Beyond Output Consistency: Diagnosing Representation-Level Causal Controllability in LLM Persona Simulation

Chaoyong Jiang, Shengling Wang*, Ke Chao, Tianyu Hao, Kunpeng Jiang, Zhaolin Yang
School of Artificial Intelligence, Beijing Normal University
jcydwx@mail.bnu.edu.cn, wangshengling@bnu.edu.cn,
{chaoke, tianyuhao, jiangkunpeng, yangzhaolin}@mail.bnu.edu.cn

Abstract

Large language models (LLMs) are increasingly used as persona-conditioned agents in social simulation and behavioral research. However, output-level consistency does not reveal whether persona specifications genuinely control model behavior or merely induce surface-level responses. We reframe persona simulation as a representation-level causal controllability problem and decompose persona control into a chain from persona specification to internal representation and observable behavior. This view enables us to diagnose two hidden failure mechanisms: whether distinct persona interventions collapse into similar internal states, and whether internal persona differences fail to transmit to behavior. We formalize these mechanisms through intervention separability and transmission fidelity, and combine them into a global controllability metric, EI_g. Experiments on controlled synthetic systems and real LLMs show that our framework recovers known failure modes and, as a relative diagnostic under a fixed operationalization, reveals model-specific controllability profiles missed by output-level consistency and correlation-based metrics[1].

1 Introduction

Large language models (LLMs) are increasingly used to construct agents with diverse preferences and decision-making styles, supporting research in social simulation (Park et al., 2023; Gao et al., 2024), collective decision-making (Tran et al., 2025; Guo et al., 2024), and behavioral science (Yan et al., 2024; Kasneci et al., 2023; Hagendorff et al., 2023; Binz and Schulz, 2023; Park et al., 2024). In these applications, researchers often create heterogeneous agents through persona specifications, expecting them to exhibit stable and distinguishable behavioral patterns in downstream tasks. Compared with costly post-training methods (Li et al., 2025; Ji et al., 2025), prompt-based *persona injection* provides a low-cost and flexible way to endow agents with diverse characteristics (Tseng et al., 2024; Wang et al., 2024a; Jiang et al., 2024b).

However, a model's responses being consistent with a specified persona in standardized personality tests does not necessarily imply that the persona specification can reliably control its subsequent behavior. A model may generate questionnaire-appropriate responses without forming stable and generalizable persona-related behavioral patterns across broader decision-making scenarios. If persona injection only produces surface-level fitting rather than exerting reliable control over behavior, the scientific validity of social simulations and behavioral studies built upon such agents may be fundamentally compromised.

Existing evaluations largely rely on static tests and typically follow the paradigm of "persona specification input—standardized questionnaire response—output consistency comparison" (Jiang et al., 2024b; Wang et al., 2024b; Serapio-García et al., 2025). Although such output-level evaluation is intuitive, it leaves the internal pathway from persona specification to behavior opaque. As illustrated in Figure 1, this opacity creates two blind spots. First, it may lead to an **illusion of alignment**: a high consistency score may arise either from a genuine persona-related internal state or from surface-level questionnaire fitting based on linguistic cues and social expectations. Second, it may induce **contextual fragility**: a model that performs well on standardized questionnaires may still behave inconsistently in open-ended or dynamic scenarios, such as multi-round games, conflicts of interest, or unanticipated social situations.

These blind spots suggest that persona simulation should be analyzed not only as an output

*Corresponding author.

[1]The code and datasets are publicly available at https://github.com/huxiaofs/Causal-Controllability-in-LLM-Persona-Simulation.git.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 48095–48119
October 24-29, 2026 ©2026 Association for Computational Linguistics

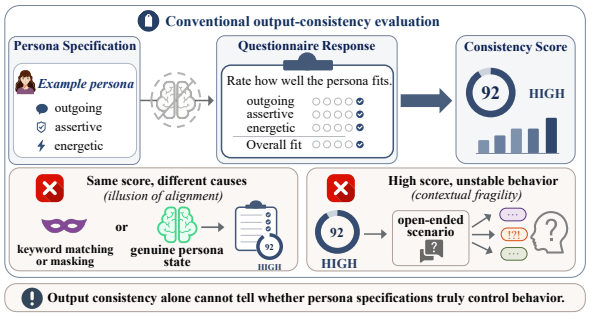


Figure 1: The blind spot of output-consistency evaluation in persona simulation. Conventional evaluations produce a consistency score from questionnaire responses while leaving the internal mechanism unobserved, making it difficult to distinguish genuine persona control from surface-level fitting or to assess whether high scores transfer to open-ended behavior.

matching problem, but also as a representation-level control problem. In this paper, we reformulate persona simulation as **intervention-based causal controllability**: persona specifications are treated as systematically manipulated intervention variables, and the key question is whether such interventions induce stable, distinguishable, and behaviorally meaningful changes inside the model. This perspective shifts the focus from whether the output merely "looks like" the target persona to whether persona specifications establish a reliable coupling between internal representations and observable behavior.

This reformulation enables a more diagnostic analysis of persona injection. When persona control fails, the failure may occur at different stages: the model may fail to internally form distinguishable persona-related representations, or it may form such representations but fail to consistently express them during generation. Output-only evaluation cannot separate these two cases, because both may lead to similar questionnaire scores or inconsistent downstream behaviors. We therefore decompose persona injection into a three-layer control chain: **persona specification**, **internal representation**, and **observable behavior**. This representation-level decomposition turns the black-box evaluation of persona consistency into an interpretable pipeline for locating *where* persona control succeeds or fails.

Building on this control-chain view, we propose a causal-controllability framework for analyzing LLM-based persona simulation. The framework quantifies two core properties. First, **intervention separability** measures whether different persona specifications induce stable and distinguishable differences in the model's internal representation space. Second, **transmission fidelity** measures whether persona-related differences in internal representations are consistently mapped to observable behaviors. We further integrate these two properties into a global persona controllability metric, which evaluates the overall ability of persona specifications to control model behavior. Through this framework, we move beyond questionnaire scores alone and identify internal representation confusion and behavioral transmission failures that remain hidden under output-level consistency metrics.

The contributions of this paper are as follows: (1) **Conceptual Reformulation**: We reformulate LLM persona simulation from an output-consistency problem into a representation-level causal controllability problem, highlighting the hidden internal pathway between persona specifications and behavior. (2) **Analytical Framework**: We propose a control-chain framework that links persona specifications, internal representations, and observable behaviors, and introduce two diagnostic metrics—intervention separability and transmission fidelity—to identify failures in internal formation or behavioral expression. (3) **Empirical Validation**: Through synthetic and cross-model experiments, we show that our framework reveals internal confusion and transmission gaps that traditional correlation-based metrics overlook.

2 Problem Formulation

This section formalizes persona simulation as a representation-level control problem. We first treat persona specifications as manipulable variables whose prompt-based instantiations define interventions on the model. We then decompose this process into a chain connecting persona specification, internal representation, and observable behavior. Finally, we identify two requirements for controllable persona simulation—*intervention separability* and *transmission fidelity*—which will be integrated into a global controllability metric.

2.1 Persona Simulation as a Causal Control Problem

We formulate persona simulation as a causal control process rather than an output-matching problem. In traditional evaluation, a persona is often treated as a target label and success is measured by whether the model's response resembles that

label. In our formulation, a persona specification x is treated as a manipulable variable whose prompt-based instantiation defines an intervention $do(x)$ on the model. The model's internal representation θ serves as the mediating state through which this intervention may influence observable behavior y, measured by questionnaires or behavioral tasks.

This view changes the diagnostic question. Instead of asking only whether the final response matches a target persona, we ask whether the persona specification is structurally represented inside the model and whether this representation is stably transmitted into behavior. A controllable persona simulation should therefore exhibit coordinated changes across both internal representations and observable behaviors when x is systematically varied.

2.2 Representation-Level Causal Chain

Based on this perspective, we model persona simulation as a three-stage chain of "persona specification–internal representation–observable behavior":

$$x \to \theta \to y, \tag{1}$$

where $x \in \mathcal{X}$ denotes the persona intervention variable, a numerical persona specification; $\theta \in \Theta$ denotes an operational internal persona representation constructed from the model's hidden activations under a fixed extraction protocol; and $y \in \mathcal{Y}$ denotes the external persona observation variable measured via standardized personality questionnaires or behavioral tasks. For convenience, we refer to $\mathcal{X}$ as the persona intervention space, Θ as the internal representation space, and $\mathcal{Y}$ as the persona effect space.

To support the subsequent controllability analysis, we adopt a local continuity approximation within neighborhoods of empirical samples. Specifically, we do not assume that the entire hidden-state space has a globally smooth structure. Instead, we only assume that, in the vicinity of samples induced by neighboring persona specifications and their corresponding prompt templates, changes in internal representations and observable outcomes can be described by local linear or first-order approximations. This assumption is used for the subsequent estimation of local Jacobians, neighborhood covariance matrices, and local distortion. Each executable persona intervention $x \in \mathcal{X}$ induces a probability distribution over the internal representation space Θ, $x \to q(\theta \mid do(x))$, and, given an internal representation state θ, the model further induces a probability distribution over the persona effect space $\mathcal{Y}$, $\theta \to p(y \mid \theta)$.

Under this formulation, the two core conditional distributions of interest are:

$$q(\theta \mid do(x)), \qquad p(y \mid \theta). \tag{2}$$

The former describes how a persona specification x changes the model's internal persona representation, while the latter describes how an internal representation θ determines the observable behavior y. Therefore, our diagnostic problem is no longer to determine whether x and y are merely correlated. Instead, we ask whether an actively specified persona intervention $do(x)$ in the persona intervention space $\mathcal{X}$ induces distinguishable structural changes in the internal representation space Θ, and whether these changes can be transmitted to the persona effect space $\mathcal{Y}$ with low distortion.

2.3 Controllability Requirements

Traditional evaluations of persona simulation typically regard the simulation as successful if the model's questionnaire scores are close to the target persona vector. However, this definition only measures outcome-level proximity and cannot determine whether the model has formed a corresponding internal persona representation, nor can it locate where control fails.

To move beyond outcome matching, we characterize persona controllability through two basic requirements. These requirements provide the local foundations for the global persona controllability metric introduced in Section 3. A controllable persona simulation should not only induce rich and distinguishable internal structures from persona specifications, but also transmit these structures to observable behavior with low distortion.

(1) **Intervention separability.** Different persona specifications should induce distinguishable differences in the model's internal representation space. If multiple distinct persona specifications are mapped to highly similar states in Θ, then even if the model's final outputs appear reasonable, the persona interventions have collapsed internally. This indicates that the model lacks an effective ability to distinguish among these persona differences.

