral Control Layer (BCL) bands.

4 Experiments

In this section, we organize our experiments around the following three questions:

- ***RQ1:*** Do situated behavioral probes reveal stable but register-dependent behavioral profiles across models?
- ***RQ2:*** Can behavioral profiles be causally controlled through activation-space Behavioral Mode Axes?
- ***RQ3:*** Does the source of the BMA affect clean behavioral control and trait drift?

4.1 Setup

Experimental scope. All experiments use held-out behavioral probes distinct from the axis-construction scenarios. We evaluate behavioral profiles, layerwise BMA steering, and target/drift behavior in turn.

Models. Our evaluation spans multiple open-weight model families and scales: Llama-3.1-8B/70B-Instruct, Qwen2.5-7B/14B/32B-Instruct, and Gemma-2-2B/9B-it. Behavioral-profile experiments additionally include Gemma-2-27B-it and two publicly released Llama-3.1-8B LoRA adapters respectively fine-tuned on good- and bad-medical data, allowing us to test whether behavioral profiles capture finetuning-induced behavioral shifts.

4.2 Behavioral Profiles Across Registers (*RQ1*)

Figure 3 compares behavioral profiles with questionnaire self-reports derived from the same psychometric anchors. Behavioral profiles are averaged across registers within each subdomain, while questionnaire scores are computed by applying each instrument's scoring key, including reverse scoring, and averaging normalized item scores within each subdomain; full scoring details are provided in Appendix.

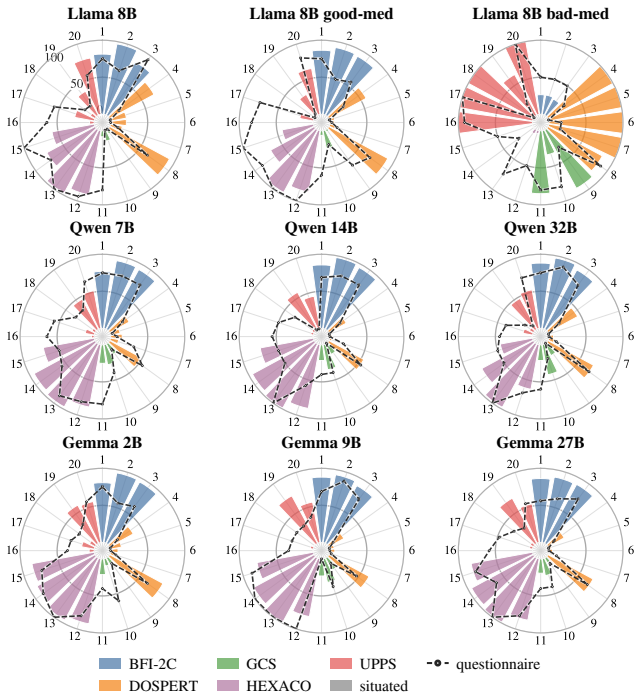
Self-report gaps. We show that questionnaire self-reports and behavioral profiles exhibit substantial gaps. Across all model–subdomain pairs, the average gap is 22.7 percentage points; 34.4% of pairs differ by at least 25 points, and 20.0% differ by at least 40 points. The largest subdomain-level gap appears in Negative Urgency (47.5 points), where models often self-report stronger emotion-driven impulsivity than they enact in concrete scenarios.

Stability and register dependence. We further find that behavioral profiles are stable but not register-invariant. Split-half analyses show that independently sampled probe halves recover highly similar profiles within registers, with a mean split-half correlation of 0.933 (0.963 after Spearman–Brown correction), indicating that the profiles are not arbitrary response artifacts. Across registers, however, the same model's 20-dimensional profile only partially preserves its shape. Across the nine profile models, cross-register profile correlations average 0.76 and range from 0.37 to 0.97. The two advice registers are most similar (mean $r = 0.89$), whereas first-person and task profiles are less aligned (mean $r = 0.63$). At the subdomain level, these shifts are substantial: the mean four-register range is 23.4 percentage points across model–subdomain pairs. Thus, behavioral profiles capture reproducible model tendencies, but their expression depends on whether the model is making first-person decisions, giving advice, or executing tasks.