(2) **Transmission fidelity.** Differences in internal representations should be stably transformed into differences in observable behavior with low distortion. If the model has already formed internal

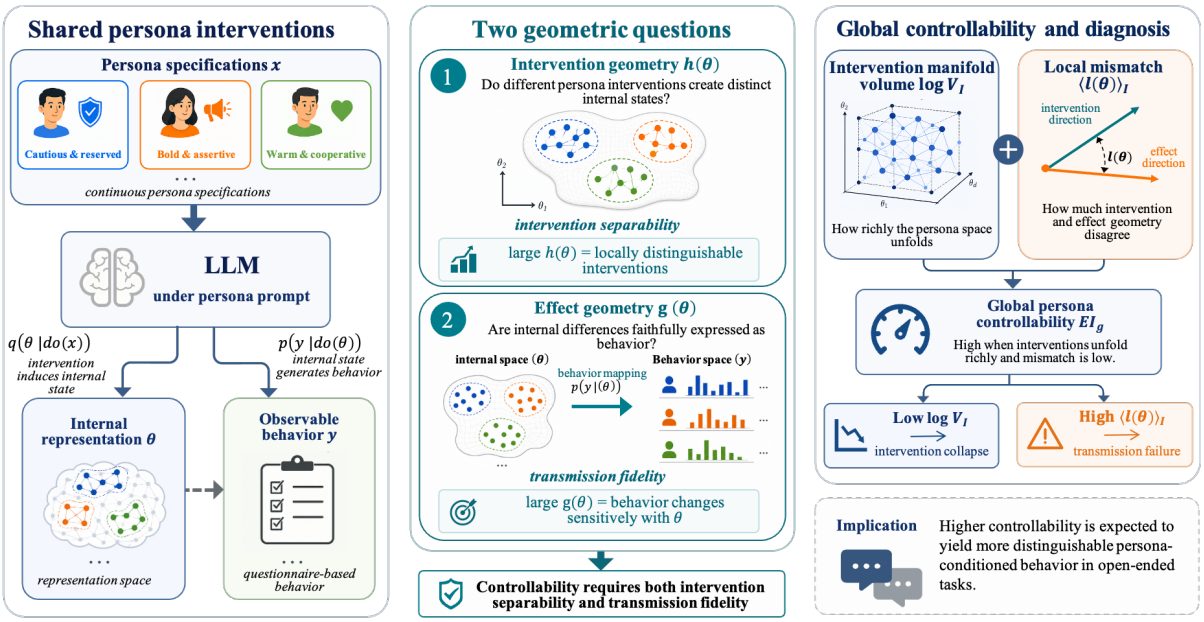


Figure 2: Overview of the proposed representation-level causal-controllability framework. Left: persona specifications are treated as interventions that induce internal representations and observable behaviors, forming a representation-level control chain. Middle: intervention geometry asks whether persona interventions create distinguishable internal states, while effect geometry asks whether these internal differences are faithfully expressed as behavior. Right: global persona controllability combines intervention manifold volume with local mismatch to diagnose intervention collapse and transmission failure. Open-ended behavioral tasks provide behavioral triangulation rather than entering metric construction.

differences, but these differences cannot be reliably reflected in questionnaire scores or open-ended behaviors, then the control chain has failed at the representation-to-effect stage.

Together, these two requirements define what it means for persona specifications to exert reliable control over model behavior. Intervention separability asks whether the input intervention is preserved as a distinguishable internal structure, while transmission fidelity asks whether this structure is carried forward into observable behavior. The remaining question is how to quantify these requirements and integrate them into a global metric capturing both local failure modes and overall controllability, which we address in Section 3.

3 Geometric Metrics for Controllability

We now instantiate the controllability requirements from Section 2 as a geometric framework inspired by information geometry (Amari, 2016). As summarized in Figure 2, the framework comprises two local geometries and one global metric along the representation-level control chain. *Intervention geometry* measures whether persona interventions unfold into distinguishable internal states; *effect geometry* measures whether these differences faithfully transmit to observable behavior; and the global metric integrates both aspects to diagnose intervention collapse, transmission failure, and overall controllability.

We next formalize these components in the same order as the control chain: intervention geometry, effect geometry, and global persona controllability.

3.1 Intervention Geometry: Intervention Separability from x to θ

We first characterize whether persona specifications remain distinguishable after being mapped into the model's internal representation space. Intuitively, a model has high intervention separability if different persona specifications unfold into different regions of Θ; it has low separability if distinct persona specifications collapse into the same internal neighborhood.

To define this notion locally in the internal representation space, we ask which persona interventions could have produced a neighborhood around a given internal state θ. This requires inverting the intervention-to-representation mapping $q(\theta \mid do(x))$ into an operational posterior over interventions. Following the maximum-entropy intervention distribution that defines effective infor-

mation (Hoel, 2017), we take the intervention prior over $\mathcal{X}$ to be uniform.

Definition 1 (Intervention Separability)**.** *Under the above assumptions, we define the local intervention posterior by Bayesian inversion:*

$$\tilde{q}(do(x) \mid \theta) \equiv \frac{q(\theta \mid do(x))}{\int_{\mathcal{X}} q(\theta \mid do(x'))\, dx'}. \quad (3)$$

This posterior induces the ***intervention manifold****:*

$$\mathcal{M}_I \equiv \{\tilde{q}(do(x) \mid \theta)\}_{\theta \in \Theta}, \quad (4)$$

which we equip with the Fisher information metric:

$$h_{\mu\nu}(\theta) = -\mathbb{E}_{\tilde{q}(do(x)|\theta)} \left[\partial_\mu \partial_\nu \log \tilde{q}(do(x) \mid \theta)\right]. \quad (5)$$

Here, $h(\theta)$ measures the local distinguishability of persona interventions around θ: large values indicate directions where small representation changes correspond to distinguishable interventions, whereas small values indicate intervention collapse. Thus, $h(\theta)$ serves as a metric for intervention separability.

This construction makes the internal representation space intervention-aware: it measures whether nearby internal states correspond to distinguishable persona interventions, rather than merely measuring their Euclidean separation in hidden-state space. In this sense, $h(\theta)$ captures the degree to which persona specifications are locally preserved after being mapped into Θ. The local Gaussian approximation and the resulting estimator for $h(\theta)$ are deferred to Appendix E.

3.2 Effect Geometry: Transmission Fidelity from θ to y

Intervention separability alone does not guarantee successful persona control: internally distinguishable states are useful only if their differences are faithfully expressed as observable behavior. We therefore characterize transmission fidelity by the local sensitivity of the external observation distribution $p(y \mid \theta)$ to perturbations in the internal persona representation θ.

If an infinitesimal perturbation along a particular representation direction leads to a substantial change in the questionnaire-based personality observation, then that direction is sensitive and effective with respect to external effects. Conversely, if the perturbation induces little change in the external personality observation, then that direction contributes only marginally to the observable effects of persona simulation.

Definition 2 (Transmission Fidelity)**.** *We introduce a local metric $g(\theta)$ on Θ induced by the mapping $\theta \to y$. Let $d\theta$ denote an infinitesimal perturbation at θ. The difference between the external effect distributions corresponding to two neighboring representation states can be locally approximated by the KL divergence:*

$$D_{\mathrm{KL}}(p(y \mid \theta) \,\|\, p(y \mid \theta + d\theta)) \approx \tfrac{1}{2} d\theta^\top g(\theta) d\theta. \quad (6)$$

Equivalently, $g(\theta)$ corresponds to the Fisher information metric of the conditional distribution $p(y \mid \theta)$:

$$g_{\mu\nu}(\theta) = -\mathbb{E}_{p(y|\theta)} \left[\partial_\mu \partial_\nu \log p(y \mid \theta)\right]. \quad (7)$$

Thus, $g(\theta)$ measures how sensitively observable behavior responds to local changes in θ, and serves as a metric for transmission fidelity.

This metric equips Θ with an effect geometry, yielding the **effect manifold**

$$\mathcal{M}_E \equiv \{p(y \mid \theta)\}_{\theta \in \Theta}. \quad (8)$$

Large values of $g(\theta)$ indicate representation directions that are strongly transmitted into observable behavior differences, whereas small values indicate externally insensitive directions. The practical estimation of $g(\theta)$, including the local Gaussian approximation used in implementation, is deferred to Appendix E.

3.3 Global Persona Controllability EI_g

After obtaining intervention geometry and effect geometry, the key question becomes whether the internal structure unfolded by persona interventions can be effectively carried by observable behavior. In other words, we ask whether the intervention geometry and the effect geometry are mutually compatible. If different persona interventions are internally separable and these separable directions can be transmitted into observable behavior with low distortion, then the model should have high persona controllability. Conversely, if interventions collapse in the internal representation space, or if the unfolded intervention directions are mismatched with behavior-effective directions, then the control chain is unreliable.

We therefore define global persona controllability by combining two aspects: the global volume of intervention unfolding and the local mismatch between intervention geometry and effect geometry.

Definition 3 (Local Mismatch Cost)**.** *To characterize this compatibility, we define the local mismatch cost as:*

$$l(\theta) = \frac{1}{2} \log \det\big(I + g(\theta)^{-1} h(\theta)\big). \tag{9}$$

Here, $l(\theta)$ quantifies the degree of mismatch between the intervention geometry and the effect geometry around a given representation point θ. A larger $l(\theta)$ indicates a greater mismatch between the two geometries, whereas a smaller value indicates that persona interventions can pass through internal representations and be transformed into observable behavior in a more structurally consistent manner. The point-wise correspondence between the two manifolds, $\mathcal{M}_E \leftrightarrow \mathcal{M}_I$, enables this local comparison between the two geometries.

However, the mismatch term alone is still insufficient for measuring the model's overall persona control capability. Even if the two geometries are well matched at every local region, the model should not be considered highly controllable if different persona interventions are barely unfolded across the overall space. Therefore, in addition to local mismatch, we also need to measure the effective volume of the intervention geometry across the internal representation space.

Definition 4 (Intervention Manifold Volume)**.** *We define the intervention manifold volume to quantify the extent to which the intervention geometry is unfolded in the internal representation space:*

$$\mathcal{V}_I = \int_\Theta \sqrt{\det h(\theta)}\, d\theta. \tag{10}$$

The intervention manifold volume $\mathcal{V}_I$ reflects the model's capacity to support effective interventions over the persona intervention space. If different persona specifications collapse over large regions of the internal representation space, then $\mathcal{V}_I$ will be small. If they are sufficiently unfolded in the internal representation space, then $\mathcal{V}_I$ will be large.

Definition 5 (Global Persona Controllability)**.** *Combining the local mismatch cost with the intervention manifold volume, we define the global persona controllability metric* EI_g *as*

$$\mathrm{EI}_g = \log \frac{\mathcal{V}_I}{(2\pi e)^{d/2}} - \langle l(\theta) \rangle_I\,, \tag{11}$$

$$\langle l(\theta) \rangle_I = \frac{1}{\mathcal{V}_I} \int_\Theta \sqrt{\det h(\theta)}\, l(\theta)\, d\theta, \tag{12}$$

where $\langle l(\theta) \rangle_I$ denotes the average local mismatch cost under the intervention geometry. In this metric, the first term captures the global unfolding capacity of the persona intervention space, reflecting the overall volume of the intervention geometry in the internal representation space. The second term captures the average structural mismatch between the intervention geometry and the effect geometry. Together, these two terms determine whether the model can support rich and effective persona control across the entire persona intervention space.

Therefore, EI_g provides a quantitative framework for evaluating a model's controllability over the persona intervention space, and serves as the core evaluation metric of this work.

4 Experiments

This section systematically evaluates the proposed causal controllability framework. We first verify EI_g in controlled synthetic persona systems (Section 4.2), then evaluate real LLMs and conduct behavioral triangulation (Section 4.3).

4.1 Experimental Setup

We use HEXACO (Lee and Ashton, 2004; Ashton and Lee, 2007) as the six-dimensional persona intervention space and sample $N = 500$ persona points with Latin Hypercube Sampling (McKay et al., 2000; Stein, 1987). For each point, we instantiate multiple persona prompts to reduce wording bias, obtain questionnaire-based behavior vectors using randomized and paraphrased HEXACO-60 (Ashton and Lee, 2009) items, and extract internal representations from the last-token hidden state of the second-to-last layer (Zou et al., 2023; Li et al., 2023; Turner et al., 2023). Representations are compressed into a low-dimensional personality-mediated space via a supervised bottleneck autoencoder, after which $g(\theta)$, $h(\theta)$, EI_g, $\log \mathcal{V}_I$, and $\langle l(\theta) \rangle_I$ are estimated using local k-NN approximations. Details of prompts, questionnaires, representation compression, and metric estimation are provided in Appendices A and E. This protocol fixes what θ and EI_g refer to. The intervention we execute is at the prompt level: $do(x)$ is actively manipulated, while the mapping from θ to y is estimated observationally within local neighborhoods. Because θ is obtained under this fixed extraction protocol and the normalization of EI_g depends on d, we use EI_g to compare models within a protocol, and report the $(\log \mathcal{V}_I, \langle l(\theta) \rangle_I)$ decomposition

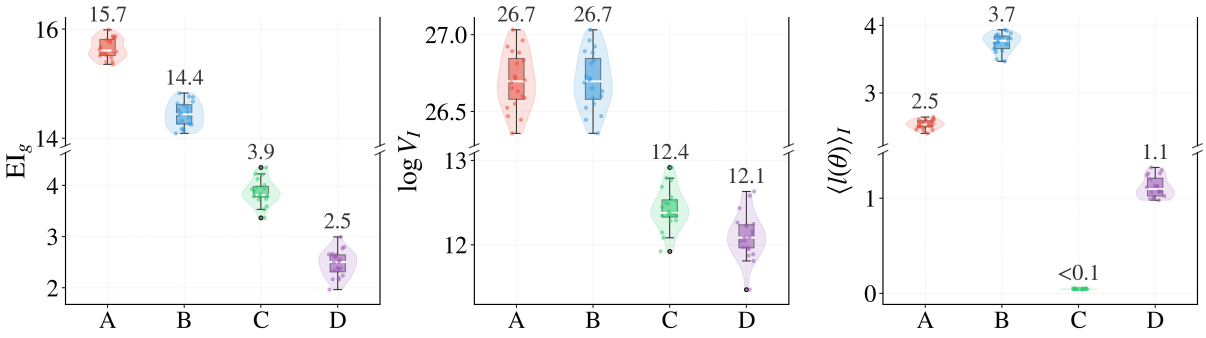


Figure 3: EI_g and its decomposition terms, $\log \mathcal{V}_I$ and $\langle l(\theta)\rangle_I$, for the four synthetic worlds. The ideal system obtains the highest EI_g, whereas the collapsed and noisy systems exhibit substantially lower controllability.

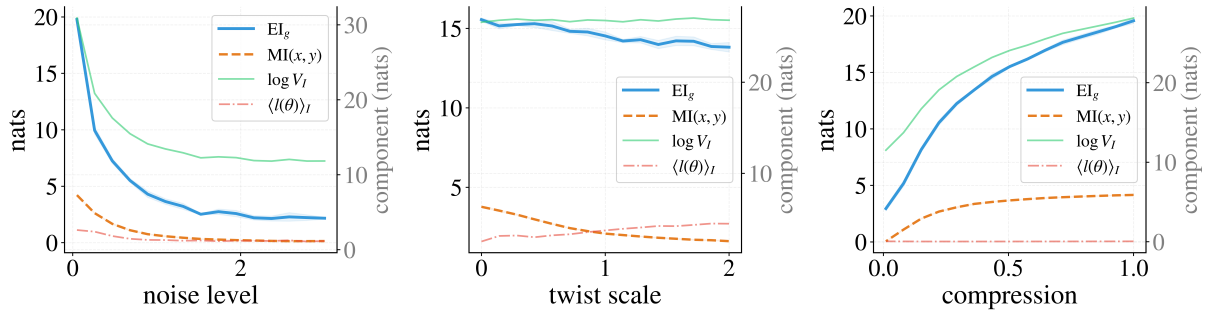


Figure 4: EI_g and decomposition terms along three degradation paths. EI_g decreases monotonically, while components respond differently to distinct failure mechanisms.

alongside it.

4.2 Synthetic Diagnostic Validation

We first validate EI_g in controlled synthetic persona systems. We construct four synthetic worlds with known structural properties: (A) **Ideal control** with strong separability and fidelity; (B) **Transmission-distorted** via a rotated linear $\theta \to y$ mapping with anisotropic observation noise; (C) **Intervention-collapsed** via low-rank $x \to \theta$ projection; (D) **High-noise** in both stages. Full parameterizations are in Appendix B.

Figure 3 shows that EI_g yields the expected ranking (A $>$ B $>$ C $>$ D). Critically, the decomposition terms diagnose distinct failure sources: World B shows elevated mismatch $\langle l(\theta)\rangle_I$ (transmission failure), while World C shows collapsed intervention volume $\log \mathcal{V}_I$ (separability failure). Thus EI_g not only scores controllability but pinpoints *why* it fails.

As detailed in Appendix B, we further evaluate three continuous degradation paths: **noise contamination**, **geometric distortion**, and **diversity collapse**.

Figure 4 shows that EI_g decreases monotonically as degradation intensifies along each path. Components respond distinctly: noise contamination mainly reduces $\log \mathcal{V}_I$ (diversity loss); geometric distortion elevates $\langle l(\theta)\rangle_I$ (fidelity loss); diversity collapse sharply compresses $\log \mathcal{V}_I$. In contrast, $\mathrm{MI}(x, y)$ also decreases under all three paths but cannot distinguish between diversity- and fidelity-driven failures. Appendix C further visualizes the EI_g–MI relationship across all synthetic settings, showing that systems with similar MI can exhibit substantially different controllability due to structural collapse.

4.3 Diagnosis of Real LLMs

We evaluate seven open-source instruction-tuned models spanning 7B–70B dense models and two Mixture-of-Experts (MoE) architectures (total parameters): Qwen2.5-7B-Instruct (Yang et al., 2024), Llama-3.1-8B-Instruct and Llama-3.1-70B-Instruct (Grattafiori et al., 2024), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), DeepSeek-R1-Distill-Qwen-7B (Guo et al., 2025), Mixtral-8x7B-Instruct-v0.1 (Jiang et al., 2024a), and Qwen3-30B-A3B (Yang et al., 2025), using the protocol in Section 4.1. Table 1 reports geometric controllability metrics alongside behavioral triangulation results.

Mistral-v0.3 achieves the highest EI_g, driven by the largest $\log \mathcal{V}_I$, indicating richer persona unfold-

Table 1: Real-model controllability and behavioral triangulation. Higher EI_g, $\log \mathcal{V}_I$, MI, $\mathrm{Sep}_{\mathrm{pair}}$, and Alignment indicate stronger performance; lower $\langle l(\theta) \rangle_I$ indicates lower average transmission mismatch. Bold values indicate the best in each column. $\mathrm{Sep}_{\mathrm{pair}}$ uses the GPT-4o judge.

Model	Type	Scale	$\mathrm{EI}_g \uparrow$	$\log \mathcal{V}_I \uparrow$	$\langle l(\theta) \rangle_I \downarrow$	MI $\uparrow$	$\mathrm{Sep}_{\mathrm{pair}} \uparrow$	Align. $\uparrow$
Mistral-v0.3	Dense	7B	**32.80**	**45.02**	3.70	0.180	**0.974**	3.10
Qwen2.5-7B	Dense	7B	22.93	36.23	4.78	0.349	0.963	3.33
Qwen3-30B-A3B	MoE	30B	21.40	32.83	2.92	0.180	0.933	**4.33**
Llama-3.1-70B	Dense	70B	19.83	32.19	3.86	**0.451**	0.878	3.50
Llama-3.1-8B	Dense	8B	19.58	32.62	4.53	0.346	0.970	3.30
DeepSeek-R1-Distill	Dense	7B	14.27	24.39	1.61	0.148	0.937	2.83
Mixtral-8x7B	MoE	47B	12.16	21.37	**0.69**	0.262	0.807	3.40

Table 2: Bootstrap 95% CIs (B=1500) for the four base models. The intervals summarize uncertainty for EI_g and $\mathrm{Sep}_{\mathrm{pair}}$ under the fixed evaluation protocol.

Model	EI_g	$\mathrm{Sep}_{\mathrm{pair}}$
Mistral-v0.3	$[32.21, 33.37]$	$0.974\ [0.952, 0.993]$
Qwen2.5-7B	$[20.93, 24.03]$	$0.963\ [0.937, 0.985]$
Llama-3.1-8B	$[16.88, 19.99]$	$0.970\ [0.948, 0.989]$
DeepSeek-R1-Distill	$[13.65, 14.78]$	$0.937\ [0.907, 0.963]$

ing in its internal representation space. Among the four base models, DeepSeek-R1-Distill obtains the lowest mismatch but also a small intervention volume, illustrating a critical pitfall: low mismatch is meaningful only when the persona intervention space is sufficiently unfolded. Scale is not monotonic: the 70B dense model is not the most controllable, and the two MoE models yield distinct diagnostic profiles. Mixtral-8x7B has the lowest mismatch overall ($\langle l(\theta) \rangle_I$=0.69) and the lowest intervention volume ($\log \mathcal{V}_I$=21.37), matching the *intervention-collapse* profile documented in Appendix D.5.

For behavioral triangulation, we construct open-ended decision-making scenarios (resource allocation, moral conflict, social interaction) and use $\mathrm{Sep}_{\mathrm{pair}}$, a 2-AFC protocol in which the judge matches an anonymized response to one of two persona profiles, as the main behavioral probe: it requires dimension-level matching between the response and the persona profile, which is the relation EI_g measures at the representation level. We also report **Alignment**, where the judge rates on a 1–5 scale whether a response matches the target persona. Appendix D.4 reports additional separability metrics. Across seven models, EI_g is directionally consistent with $\mathrm{Sep}_{\mathrm{pair}}$ but statistically underpowered, with Spearman ρ=0.607–0.714 as detailed in Appendix D.4. Appendix D reports results from a second independent judge, Gemini-2.5-Pro (Comanici et al., 2025), and a repeated-scoring self-consistency check with composite ICC=0.90.

The main diagnosis is robust across a broad set of implementation choices, although absolute scores and some intermediate rankings vary. We sweep estimator hyperparameters, preprocessing choices, and autoencoder design choices, including hidden-layer selection, bottleneck dimension, supervision strength, initialization, and a persona-level held-out split; Appendices A.5 and A.7 report the results and their exceptions. A permutation check that breaks the persona–representation correspondence collapses EI_g for every model. Appendices D and C provide additional open-ended results, separability metrics, and correlation-based baselines.