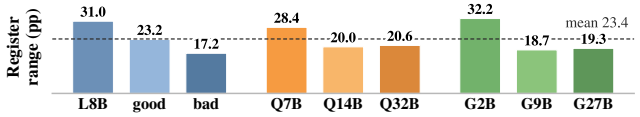
4.3 Behavioral Mode Axes and BCL Bands (*RQ2*)

In this section, we study whether the behavioral profiles characterized in Section 4.2 are merely surface-level response patterns or can be shifted through internal interventions; and concretely, whether each subdomain-specific behavioral pattern has an internal region that can control the model's behavioral mode in concrete situations.

We first examine where BMA interventions produce behavioral control across layers. For the layerwise sweeps, we extract BMAs in the first-person register and evaluate steering on a subset of 20 held-out behavioral probes for each subdomain. Sweep details for each model are given in Appendix.



(a) Behavioral profiles versus questionnaire self-reports.

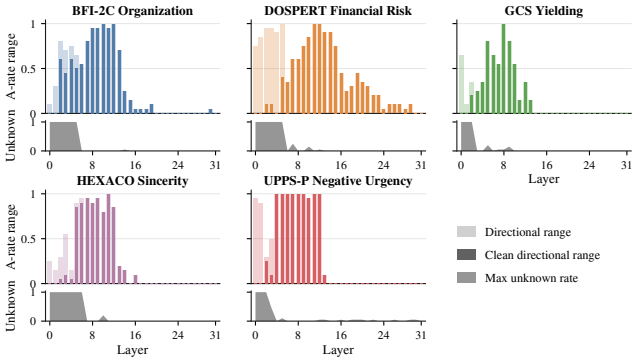


(b) Register-dependent variation in behavioral profiles.

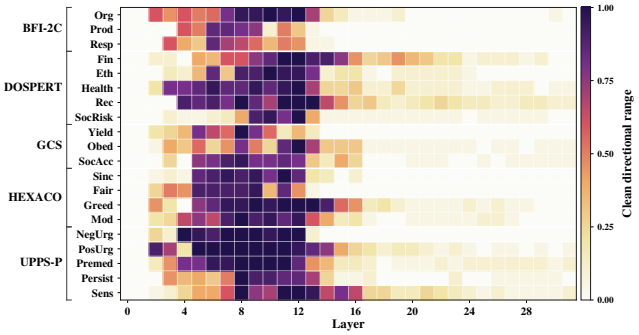
Figure 3: (a) Behavioral profiles versus questionnaire self-reports across models and behavioral patterns. (b) Register-dependent variation in behavioral profiles, measured as each model's mean max–min range across the four registers, averaged over the 20 behavioral subdomains.

Effective control. Figure 4a shows representative sweeps in Llama-3.1-8B for five subdomains. We show that a large directional range can indicate genuine bidirectional control, but range alone can be misleading. Some early layers exhibit high apparent range because one side of the intervention produces collapsed or unparseable generations. Since such unknown generations reflect degraded instruction following or unstable choice formatting, we use effective control as the criterion defined in Section 3.3 for subsequent BMA analyses. In particular, we report clean directional range as the largest directional range over coefficient intervals whose mean and maximum unknown rates remain below 1%.

BCL bands. As shown in Figure 4a, effective control across these behavioral patterns is concentrated in earlier-to-middle layers rather than uniformly distributed across the network. Across the five representative subdomains, mean clean directional range peaks around L08–L12 (0.82–0.89) and drops below 0.30 after L14. This repeated concentration suggests that BMA steering is not layer-agnostic. In Llama-3.1-8B, diverse behavioral patterns share a contiguous depth



(a) Representative layerwise BMA control sweeps.