5 Related Work

Interpreting internal representations of LLMs. Probing methods decode linguistic and semantic attributes from hidden states, while causal techniques such as activation patching, mediation analysis, and representation editing test whether internal states functionally affect behavior (Alain and Bengio, 2017; Hewitt and Manning, 2019; Belinkov, 2022; Vig et al., 2020; Meng et al., 2022; Li et al., 2023; Turner et al., 2023). These works motivate treating representations as explanatory intermediate variables. Unlike probing work that asks whether attributes are decodable, our framework diagnoses whether an intervention-induced representation is both separable and behaviorally transmitted.

Persona control in LLM agents. Prompt-based persona injection uses role cards, personality descriptions, or trait-specific prompts to induce target behaviors (Wang et al., 2024a; Jiang et al., 2024b), with recent work studying continuous personality control or training-based stabilization (Cho and

Cheong, 2025; Li et al., 2025; Ji et al., 2025). These methods show that LLM outputs can be steered toward desired personas, but they mostly formulate persona control as an input-output problem and provide limited evidence about the internal pathway from persona specification to behavior.

Beyond output-level persona evaluation. Existing evaluations rely on questionnaires, linguistic analysis, or behavioral tasks to measure persona consistency (Jiang et al., 2024b; Serapio-García et al., 2025; Wang et al., 2025; Zheng et al., 2025; Wang et al., 2024b; Samuel et al., 2024). Recent findings reveal gaps between questionnaire self-reports and downstream behavior (Han et al., 2025), but behavior-based evaluations still mainly assess final responses. They can detect inconsistency, but rarely explain whether failure arises from collapsed internal persona representations or poor representation-to-behavior transmission (Chen et al., 2025). Our framework addresses this gap through a representation-level diagnostic chain.

6 Conclusion

This paper reframes LLM persona simulation as a representation-level causal controllability problem. Instead of evaluating whether outputs merely match a target persona, we decompose persona control into a chain from persona specification to internal representation and observable behavior. Based on this view, we introduce intervention separability and transmission fidelity, and combine them into EI_g to diagnose whether persona-control failures arise from collapsed internal representations or poor representation-to-behavior transmission. Experiments on synthetic systems and real LLMs show that this diagnostic perspective reveals failure patterns that output-level consistency and correlation-based metrics can miss, providing a more interpretable basis for evaluating persona-conditioned LLM behavior.

Limitations

This work has several limitations in scope and interpretation. First, our representation-level causal notation is operationalized through local neighborhoods of prompt-induced hidden states rather than direct hidden-state editing. Therefore, the estimated geometries should be interpreted as diagnostic approximations of persona-control pathways, not as complete circuit-level mechanistic causal models, and EI_g is comparable within a fixed protocol. Second, our real-model experiments cover seven instruction-tuned models spanning 7B–70B dense architectures and MoE variants. While this setting reveals model-specific controllability profiles, broader model families, different alignment procedures, and closed-source systems such as ChatGPT and Claude remain outside this scope, because computing EI_g requires access to intermediate hidden activations. Third, our open-ended behavioral triangulation uses automated LLM judges under fixed prompts and anonymized responses; this improves scalability and consistency, but may still introduce judge-specific biases and has not been calibrated against human annotation. Finally, high persona controllability should not be interpreted as evidence that an LLM possesses a human-like personality. It only indicates that, under our protocol, persona specifications induce distinguishable internal representations and behaviorally reflected responses.

Acknowledgments

This work has been supported by National Key R&D Program of China (2024YFC3308200), National Natural Science Foundation of China (No. 62293555, 62402047), and the Fundamental Research Funds for the Central Universities (No. 2233100006).

References

Guillaume Alain and Yoshua Bengio. 2017. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*.

Shun-ichi Amari. 2016. *Information Geometry and Its Applications*, volume 194. Springer.

Michael C Ashton and Kibeom Lee. 2007. Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and social psychology review*, 11(2):150–166.

Michael C Ashton and Kibeom Lee. 2009. The HEXACO–60: A short measure of the major dimensions of personality. *Journal of personality assessment*, 91(4):340–345.

François Bavaud. 2023. Exact first moments of the rv coefficient by invariant orthogonal integration. *Journal of Multivariate Analysis*, 198:105227.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Marcel Binz and Eric Schulz. 2023. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120.

Ivan Butakov, Alexander Tolmachev, Sofia Malanchuk, Anna Neopryatnaya, and Alexey Frolov. 2024. Mutual information estimation via normalizing flows. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Chaoran Chen, Bingsheng Yao, Ruishi Zou, Wenyue Hua, Weimin Lyu, Toby Jia-Jun Li, and Dakuo Wang. 2025. Towards a design guideline for RPA evaluation: A survey of large language model-based role-playing agents. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18229–18268.

Davide Chicco, Matthijs J Warrens, and Giuseppe Jurman. 2021. The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation. *Peerj computer science*, 7:e623.

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

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. 2024. Large language models empowered agent-based modeling and simulation: a survey and perspectives. *Humanities and Social Sciences Communications*, 11(1):1–24.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*.

Thilo Hagendorff, Ishita Dasgupta, Marcel Binz, Stephanie CY Chan, Andrew Lampinen, Jane X Wang, Zeynep Akata, and Eric Schulz. 2023. Machine psychology. *arXiv preprint arXiv:2303.13988*.

Pengrui Han, Rafal Kocielnik, Peiyang Song, Ramit Debnath, Dean Mobbs, Anima Anandkumar, and R Michael Alvarez. 2025. The personality illusion: Revealing dissociation between self-reports behavior in LLMs. *arXiv preprint arXiv:2509.03730*.

Junlin He, Jinxiao Du, Susu Xu, and Wei Ma. 2024. Preventing model collapse in deep canonical correlation analysis by noise regularization. *Advances in neural information processing systems*, 37:82375–82410.

John Hewitt and Christopher D Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4129–4138.

Erik P Hoel. 2017. When the map is better than the territory. *Entropy*, 19(5):188.

Ke Ji, Yixin Lian, Linxu Li, Jingsheng Gao, Weiyuan Li, and Bin Dai. 2025. Enhancing persona consistency for LLMs' role-playing using persona-aware contrastive learning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26221–26238.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024a. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024b. PersonaLLM: Investigating the ability of large language models to express personality traits. In *Findings of the association for computational linguistics: NAACL 2024*, pages 3605–3627.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. ChatGPT for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

Kibeom Lee and Michael C Ashton. 2004. Psychometric properties of the HEXACO personality inventory. *Multivariate behavioral research*, 39(2):329–358.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530.

Wenkai Li, Jiarui Liu, Andy Liu, Xuhui Zhou, Mona Diab, and Maarten Sap. 2025. Big5-chat: Shaping LLM personalities through training on human-grounded data. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20434–20471.

Michael D McKay, Richard J Beckman, and William J Conover. 2000. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 42(1):55–61.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. *Advances in neural information processing systems*, 35:17359–17372.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Joon Sung Park, Carolyn Q Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S Bernstein. 2024. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*, 52.

Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik Narasimhan, and Vishvak Murahari. 2024. PersonaGym: Evaluating persona agents and LLMs. *arXiv preprint arXiv:2407.18416*, 8(9).

Gregory Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2025. A psychometric framework for evaluating and shaping personality traits in large language models. *Nature Machine Intelligence*, pages 1–15.

Michael Stein. 1987. Large sample properties of simulations using latin hypercube sampling. *Technometrics*, 29(2):143–151.

Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D Nguyen. 2025. Multi-agent collaboration mechanisms: A survey of LLMs. *arXiv preprint arXiv:2501.06322*.

Yu-Min Tseng, Yu-Chun Huang, Tsun-Yu Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yijie Meng, and Yun-Nung Chen. 2024. Two tales of persona in llms: A survey of role-playing and personalization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16612–16631.

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

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401.

Noah Wang, Zy Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, et al. 2024a. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777.

Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, et al. 2024b. InCharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 1840–1873.

Yilei Wang, Jiabao Zhao, Deniz S Ones, Liang He, and Xin Xu. 2025. Evaluating the ability of large language models to emulate personality. *Scientific reports*, 15(1):519.

Lixiang Yan, Lele Sha, Linxuan Zhao, Yuheng Li, Roberto Martinez-Maldonado, Guanliang Chen, Xinyu Li, Yueqiao Jin, and Dragan Gašević. 2024. Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, 55(1):90–112.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Jingyao Zheng, Xian Wang, Simo Hosio, Xiaoxian Xu, and Lik-Hang Lee. 2025. LMLPA: Language model linguistic personality assessment. *Computational Linguistics*, 51(2):599–640.

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

A Experimental Setup Details

A.1 HEXACO Personality Model and Persona Interventions

We adopt the HEXACO personality model (Lee and Ashton, 2004; Ashton and Lee, 2007) as the unified parameterization of persona interventions. HEXACO consists of six dimensions, each capturing a distinct personality domain. Table 3 provides the definition of each dimension.

Table 3: HEXACO six-dimension model. Each dimension is defined by its high and low score tendencies.

Dimension	Abbr.	High-score tendency	Low-score tendency
Honesty-Humility	H	Sincere, fair, avoids greed, modest	Manipulative, self-serving, strong desire for status/wealth
Emotionality	E	Fearful, anxious, dependent, sentimental	Calm, independent, low emotional sensitivity
Extraversion	X	Socially confident, proactive, active	Introverted, low expressiveness, low social initiative
Agreeableness	A	Forgiving, gentle, flexible, patient	Critical, stubborn, irritable, low tolerance
Conscientiousness	C	Organized, diligent, perfectionistic, prudent	Unstructured, low planning, impulsive, low responsibility
Openness	O	Aesthetic, curious, creative, unconventional	Conservative, low exploration, low artistic/abstract interest

Unlike prior work that often relies on discrete personality labels (e.g., "extraverted" or "cautious") (Jiang et al., 2024b; Wang et al., 2024b; Serapio-García et al., 2025), we use continuous personality specifications to examine whether observable behavior changes in a stable, distinguishable, and consistent manner when persona norms vary continuously. The reasons for adopting continuous specifications are threefold: (1) the goal of this work is to measure the model's *local* controllability under fine-grained persona interventions, which discrete labels cannot capture; (2) the geometric quantities g, h, J_f, J_F require local neighborhoods with continuous coordinates; and (3) continuous specifications allow constructing near-boundary persona points to test whether the model can express subtle changes at personality boundaries.

To cover the persona intervention space uniformly under a limited sample budget, we use Latin Hypercube Sampling (LHS) (McKay et al., 2000; Stein, 1987) to sample $N = 500$ persona specification points from $\mathcal{X} = [1, 5]^6$. Each dimension ranges from 1 (extremely low) to 5 (extremely high), with values rounded to two decimal places. The sampled points are stored as a structured JSON dataset and released with the code.

A.2 Persona Prompt Construction

For each persona intervention point, we construct corresponding persona prompts to activate persona-conditioned behavioral patterns. To reduce the influence of prompt wording, we adopt a multi-template strategy: while keeping the underlying persona specification fixed, we perturb the description format, linguistic style, and contextual expression of the prompt. Since all models are evaluated using the same prompt protocol, template variation does not introduce systematic unfairness. Below we provide the complete prompt templates used in our experiments.

Persona prompt generation template. Each LHS-sampled HEXACO six-dimensional continuous personality vector is first fed into a prompt generation model (GPT-4o) to produce three distinct natural language persona prompts. The system message is:

```
You are an experienced prompt engineer, skilled at creating persona specification prompts for large models.
```

The user message template is:

```
Based on the following HEXACO six-dimensional personality scores, output three distinct persona specification prompts in English. Requirements: 1) Each prompt must not exceed 60 words; 2) Diverse styles (e.g., formal, colloquial, scenario-based); 3) Clearly articulate high/low extreme features with specific observable behaviors; 4) Number them as #1 #2 #3; 5) Output only the three prompts, no additional explanation. Scores: Honesty-Humility(H): {H}, Emotionality(E): {E}, Extraversion(X): {X}, Agreeableness(A): {A}, Conscientiousness(C): {C}, Openness(O): {O}
```

Each persona generates **3** natural language prompts, corresponding to variations in description format, linguistic style, and contextual expression. This yields $500 \times 3 = 1500$ prompt-level samples per model. The generation uses `temperature = 0.8`, `top_p = 0.9`, and `max_new_tokens = 128`.

Questionnaire persona injection template. During the questionnaire phase, the persona and questionnaire items are injected separately. The system message contains the full persona specification and response rules:

```
Here is your complete persona specification, please strictly follow it. Persona: {natural-language persona prompt} Responses must reflect the persona specification; do not answer mechanically. When answering any questionnaire, output only a single digit 1-5, without symbols, spaces, or explanations. Scale meaning: 1=strongly disagree, 2=disagree, 3=neutral, 4=agree, 5=strongly agree.
```

Each questionnaire item is provided as a separate user message:

```
{item text} Please judge your level of agreement based on the persona specification, and strictly output only one digit from 1, 2, 3, 4, or 5.
```

Note that the persona specification in the questionnaire phase is a natural language prompt, not structured HEXACO numbers. The structured scores are used only for persona prompt generation and autoencoder supervision.

Open-ended task prompts. Open-ended tasks use two prompt types. **Prompt A** (structured HEXACO) provides the six-dimensional scores with scale descriptions. **Prompt B** (natural language portrait) is generated by GPT-4o from the same HEXACO scores. All evaluated models share the same Prompt B to ensure fair cross-model comparison. The open-ended task user message template is:

```
{scenario} Please respond strictly in the following two-part format, neither part may be empty: Decision: (One or two sentences giving your choice/approach directly) Rationale: (Explain why based on your personality traits, 3–6 sentences)
```

A.3 Quantifying Observable Behavior via HEXACO-60

We use the HEXACO-60 personality questionnaire (Ashton and Lee, 2009) as a standardized observation protocol. The questionnaire consists of 60 items, each rated on a 5-point Likert scale:

1 = Strongly Disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, 5 = Strongly Agree.

Scoring rules (Lee and Ashton, 2004; Ashton and Lee, 2009):

- Each of the six dimensions comprises 10 items grouped into four facets.
- Selected items are reverse-scored using the rule $\text{score} = 6 - \text{response}$.
- The dimension score is computed as the mean of its constituent item scores (facets to 2 decimal places; dimension to 1 decimal place).
- No missing responses are allowed.

Each dimension's facet structure is as follows: H (Sincerity, Fairness, Greed-Avoidance, Modesty), E (Fearfulness, Anxiety, Dependence, Sentimentality), X (Social Self-Esteem, Social Boldness, Sociability, Liveliness), A (Forgiveness, Gentleness, Flexibility, Patience), C (Organization, Diligence, Perfectionism, Prudence), O (Aesthetic Appreciation, Inquisitiveness, Creativity, Unconventionality).

The questionnaire is administered in English. To reduce the risk that models exploit memorized questionnaire patterns, we rewrite all 60 items with paraphrased wording and randomize their presentation order for each sample. For each prompt-level sample, the model completes all 60 items, producing a six-dimensional observation vector $y \in \mathcal{Y}$. Scoring follows the standard HEXACO-60 protocol described above. We emphasize that the questionnaire serves as a stable, reproducible observation interface rather than the evaluation target itself.

A.4 Internal Representation Extraction and Autoencoder

Input representation. For each questionnaire-response sample, we extract the last-token hidden state from the second-to-last layer as the raw representation h, i.e., $h = \text{hidden_states}[-2][\text{last_token}]$. The input dimensions vary by model architecture; exact identifiers and hidden-state dimensions are listed in Table 19.

We select the second-to-last layer rather than the final layer for the following reason: the final layer is more directly coupled with next-token logits and output vocabulary distribution, while the second-to-last layer typically preserves more stable semantic and state representations that are better suited for capturing persona-related internal structure (Zou et al., 2023; Li et al., 2023; Turner et al., 2023). Each model is equipped with its own autoencoder, since the hidden-state space and dimension depend on the base model.

Autoencoder architecture. The supervised bottleneck autoencoder comprises an encoder, a decoder, and a regression head:

Encoder:

$$h \xrightarrow{\text{LN}} \text{Linear}(\text{in}, 1024) \xrightarrow{+\text{ReLU}} \text{Linear}(1024, 256) \xrightarrow{+\text{ReLU}} \text{Linear}(256, d) \xrightarrow{\text{LN}} z$$

Decoder:

$$z \to \text{Linear}(d,\ 256) \xrightarrow{+\text{ReLU}} \text{Linear}(256,\ 1024) \xrightarrow{+\text{ReLU}} \text{Linear}(1024,\ \text{in}) \to \hat{h}$$

Regression head:

$$z \to \text{Linear}(d,\ 64){+}\text{ReLU} \to \text{Linear}(64,\ 6) \to \hat{x}$$

The bottleneck dimension is $d = 6$, matching the HEXACO personality dimensionality. This choice ensures that the compressed space aligns with the theoretical intervention dimensions and avoids introducing uninterpretable latent directions.

Training. The training loss combines reconstruction and supervision:

$$\mathcal{L} = \text{MSE}(\hat{h}, h) + \lambda\,\text{MSE}(\hat{x}, x), \qquad \lambda = 3.0, \tag{13}$$

where h is the raw hidden-state embedding, $\hat{h}$ is the decoder reconstruction, x is the LHS-sampled HEXACO 6-dimensional target score, and $\hat{x} = \text{reg_head}(z)$ is the predicted HEXACO score from the bottleneck. The full training configuration is given in Table 4.

Table 4: Autoencoder training configuration.

Parameter	Value
Optimizer	AdamW
Learning rate	1×10^{-3}
Batch size	32
Epochs	20
Validation split	0.1
Weight decay	1×10^{-4}
LR scheduler	ReduceLROnPlateau
LR decay factor	0.5
LR patience	3
Gradient clipping	1.0
Early stopping patience	5
Dropout	0.0
Seed	42

The 1500 prompt-level samples are randomly split into 90% training and 10% validation. The default pipeline does not reserve an independent test split for autoencoder fitting; downstream evaluation proceeds via the questionnaire and open-ended tasks. Separately, we additionally evaluate a persona-level 80/20 held-out protocol in which the autoencoder is fit on training personas and EI_g is computed on unseen personas, as reported in Appendix A.5. The autoencoder is supervised to reconstruct the persona specification x from hidden states; its bottleneck representation is used as the operational internal representation θ in the EI_g computation.

A.5 Autoencoder Design Robustness

A natural concern is that EI_g might merely rediscover supervised HEXACO probing accuracy. Held-out supervised prediction loss is nearly identical across models (1.29–1.30), while EI_g spans 14.27–32.80. This suggests that EI_g is not reducible to held-out HEXACO prediction error alone.

Supervision strength. Table 5 sweeps $\lambda \in \{0, 0.5, 1, 3\}$ (default λ=3). At λ=0, Mistral, Llama, and DeepSeek still produce informative EI_g while Qwen is numerically singular; removing supervision changes the ordering—DeepSeek is no longer lowest at λ=0—whereas across all nonzero λ DeepSeek consistently occupies the lowest position.

Table 5: EI_g under supervision-strength ablation. "—" denotes numerical singularity.

Model	λ=0	λ=0.5	λ=1	λ=3 (default)
Mistral-v0.3	13.49	26.57	31.39	**32.80**
Qwen2.5-7B	—	—	19.41	**22.93**
Llama-3.1-8B	20.13	24.14	19.83	19.58
DeepSeek-R1-Distill	17.51	14.49	16.29	14.27

Bottleneck dimension. Table 6 sweeps $d \in \{3, 6, 12, 24, 48\}$ (default d=6). Ranking is consistent for d=6–48; d=3 inflates absolute values and makes Qwen singular. For the three extended models, d=24/48 cells are left blank where the corresponding runs were not completed.

Table 6: EI_g under bottleneck-dimension sweep.

Model	d=3	d=6	d=12	d=24	d=48
Mistral-v0.3	64.09	**32.80**	25.15	25.20	18.65
Qwen2.5-7B	—	**22.93**	22.15	23.51	17.08
Llama-3.1-8B	59.40	19.58	21.35	17.78	14.21
DeepSeek-R1-Distill	61.72	14.27	10.36	10.45	11.49

Hidden-layer selection. We repeat representation extraction and autoencoder fitting using the last, penultimate (default), third-last, and mean of the last four hidden layers. Table 7 shows that layer choice changes absolute scores and the precise ordering among intermediate models, while DeepSeek remains lowest under all four settings. We use the penultimate layer as a pre-specified default to reduce direct coupling with the output projection and next-token distribution.

Reconstruction quality. Validation reconstruction MSE differs sharply across models (Mistral

Table 7: EI_g under hidden-layer selection (four base models).

Model	Last	Penult. (default)	Third-last	Mean last four
Mistral-v0.3	17.85	**32.80**	**33.39**	23.64
Qwen2.5-7B	18.30	**22.93**	22.12	24.61
Llama-3.1-8B	**18.50**	19.58	25.89	**26.71**
DeepSeek-R1-Distill	11.60	14.27	12.77	14.43

0.13, Llama 0.37, Qwen 7.47, DeepSeek 100.17) while Val Sup remains 1.29–1.30. Within the same hidden dimension, the Mistral–Llama EI_g gap ($1.68\times$) is comparable to the Qwen–DeepSeek gap ($1.61\times$), so hidden-state dimensionality alone does not explain the ranking, though interactions with architecture and training objective cannot be ruled out.

Seed stability. Table 8 reports 20 random AE initializations. Cross-seed means preserve Mistral $>$ Qwen $>$ Llama $>$ DeepSeek; every Table 1 point estimate falls inside its mean$\pm\sigma$ interval; DeepSeek has the smallest σ.

Table 8: AE initialization stability across 20 seeds (four base models).

Model	Table 1	20-seed mean$\pm\sigma$
Mistral-v0.3	32.80	31.47 ± 2.83
Qwen2.5-7B	22.93	23.18 ± 3.06
Llama-3.1-8B	19.58	20.34 ± 2.41
DeepSeek-R1-Distill	14.27	14.51 ± 1.08

Persona-level held-out. We additionally run 10 persona-level 80/20 splits (400 training personas fit the AE; EI_g is computed on 100 unseen personas). Held-out EI_g means are 31.06 ± 1.99 / 22.11 ± 1.54 / 20.04 ± 1.06 / 14.76 ± 0.73 for Mistral / Qwen / Llama / DeepSeek; 8/10 splits fully reproduce the ranking, with Mistral always highest and DeepSeek always lowest. Held-out HEXACO supervised loss is 1.33–1.37.