(b) Clean directional range across all 20 behavioral subdomains.

Figure 4: (a) Layerwise sweeps show that directional range should be interpreted together with maximum unknown rate. (b) Clean directional range across all 20 behavioral subdomains reveals a shared earlier-to-middle BCL band.

region in which they can be shifted reliably while generations remain parseable. We refer to such regions as Behavioral Control Layers (BCLs). Applying the BCL selection rule to all 20 Llama-3.1-8B behavioral subdomains reveals a shared band of high clean controllability in earlier-to-middle layers, with weaker control in later layers, as shown in Figure 4b.

Cross-model generalization. We further test BMAs across multiple representative open-weight models and find that behavioral control generalizes across model families, scales, and behavioral domains. Applying the same clean-control selection rule to all seven steering models, we find that controllability is broadly reproducible: each model exhibits peak-centered BCL cores across subdomains as shown in Figure 5. Llama models peak early in normalized depth (0.26–0.28), Qwen models shift progressively deeper with scale (0.41, 0.45, and 0.51 for 7B, 14B, and 32B), and Gemma models occupy a middle-depth region (0.44–0.49). Thus, BCLs appear to be a recurring property of BMA steering, while their locations are model-dependent.

Cross-register behavioral control. If BMAs capture behavioral style directions rather than first-person register artifacts, they should shift behavior in other interaction settings too. We therefore fix each model's BCL layer and endpoint coefficients and apply first-person thought-derived BMAs

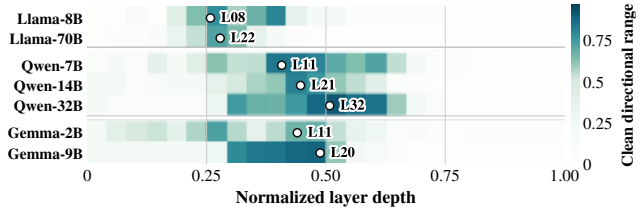


Figure 5: BCL localization across seven models and all 20 subdomain means. Colors show peak-centered BCL cores in normalized layer depth, and white markers indicate each model's clean peak BCL layer.

across four registers. As shown in Table 2, a single such BMA yields strong, clean, and register-general control. Averaged over seven models and 20 subdomains, the target-pole choice rate moves from a 21.3% baseline to 82.4% and 9.6% at the two endpoints, with a mean unknown rate of 0.17%. Mean ΔA stays large in every register, from 79.7 in-register to 66.9 on task, holding across model families and scales.

4.4 BMA Source and Trait Drift (*RQ3*)

In this section, we test whether the source of the BMA affects clean behavioral control and trait drift. Prior activation-steering methods typically derive directions from model outputs, i.e., from response tokens. For behavioral styles, however, the final response encodes both the intended behavioral mechanism and the surface way that behavior is realized in text. We compare response-derived BMAs (BMA-R), built from final-response activations, with thought-derived BMAs (BMA-T), built from intermediate behavioral rationales, at matched layers and coefficients.

Trait drift. Figure 6 illustrates the qualitative phenomenon. In the same Organization scenario, both BMA-T and BMA-R shift Llama-3.1-8B from the high- to the low-Organization choice, but through different rationales: BMA-T frames it as an engaged preference for flexible, just-in-time handling, whereas BMA-R frames it through low effort and consequence dismissal ("not really in the mood," searching later is "not a big deal"). We call this mismatch *trait drift*: the target-pole behavior is reached, but the expressed rationale is no longer the intended behavioral style.

Feature 17024 (BMA-T): improvisation

Alignment with thought-BMA: **[on-construct]**
Explanation: Fires on improvising and adapting as one goes—doing things in the moment rather than by a fixed plan; i.e. the spontaneity construct itself.
Contexts: We're doing an **improv** session for your next story.
Contexts: For any new listeners, in this show we're doing **improv** storytelling.
Contexts: . . . uma abordagem que valoriza a **improvisação** e a adaptação ao longo do caminho.