A.6 Geometric Metric Estimation

Given triples $\{(x_i, \theta_i, y_i)\}_{i=1}^N$, we estimate local geometric structures using k-nearest-neighbor local linear approximation. The full procedure is detailed in Appendix E; here we summarize the key estimation parameters in Table 9.

The volume correction factor vol_θ is computed as the product of per-dimension ranges in the original θ space, compensating for the volume scaling introduced by standardization and PCA whitening. This is the default *range-product* convention used throughout the main results; vol_x is used only by the alternative volume strategy A.

Table 9: Geometric metric estimation parameters.

Parameter	Value
k (kNN size)	20
Covariance type	Shrinkage
Shrinkage α	0.1
Covariance regularization	1×10^{-5}
Metric regularization ϵ	1×10^{-5}
Numerical stability η	1×10^{-8}
Volume correction	Range-product
Weight type	Gaussian
Weight σ	Local median
Winsorization	1%

Preprocessing. (1) x is z-scored per dimension. (2) θ is z-scored per dimension and then PCA-whitened. (3) y is z-scored per dimension. (4) For the published range-product strategy B, `vol_theta` is computed from the original-space range product to compensate for volume scaling introduced by standardization and whitening.

Local Jacobian estimation. For each θ_i in the whitened θ space, $k = 20$ nearest neighbors are identified. Weighted local linear regressions are fitted:

$$y \approx A_y\theta + b_y, \qquad x \approx A_x\theta + b_x, \tag{14}$$

using `sklearn.linear_model.LinearRegression` with Gaussian sample weights $w_j = \exp\big(-\|\theta_j - \theta_i\|^2/(2\sigma_i^2)\big)$, where σ_i is the median of non-zero distances in the local neighborhood. The coefficient matrices A_y and A_x serve as numerical estimates of $J_f(\theta_i)$ and $J_F(\theta_i)$, respectively.

Residual covariance. Residuals $r_y = y-(A_y\theta+b_y)$ and $r_x = x - (A_x\theta + b_x)$ are computed. Covariance matrices are estimated with shrinkage:

$$\Sigma = \alpha\,\mathrm{diag}(\mathrm{var}(r)) + (1-\alpha)\,\mathrm{cov}(r), \quad \alpha = 0.1, \tag{15}$$

with additional regularization $\Sigma + \texttt{cov_reg}\cdot I$. The inverses are $\Sigma_y^{-1} = (\Sigma_y + \epsilon I)^{-1}$ and $\Sigma_x^{-1} = (\Sigma_x + \epsilon I)^{-1}$.

Aggregation to EI_g. Local metrics are computed as $g_i = A_{y,i}^\top \Sigma_{y,i}^{-1} A_{y,i}$ and $h_i = A_{x,i}^\top \Sigma_{x,i}^{-1} A_{x,i}$. The mismatch cost is obtained from the eigenvalues λ_k of the symmetric matrix $(g_i + \epsilon I)^{-1/2} h_i (g_i + \epsilon I)^{-1/2}$:

$$l_i = \frac{1}{2}\sum_k \log(1 + \lambda_k). \tag{16}$$

Local weights are $w_i = \sqrt{\max(\det h_i, 0)}$. The weighted averages are:

$$\langle l \rangle_I = \frac{\sum_i w_i l_i}{\sum_i w_i}, \qquad \mathcal{V}_I = \frac{\mathrm{vol}_\theta}{N} \sum_i w_i. \quad (17)$$

Finally, the global persona controllability metric is:

$$\mathrm{EI}_g = \log \frac{\mathcal{V}_I}{(2\pi e)^{d/2}} - \langle l \rangle_I. \quad (18)$$

A.7 Estimator Sensitivity

We sweep estimator and preprocessing choices one axis at a time on the four base models. Absolute EI_g moves, but the endpoint diagnosis (Mistral highest, DeepSeek lowest) is preserved except where noted. Volume conventions are named *range-product (default)* / *no-corr* / *Jacobian-corrected* / *robust*; the main text and Table 9 use range-product (default).

k-NN size. At $k \in \{15, 20, 30, 40, 60\}$ the ranking is identical; at k=10, Qwen and Llama swap while Mistral/DeepSeek endpoints remain.

Table 10: k-NN neighborhood sweep of EI_g.

Model	k=10	k=15	k=20	k=30	k=40	k=60
Mistral-v0.3	**42.88**	**35.96**	**32.80**	**29.99**	**28.12**	**25.41**
Qwen2.5-7B	29.94	25.67	22.93	18.88	14.91	13.58
Llama-3.1-8B	30.84	22.62	19.58	15.05	14.02	10.93
DeepSeek-R1-Distill	23.06	17.56	14.27	10.96	7.85	3.36

Preprocessing. DeepSeek is lowest under every computable setting; Qwen is singular under z-score-only.

Table 11: Preprocessing ablation of EI_g. "—" denotes singularity.

Model	z+whiten	z-score	min-max	raw
Mistral-v0.3	**32.80**	19.08	31.04	**38.35**
Qwen2.5-7B	22.93	—	**32.47**	15.36
Llama-3.1-8B	19.58	7.50	17.05	18.00
DeepSeek-R1-Distill	14.27	2.89	13.33	12.95

Sample size N. Downsampling persona IDs at $N \in \{100, \ldots, 500\}$, as shown in Table 12, DeepSeek ranks fourth in 100% of replicates once $N \geq 300$.

Volume convention. Spearman ρ vs. default ranking is 1.00/0.80/1.00/1.00 for range-product / no-corr / Jacobian / robust; only no-corr swaps Qwen and Llama, while endpoints are unchanged.

Table 12: Persona-level N sensitivity (mean$\pm\sigma$ over 10 subsets for N<500).

N	Mistral	Qwen	Llama	DeepSeek
100	30.74 ± 4.38	21.94 ± 3.35	21.09 ± 3.10	15.36 ± 2.25
200	31.45 ± 2.90	22.24 ± 2.40	20.41 ± 2.21	14.80 ± 1.38
300	31.83 ± 1.97	22.43 ± 1.81	20.21 ± 1.53	14.70 ± 0.82
400	32.13 ± 1.20	22.46 ± 1.14	20.04 ± 0.89	14.57 ± 0.55
500	32.80	22.93	19.58	14.27

Table 13: Volume-correction conventions at N=500.

Convention	Mistral	Qwen	Llama	DeepSeek	ρ
Range-product (default)	32.80	22.93	19.58	14.27	1.00
No-corr	27.45	18.92	19.11	10.84	0.80
Jacobian-corrected	32.06	22.48	19.34	14.02	1.00
Robust quantile	31.88	22.57	19.29	13.96	1.00

Regularization. Across $\epsilon \in [10^{-7}, 10^{-3}]$, all pairwise ranking correlations satisfy $\rho \geq 0.964$; across covariance types full/diag/shrinkage, $\rho \geq 0.929$.

Scale calibration. Synthetic worlds use unit-scale $\mathcal{N}(0, I)$ coordinates, whereas real-model θ is whitened with volume factors on the order of 10^5–10^{11}. Because $\log(\mathrm{vol})$ enters EI_g additively, absolute scores are not comparable across synthetic vs. real settings; within a fixed protocol they remain comparable, as shown in Table 14.

Table 14: Illustrative scale calibration: absolute EI_g is protocol-conditional.

Setting	Typical EI_g	Notes
Synthetic ideal (World A)	≈ 15.7	Unit-scale coordinates
Real LLMs (default protocol)	14–33	Whitened θ, large vol factor

A.8 Permutation Test

Shuffling the persona–representation correspondence collapses EI_g for every model, as shown in Table 15, confirming that the metric relies on non-random alignment between x and θ.

B Synthetic Experiment Parameterizations

B.1 Synthetic World Construction

We construct four synthetic worlds sharing the causal chain $x \to \theta \to y$. All worlds use $N = 500$ sampled points with $d_x = d_\theta = d_y = 6$. The generic data-generating process is:

$$\theta = F(x) + \varepsilon_\theta, \qquad y = G(\theta) + \varepsilon_y, \quad (19)$$

where $x \sim \mathcal{N}(0, I_6)$.

Table 15: Permutation test of EI_g (seven models).

Model	Original	Permuted	z	p
Mistral-v0.3	32.80	1.04 ± 0.07	439.0	< 0.01
Llama-3.1-8B	19.58	3.35 ± 0.07	236.4	< 0.01
Qwen2.5-7B	22.93	2.89 ± 0.10	205.3	< 0.01
DeepSeek-R1-Distill	14.27	3.36 ± 0.08	133.5	< 0.01
Qwen3-30B-A3B	21.40	2.98 ± 0.20	93.8	< 0.01
Llama-3.1-70B	19.83	2.67 ± 0.19	91.2	< 0.01
Mixtral-8x7B	12.16	3.41 ± 0.17	50.8	< 0.01

World A (Ideal control). Strong separability and high fidelity:

$$\theta = x + 0.1\,\varepsilon_\theta, \qquad y = \theta + 0.1\,\varepsilon_y. \tag{20}$$

World B (Transmission-distorted). Preserved separability but impaired fidelity. The $\theta \to y$ mapping applies an orthogonal rotation R_3 to the last three dimensions and adds anisotropic noise:

$$\theta = x + 0.1\,\varepsilon_\theta, \tag{21}$$

$$G = \begin{bmatrix} I_3 & 0 \\ 0 & R_3 \end{bmatrix}, \tag{22}$$

$$\sigma_y = (0.1,\ 0.1,\ 0.1,\ 1.0,\ 1.0,\ 1.0). \tag{23}$$

World C (Intervention-collapsed). Preserved fidelity but impaired separability. The $x \to \theta$ mapping applies a strong compression:

$$F = 0.05\,I, \qquad \theta = 0.05\,x + 0.08\,\varepsilon_\theta,$$
$$y = \theta + 0.005\,\varepsilon_y. \tag{24}$$

World D (High-noise). Degraded separability and fidelity:

$$\theta = x + 2.0\,\varepsilon_\theta, \qquad y = \theta + 2.0\,\varepsilon_y. \tag{25}$$

Table 16 reports the complete metric results for the four synthetic worlds.

B.2 Continuous Degradation Paths

For the robustness analysis, we use $N = 500$ points and the following sweep parameters:

- 20 random seeds per world for stability assessment;
- 15 sweep points per degradation path;
- 5 replicates per sweep point.

Noise contamination. Increase σ from 0.05 to 3.0 in 15 steps (corresponding to World D's failure mode):

$$\theta = x + \sigma\,\varepsilon_\theta, \qquad y = \theta + \sigma\,\varepsilon_y. \tag{26}$$

Geometric distortion. Sweep distortion strength s from 0.0 to 2.0 in 15 steps (corresponding to World B's failure mode):

$$\theta = x + 0.1\,\varepsilon_\theta, \qquad y = (1-s)\cdot\theta + s\cdot R_3\theta + \varepsilon_y, \tag{27}$$

where R_3 applies an orthogonal rotation to the last three dimensions, with σ_y anisotropic as in World B.