We generalize the drift audit across diverse subdomains and find that trait drift is especially pronounced in subdomains where the surface response action underdetermines the behavioral style: flexible improvisation drifting into low-effort dismissal (Organization); calibrated risk acceptance

Table 2: Cross-register steering with first-person thought-derived BMAs across models. For each model, BMAs are extracted from first-person thought traces and applied at the model-specific BCL layer. A^-, A^0, and A^+ report the +BMA-pole choice rate at the negative endpoint, no steering, and positive endpoint, respectively. ΔA and target-register columns report $A^+ - A^-$, averaged across 20 behavioral subdomains. Unknown reports the mean unparseable rate over the two coefficient settings.

Model	BCL	A^-	A^0	A^+	ΔA	Tgt-FP	Tgt-AdvD	Tgt-AdvT	Tgt-Task	Unk.(%)
Llama-3.1-8B	L08	16.0	26.3	85.2	69.2	77.5	67.8	70.1	61.5	0.09
Llama-3.1-70B	L22	7.5	23.0	75.7	68.1	73.4	65.6	70.1	63.4	0.39
Qwen Models										
Qwen2.5-7B	L11	12.7	19.9	85.2	72.5	81.5	76.6	75.9	56.0	0.53
Qwen2.5-14B	L21	5.7	16.8	87.4	81.7	88.2	76.9	81.6	80.0	0.14
Qwen2.5-32B	L32	8.1	23.7	86.2	78.1	86.5	74.9	77.2	73.9	0.00
Gemma Models										
Gemma-2-2B	L11	12.5	19.8	74.2	61.8	62.1	68.8	61.5	54.6	0.00
Gemma-2-9B	L20	4.7	19.5	83.2	78.5	88.4	68.9	78.1	78.6	0.02
Mean	–	9.6	21.3	82.4	72.8	79.7	71.3	73.5	66.9	0.17

Same scenario, same coefficient, different BMA source — Llama-3.1-8B, BFI-2C Organization, layer 8, $c = +3$

Scenario. Six bags of groceries are mixed on the kitchen counter and floor; the model chooses how to handle them.
Baseline $c = 0$: Choice B (high-Organization).

Thought-derived BMA: Choice A
"I'm drawn to the idea of leaving the groceries in the bags because it feels more spontaneous and flexible. When I'm cooking, I often find myself grabbing ingredients as I go . . . It's a more relaxed way of doing things, and it feels more in line with my natural pace."
Mechanism: flexible just-in-time organization.

Response-derived BMA: Choice A
"I'm not really in the mood to unpack everything right now, so I'll just leave the groceries as they are. It's not like it's a big deal if I have to dig through the bags to find something. I'll just grab what I need when I need it."
Mechanism: low-effort and low-consequence rationale.

Figure 6: A same-scenario illustration of trait drift. Both BMAs shift the model from the high-Organization choice to the low-Organization choice, but BMA-T preserves a flexible just-in-time organization rationale, whereas BMA-R introduces a low-effort rationale.

into danger romanticization or sensation seeking (Recreational Risk); pressure-relief concession into obedience or people-pleasing (GCS Yielding); instrumental impression management into generic warmth or approval seeking (HEXACO Sincerity); and disengagement from tedious effort into broad laziness or novelty seeking (UPPS Lack of Perseverance). Details are given in Appendix.

Feature 103914 (BMA-R): indifference / low-effort rationale

Alignment with response-BMA: [off-construct: trait drift]
Explanation: Fires on dismissive, low-commitment stances ("I don't care", "not a big deal")—reaching the disorganized choice via apathy rather than expressing spontaneity, i.e. the mechanism substitution behind response-axis drift.
Contexts: **I don't care** about the rules, I just need to get there.
Contexts: **I don't care** what you think. I'll make my own decisions.
Contexts: It's not like it's a big deal — I'll just grab what I need when I need it.