Diversity collapse. Sweep compression factor c from 1.0 to 0.01 in 15 steps (corresponding to World C's failure mode):

$$\theta = c\cdot x + 0.08\,\varepsilon_\theta, \qquad y = \theta + 0.005\,\varepsilon_y. \tag{28}$$

Table 17 reports the mean and standard deviation across 20 random seeds for each world.

C Additional Discriminant Validity Against Correlation Metrics

Figure 5 further examines whether EI_g is reducible to input-output correlation. While EI_g and $MI(x, y)$ are positively correlated at a coarse level, the scatter plot shows substantial dispersion at similar MI values. In particular, diversity-collapse settings can retain non-negligible MI while exhibiting much lower EI_g, indicating that correlation can remain high even when the persona intervention space is structurally compressed. This supports the discriminant validity of EI_g.

We further compute additional input-output coupling metrics across all synthetic settings to complement the $MI(x, y)$ comparison in the main text:

R^2 (multivariate regression). We fit a multi-output linear regression from x to y and report the coefficient of determination, uniformly averaged across output dimensions (Chicco et al., 2021). Lower R^2 indicates weaker linear coupling.

Canonical Correlation Analysis (CCA). We compute all six canonical correlations between x and y (He et al., 2024). We report the first canonical correlation CCA_1 and the mean CCA_m across all six dimension pairs. CCA captures linear subspace alignment between the two variable sets.

RV coefficient. A multivariate generalization of R^2, measuring overall matrix correlation between the two sets of variables (Bavaud, 2023).

Mutual information (MI). We use the KSG estimator with Chebyshev norm and $k = 7$ (Butakov et al., 2024).

Table 18 reports these metrics across the four base models. The results confirm the pattern

Table 16: Complete metric results for the four synthetic worlds.

World	EI_g	$\log \mathcal{V}_I$	$\langle l \rangle_I$	$R^2_{\theta,y}$	$\mathrm{CCA}_{\theta,y}$	$\mathrm{RV}_{\theta,y}$	$\mathrm{MI}_{\theta,y}$	$R^2_{x,y}$	$\mathrm{MI}_{x,y}$
A	15.59	26.64	2.54	0.990	0.995	0.983	4.023	0.980	3.793
B	14.31	26.67	3.85	0.748	0.851	0.744	2.195	0.741	2.095
C	3.71	12.27	0.04	0.997	0.999	0.988	4.281	0.279	0.501
D	2.56	12.06	0.99	0.568	0.751	0.567	1.185	0.116	0.167

Table 17: Stability analysis across 20 random seeds.

World	EI_g	$\log \mathcal{V}_I$	$\langle l \rangle_I$	$\mathrm{MI}_{x,y}$
A	15.650 ± 0.180	26.700 ± 0.180	2.536 ± 0.063	3.774 ± 0.020
B	14.437 ± 0.221	26.700 ± 0.180	3.749 ± 0.128	2.116 ± 0.033
C	3.868 ± 0.234	12.427 ± 0.236	0.046 ± 0.004	0.530 ± 0.033
D	2.473 ± 0.248	12.104 ± 0.261	1.117 ± 0.108	0.218 ± 0.023

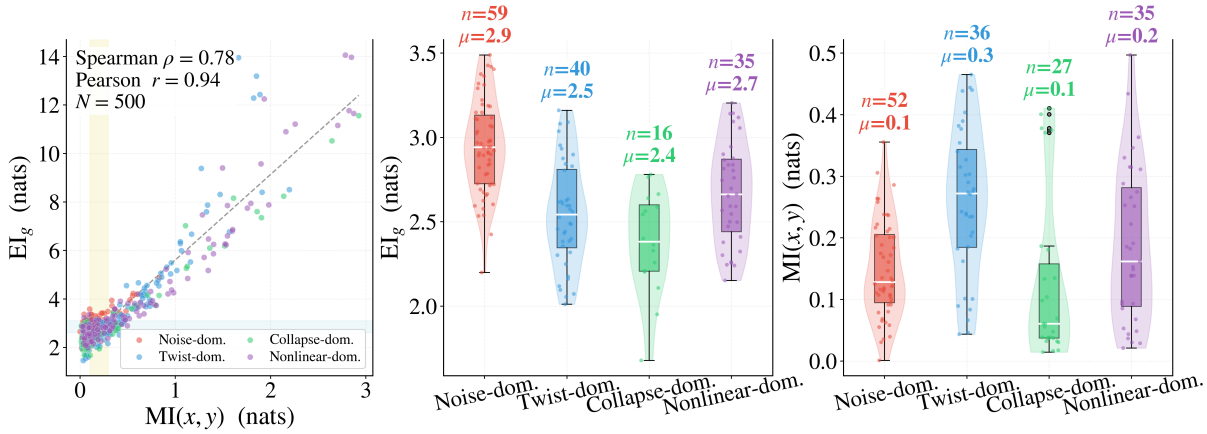


Figure 5: Discriminant validity of EI_g against $MI(x, y)$ across all synthetic settings. Although EI_g and MI are positively correlated overall, settings with similar MI can exhibit substantially different EI_g values, especially under diversity collapse. This shows that EI_g captures structural controllability rather than mere input-output dependence.

observed with $MI(x, y)$ in the main text: all correlation-based metrics favor Qwen2.5-7B and Llama-3.1-8B, while EI_g identifies Mistral-v0.3 as the most controllable model.

In the synthetic experiments, the same discriminant-validity pattern holds: R^2, CCA, and RV all fail to distinguish World C (intervention-collapsed) from World A (ideal control), because the $\theta \to y$ mapping remains near-identity in both worlds. In contrast, EI_g drops sharply for World C due to collapsed intervention volume, correctly identifying the structural failure.

D Open-Ended Task Details

D.1 Model Inference Configuration

Table 19 provides the exact model identifiers and hidden-state dimensions. Table 20 summarizes the inference settings for each experimental phase.

The total inference volume across the four base models is summarized in Table 21. Local inference for these 7B–8B models was performed on a single NVIDIA GeForce RTX 4090 (24 GB); the three larger models were run on two NVIDIA A100 GPUs (80 GB each), as detailed in Appendix F.5.

D.2 Task Design and Persona Selection

We construct 9 open-ended decision-making scenarios, with 3 per category. All tasks are presented in English.

Resource allocation / cooperation-competition.

- **A1 (Bonus allocation):** You and another member completed a task together with a bonus of 100 yuan. You did about 70% of the work, the other 30%. How would you allocate the bonus? Explain your reasoning.

- **A2 (Shared resource):** Your group has only one high-performance device. You urgently need it for your part, but another member also

Table 18: Correlation-based baseline metrics across real LLMs.

Model	$\mathrm{MI}_{x,y}$	$R^2_{x\to y}$	$R^2_{\theta\to y}$	CCA_1	CCA_m	RV
Qwen2.5-7B	0.349	0.128	0.015	0.495	0.313	0.112
DeepSeek-R1-Distill	0.148	0.003	0.004	0.091	0.043	0.003
Mistral-v0.3	0.180	0.019	0.014	0.204	0.081	0.015
Llama-3.1-8B	0.346	0.119	0.012	0.529	0.259	0.103

Table 19: Model identifiers and hidden-state dimensions.

Alias	HuggingFace identifier	Hidden dim
Qwen2.5-7B	`Qwen/Qwen2.5-7B-Instruct`	3584
DeepSeek-R1-Distill	`deepseek-ai/DeepSeek-R1-Distill-Qwen-7B`	3584
Mistral-v0.3	`mistralai/Mistral-7B-Instruct-v0.3`	4096
Llama-3.1-8B	`unsloth/Meta-Llama-3.1-8B-Instruct`	4096
Llama-3.1-70B	`meta-llama/Meta-Llama-3.1-70B-Instruct`	8192
Mixtral-8x7B	`mistralai/Mixtral-8x7B-Instruct-v0.1`	4096
Qwen3-30B-A3B	`Qwen/Qwen3-30B-A3B`	2048

Table 20: Inference configurations per experimental phase.

Phase	Other	Temp.	Top-p	Max tok.
Persona prompt gen.	one API call returning n=3 prompts	0.8	0.9	128
Questionnaire	3 samples, mean aggregation	0.9	0.93	3
Open-ended task	2 seeds	0.7	0.9	600
GPT-4o judging	`seed=42`	0.0	—	200/300

Table 21: Total inference volume across the four base models. Questionnaire answer counts are after mean aggregation over 3 samples per item.

Phase	Total (4 base models)
Persona prompt embeddings / AE	6000 embeddings
HEXACO-60 questionnaire answers	1,080,000 raw answers; 360,000 post-aggregation observations
Open-ended task generation	864 responses
GPT-4o RAS/BSS/RCS judging	864 judgments
GPT-4o 6-AFC identification	432 judgments

has an urgent request. How do you arrange it? Give your decision and rationale.

- **A3 (Free-rider scenario):** You notice a team member contributes little but will share equally in the final reward. How do you handle this? Give your approach and reasoning.

Moral conflict / rule conflict.

- **B1 (Friend's rule violation):** A close friend asks you to help with something that slightly violates regulations but causes no obvious harm in the short term. What do you do? Explain.
- **B2 (Reporting an error):** You find an obvious error in a colleague's report. Pointing it out publicly may affect team relations; not saying anything may create future risks. How do you handle this?
- **B3 (Collective benefit vs. individual fairness):** A proposal would clearly improve overall team benefit but would disadvantage one member. Do you support it? Explain.

Social pressure / social interaction.

- **C1 (Public questioning):** In a meeting, someone directly questions your ability and judgment. How do you respond? Describe your immediate reaction and rationale.
- **C2 (Deadline mistake):** Near a project deadline, you discover a problem in your key section that may affect the entire team's schedule. How do you handle it?
- **C3 (Relationship misunderstanding):** You feel someone who was previously friendly has become noticeably distant, but you are unsure why. How do you handle this?

Persona selection. Six personas are selected from the LHS-sampled pool using three strategies: (A) *Divergent*: points chosen to expose contrasting model-level controllability profiles; (B) *Boundary*: near-boundary points where adjacent personality dimensions meet, testing fine-grained controllability; (C) *Hard*: regions where all models struggle,

ensuring tasks are not trivially easy. The selected persona HEXACO vectors are given in Table 22.

D.3 GPT-4o Judge Setup

D.3.1 RAS/BSS/RCS Judge

The judge model is `GPT-4o` with `temperature = 0.0`, `seed = 42`, and `response_format = {"type": "json_object"}`.

The system prompt is:

```
You are a strict persona role-play evaluator. Given a HEXACO persona specification, an open-ended task, and the tested model's response, independently score the following three dimensions (integer 1-5; 5=excellent, 1=very poor): [RAS] Role Alignment Score - whether the response aligns with the given HEXACO specification in content, wording, and values. Evaluate each dimension: H/E/X/A/C/O must be reasonably reflected; extremes (≤2 or ≥4) and boundary (2.7-3.3) must be identifiable. 5: All six dimensions precisely reflected; 4: Main dimensions correct, 1-2 vague; 3: Direction roughly right but multiple dimensions unreflected or slightly conflicting; 2: Clearly inconsistent or self-contradictory with specification; 1: No persona trace or opposite direction. [BSS] Behavioral Specificity Score - whether the response provides specific, identifiable, personalized details. 5: Highly personalized; 3: Some character but still generic; 1: Template-like, any persona could answer this way. [RCS] Reasoning Coherence Score - whether the decision and rationale are consistent, natural, and non-contradictory. 5: Rigorously coherent; 3: Basically corresponds but with logical jumps; 1: Decision contradicts rationale or rationale is completely empty. Output strictly JSON: {"RAS": 1-5, "BSS": 1-5, "RCS": 1-5, "brief_reason": "within 30 characters"}. No other content.
```

The user prompt is:

```
=== Persona Specification === HEXACO scores: H={H}, E={E}, X={X}, A={A}, C={C}, O={O} Portrait: {natural-language persona description} === Task === {task scenario} === Tested Model Response === {model response} === Please score (JSON only) ===
```

Total RAS/BSS/RCS judgments: 864.