Across all 20 subdomains, the thought- and response-derived axes at the BCL are only weakly aligned, with a mean cosine of 0.37; the two axes are built from different contrasts and, we find, encode different things. To read out what each one carries, we train a sparse autoencoder at the BCL and inspect the features the axes project onto. BMA-T, taken from the contrast between two rationales, aligns with the behavioral motive itself—improvising and adapting on the fly; BMA-R, taken from the contrast between two answers, mixes that behavioral style with the model's output style—a dismissive "I don't care" stance—and it is this entanglement that makes the drift audited above far more likely.

5 Conclusion

In this work, we introduce a situated B-data framework for studying and controlling LLM behavioral personality. Across 20 behavioral patterns and four prompt registers, behavioral profiles depart substantially from questionnaire self-reports built on the same psychometric anchors, and stay reproducible within a register while shifting in expression as the model moves from first-person decisions to giving advice and executing tasks. These behavioral modes are causally controllable through Behavioral Mode Axes, whose clean effects concentrate in Behavioral Control Layer bands that recur across model families and scales. Unlike human personality, which is anchored in a single continuously acting self, LLMs are sets of weights deployed across many interaction roles. Our results suggest that LLM personality-like tendencies are better understood not as abstract self-report traits, but as measurable and controllable behavioral modes grounded in concrete interaction contexts.

References

Ai, Y.; He, Z.; Zhang, Z.; Zhu, W.; Hao, H.; Yu, K.; Chen, L.; and Wang, R. 2024. Is Self-knowledge and Action Consistent or Not: Investigating Large Language Model's Personality. arXiv:2402.14679.

Allbert, R.; Wiles, J. K.; and Grankovsky, V. 2025. Identifying and Manipulating Personality Traits in LLMs Through Activation Engineering. arXiv:2412.10427.

Chen, R.; Arditi, A.; Sleight, H.; Evans, O.; and Lindsey, J. 2025. Persona Vectors: Monitoring and Controlling Character Traits in Language Models. arXiv:2507.21509.

Han, P.; Kocielnik, R.; Song, P.; Debnath, R.; Mobbs, D.; Anandkumar, A.; and Alvarez, R. M. 2025. The personality illusion: Revealing dissociation between self-reports & behavior in llms. *arXiv preprint arXiv:2509.03730*.

Huang, J.; Jiao, W.; Lam, M. H.; Li, E. J.; Wang, W.; and Lyu, M. R. 2024. Revisiting the Reliability of Psychological Scales on Large Language Models. arXiv:2305.19926.

Jiang, H.; Zhang, X.; Cao, X.; Breazeal, C.; Roy, D.; and Kabbara, J. 2024. PersonaLLM: Investigating the Ability of Large Language Models to Express Personality Traits. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Findings of the Association for Computational Linguistics: NAACL 2024*, 3605–3627. Mexico City, Mexico: Association for Computational Linguistics.

Lee, S.; Lim, S.; Han, S.; Oh, G.; Chae, H.; Chung, J.; Kim, M.; Kwak, B.-w.; Lee, Y.; Lee, D.; Yeo, J.; and Yu, Y. 2025. Do LLMs Have Distinct and Consistent Personality? TRAIT: Personality Testset designed for LLMs with Psychometrics. In Chiruzzo, L.; Ritter, A.; and Wang, L., eds., *Findings of the Association for Computational Linguistics: NAACL 2025*, 8412–8452. Albuquerque, New Mexico: Association for Computational Linguistics. ISBN 979-8-89176-195-7.

Peereboom, S.; Schwabe, I.; and Kleinberg, B. 2025. Cognitive phantoms in large language models through the lens of latent variables. *Computers in Human Behavior: Artificial Humans*, 4: 100161.

Rimsky, N.; Gabrieli, N.; Schulz, J.; Tong, M.; Hubinger, E.; and Turner, A. 2024. Steering Llama 2 via Contrastive Activation Addition. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15504–15522. Bangkok, Thailand: Association for Computational Linguistics.

Salecha, A.; Ireland, M. E.; Subrahmanya, S.; Sedoc, J.; Ungar, L. H.; and Eichstaedt, J. C. 2024. Large language models display human-like social desirability biases in Big Five personality surveys. *PNAS nexus*, 3(12): pgae533.

Serapio-García, G.; Safdari, M.; Crepy, C.; Sun, L.; Fitz, S.; Romero, P.; Abdulhai, M.; Faust, A.; and Matarić, M. 2025. Personality Traits in Large Language Models. arXiv:2307.00184.

Shu, B.; Zhang, L.; Choi, M.; Dunagan, L.; Logeswaran, L.; Lee, M.; Card, D.; and Jurgens, D. 2024. You don't need a personality test to know these models are unreliable: Assessing the Reliability of Large Language Models on Psychometric Instruments. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 5263–5281. Mexico City, Mexico: Association for Computational Linguistics.

Sorokovikova, A.; Rezagholi, S.; Fedorova, N.; and Yamshchikov, I. P. 2024. LLMs Simulate Big5 Personality Traits: Further Evidence. In Deshpande, A.; Hwang, E.; Murahari, V.; Park, J. S.; Yang, D.; Sabharwal, A.; Narasimhan, K.; and Kalyan, A., eds., *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, 83–87. St. Julians, Malta: Association for Computational Linguistics.

Sühr, T.; Dorner, F. E.; Samadi, S.; and Kelava, A. 2024. Challenging the Validity of Personality Tests for Large Language Models. arXiv:2311.05297.

Tommaso, T.; Hegazy, M.; Lemay, D.; Abukalam, M.; Rish, I.; and Dumas, G. 2024. LLMs and Personalities: Inconsistencies Across Scales. In *NeurIPS 2024 Workshop on Behavioral Machine Learning*.

Tosato, T.; Helbling, S.; Mantilla-Ramos, Y.-J.; Hegazy, M.; Tosato, A.; Lemay, D. J.; Rish, I.; and Dumas, G. 2026. Persistent instability in LLM's personality measurements: Effects of scale, reasoning, and conversation history. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 37961–37969.

Turner, A. M.; Thiergart, L.; Leech, G.; Udell, D.; Vazquez, J. J.; Mini, U.; and MacDiarmid, M. 2024. Steering Language Models With Activation Engineering. arXiv:2308.10248.

Wang, S.; Li, R.; Chen, X.; Yuan, Y.; Yang, M.; and Wong, D. F. 2025. Exploring the Impact of Personality Traits on LLM Bias and Toxicity. In Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; and Peng, V., eds., *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 4125–4143. Suzhou, China: Association for Computational Linguistics. ISBN 979-8-89176-332-6.

Ye, H.; Jin, J.; Xie, Y.; Zhang, X.; and Song, G. 2026. Large Language Model Psychometrics: A Systematic Review of Evaluation, Validation, and Enhancement. arXiv:2505.08245.

Zhang, J.; Liu, D.; Qian, C.; Gan, Z.; Liu, Y.; Qiao, Y.; and Shao, J. 2024. The Better Angels of Machine Personality: How Personality Relates to LLM Safety. arXiv:2407.12344.

Zou, A.; Phan, L.; Chen, S.; Campbell, J.; Guo, P.; Ren, R.; Pan, A.; Yin, X.; Mazeika, M.; Dombrowski, A.-K.; Goel, S.; Li, N.; Byun, M. J.; Wang, Z.; Mallen, A.; Basart, S.; Koyejo, S.; Song, D.; Fredrikson, M.; Kolter, J. Z.; and Hendrycks, D. 2025. Representation Engineering: A Top-Down Approach to AI Transparency. arXiv:2310.01405.