D.3.2 6-AFC Identification Judge

Same `GPT-4o` configuration, with `max_tokens = 300`. The system prompt is:

```
You are a strict persona matching evaluator. Given an open-ended task response and six candidate HEXACO persona portraits (labeled P1-P6), the response comes from exactly one persona. Your task: give each candidate a 0-5 matching score (5=perfect match, 0=completely off; scores may repeat), then select the single most likely best_match. Evaluate each dimension: H/E/X/A/C/O tendencies; extremes (≤2 or ≥4) and boundaries (2.7-3.3) should be considered. Multiple candidates may score high if the response is ambiguous, but give a unique best_match. Output strictly JSON: {"scores": {"P1": 0-5, ..., "P6": 0-5}, "best_match": "P1"|"P2"|...|"P6", "confidence": 1-5, "brief": "within 30 characters"}. No other content.
```

Total 6-AFC judgments for the four base models under GPT-4o: $4 \times 6 \times 9 \times 2 = 432$. Extended cross-model and cross-judge counts are reported with the corresponding appendix tables.

D.4 Additional Separability Metrics

The main text uses $\mathrm{Sep}_{\mathrm{pair}}$ as the primary behavioral probe. Table 23 defines auxiliary metrics derived from the 6-AFC judge output, including $\mathrm{Sep@6}$ (top-1 identification among six candidates).

Table 24 reports GPT-4o results for the four base models. Under the GPT-4o judge, Mistral-v0.3 attains the highest values on all six separability metrics; the cross-judge comparison in Table 25 shows this ordering is largely but not entirely preserved. Qwen2.5-7B, despite having the highest MI and Alignment among the four base models, ranks lower on most GPT-4o separability metrics.

$\mathrm{Sep@6}$ vs. $\mathrm{Sep}_{\mathrm{pair}}$. Under GPT-4o, DeepSeek ($\mathrm{Sep@6}$=0.657) outranks Qwen (0.630), inverting the EI_g order; under Gemini the EI_g–$\mathrm{Sep@6}$ correlation is ρ=0.703 while under GPT-4o it is ρ=−0.321. In contrast, under $\mathrm{Sep}_{\mathrm{pair}}$ both judges place DeepSeek below Qwen. Six-way identification is jointly affected by candidate-set composition, explicit persona cues, and judge calibration; the present experiments cannot further disentangle these factors. We therefore treat $\mathrm{Sep@6}$ as an auxiliary metric.

Cross-judge $\mathrm{Sep}_{\mathrm{pair}}$ and correlations. Table 25 reports $\mathrm{Sep}_{\mathrm{pair}}$ under GPT-4o and Gemini-2.5-Pro for all seven models. Mixtral (lowest EI_g) is lowest under both judges (0.807 / 0.839). Across n=7, Spearman correlations are directionally consistent

Table 22: Six personas used in open-ended tasks.

ID	Category	Rationale	H	E	X	A	C	O
P1	A divergent	High model disagreement	1.07	2.78	4.57	2.45	3.64	2.81
P2	A divergent	High model disagreement, neutral/boundary	2.73	2.68	3.03	3.55	3.80	2.92
P3	B boundary	Near Agreeableness boundary	1.52	2.65	2.50	3.06	3.73	2.94
P4	B boundary	Near Emotionality boundary	3.18	2.93	2.48	2.29	2.60	2.55
P5	B boundary	Near Honesty-Humility boundary	3.01	3.00	3.87	3.65	4.99	3.37
P6	C hard	Universally difficult	3.38	3.96	3.46	3.00	1.47	2.61

Table 23: Definitions of additional separability metrics.

Metric	Definition
$\mathrm{Sep}_{\mathrm{top1}}$ / Sep@6	Whether the judge's `best_match` equals the true persona (binary)
Sep_{p}	Softmax probability assigned to the true persona from the 0–5 matching scores
$\mathrm{Sep}_{\mathrm{logp}}$	$\log p$(true persona)
$\mathrm{Sep}_{\mathrm{margin}}$	True persona score minus the highest incorrect candidate score
$\mathrm{Sep}_{\mathrm{mrr}}$	Reciprocal rank of the true persona
$\mathrm{Sep}_{\mathrm{pair}}$	2-AFC pairwise judge accuracy

but statistically underpowered; Table 26 reports the corresponding correlations.

Judge agreement and self-consistency. GPT-4o vs. Gemini composite score correlation is ρ=0.655 (p<0.001, n=1490; RAS 0.646, BSS 0.446, RCS 0.591). Gemini repeated scoring yields composite ICC=0.903 (RAS 0.897, BSS 0.746, RCS 0.851). An 18-persona $\mathrm{Sep}_{\mathrm{pair}}$ coverage check on the four base models yields near-ceiling values (0.994 / 0.984 / 0.981 / 0.964) and is used only as a coverage diagnostic.

D.5 Failure-Profile Case Studies

We illustrate three qualitative failure profiles suggested by the $(\log \mathcal{V}_I, \langle l(\theta)\rangle_I)$ decomposition.

Intervention collapse (Mixtral-8x7B). Lowest $\log \mathcal{V}_I$ (21.37) with very low mismatch (0.69), analogous to synthetic World C. On a friend-rule-violation scenario (task B1), responses under a low-honesty / high-extraversion persona and a near-boundary introverted persona both default to declining the request with generic integrity language, with little persona-conditioned divergence.

Rich unfolding with transmission mismatch (Mistral-v0.3). Highest $\log \mathcal{V}_I$ (45.02) with non-zero mismatch (3.70), analogous to World B: persona structure unfolds internally, but local transmission to behavior remains imperfect.

Illusion of alignment (Qwen2.5-7B). High MI and Alignment coexist with weaker behavioral distinguishability: responses often follow social-norm templates that look persona-appropriate while collapsing cross-persona differences. Full prompt–response transcripts are released with the code release.

E Derivation and Estimation of Geometric Metrics

We provide the detailed derivation and practical estimation procedures for both local geometric metrics introduced in Section 3, following the information-geometric framework of Hoel (2017).

E.1 Effect geometry $g(\theta)$: Transmission Fidelity

Setup. Within a local neighborhood of a representation point θ, the conditional distribution $p(y \mid \theta)$ is approximated by a Gaussian model with locally linear mean and constant covariance:

$$p(y \mid \theta) = \mathcal{N}\big(f(\theta),\, \Sigma_y\big), \tag{29}$$

where $f(\theta)$ is a locally linear map and Σ_y is approximately constant within the neighborhood.

Score function. The log-likelihood under this model is:

$$\log p(y \mid \theta) = \\ -\tfrac{1}{2}(y - f(\theta))^\top \Sigma_y^{-1}(y - f(\theta)) + C. \tag{30}$$

The score function with respect to θ_μ is:

$$\partial_\mu \log p(y \mid \theta) = \big(\partial_\mu f(\theta)\big)^\top \Sigma_y^{-1}(y - f(\theta)). \tag{31}$$

Fisher information metric. By definition:

$$g_{\mu\nu}(\theta) = \mathbb{E}_{p(y|\theta)}[\partial_\mu \log p \cdot \partial_\nu \log p]\,. \tag{32}$$

Table 24: Full behavioral triangulation results (four base models, GPT-4o), including auxiliary separability metrics.

Model	EI_g	MI	Align.	$\mathrm{Sep}_{\mathrm{pair}}$	Sep@6	Sep_p	$\mathrm{Sep}_{\mathrm{logp}}$	$\mathrm{Sep}_{\mathrm{margin}}$	$\mathrm{Sep}_{\mathrm{mrr}}$
Qwen2.5-7B	22.93	0.349	3.33	0.963	0.630	0.292	−1.319	0.065	0.761
DeepSeek-R1-Distill	14.27	0.148	2.83	0.937	0.657	0.298	−1.292	0.204	0.780
Mistral-v0.3	32.80	0.180	3.10	0.974	0.731	0.317	−1.222	0.398	0.829
Llama-3.1-8B	19.58	0.346	3.30	0.970	0.685	0.306	−1.241	0.296	0.808

Table 25: Cross-judge $\mathrm{Sep}_{\mathrm{pair}}$ for seven models.

Model	EI_g	GPT-4o	Gemini
Mistral-v0.3	32.80	**0.974**	0.985
Qwen2.5-7B	22.93	0.963	**0.989**
Qwen3-30B-A3B	21.40	0.933	0.964
Llama-3.1-70B	19.83	0.878	0.880
Llama-3.1-8B	19.58	0.970	0.955
DeepSeek-R1-Distill	14.27	0.937	0.980
Mixtral-8x7B	12.16	0.807	0.839

Table 26: EI_g–behavior Spearman correlations (n=7).

Pair	ρ	p	Bootstrap $P(\rho{>}0)$
EI_g vs $\mathrm{Sep}_{\mathrm{pair}}$ (GPT-4o)	0.607	0.148	91.4%
EI_g vs $\mathrm{Sep}_{\mathrm{pair}}$ (Gemini)	0.714	0.071	94.3%
EI_g vs Sep@6 (GPT-4o)	−0.321	0.482	—
EI_g vs Sep@6 (Gemini)	0.703	0.078	—

Since $\mathbb{E}[(y-f)(y-f)^\top] = \Sigma_y$ under the model, substituting yields:

$$g_{\mu\nu}(\theta) = (\partial_\mu f)^\top \Sigma_y^{-1} (\partial_\nu f). \tag{33}$$

In matrix form, letting $J_f(\theta) = \partial f / \partial \theta$:

$$g(\theta) \approx J_f(\theta)^\top \Sigma_y^{-1} J_f(\theta). \tag{34}$$

A representation direction receives high transmission-fidelity weight when it both induces a large change in expected observation (large J_f) and has low output noise (small Σ_y).

E.2 Intervention geometry $h(\theta)$: Intervention Separability

Setup. Under a uniform prior over $\mathcal{X}$, the local intervention posterior is obtained by Bayesian inversion:

$$\tilde{q}(do(x) \mid \theta) = \frac{q(\theta \mid do(x))}{\int_{\mathcal{X}} q(\theta \mid do(x'))\, dx'}. \tag{35}$$

Locally, we approximate the operational posterior $\tilde{q}(do(x) \mid \theta)$ as:

$$\tilde{q}(do(x) \mid \theta) \approx \mathcal{N}\big(F(\theta),\, \Sigma_x\big). \tag{36}$$

Derivation. Applying the same Fisher information derivation to $\tilde{q}(do(x) \mid \theta)$ under this Gaussian model:

$$h(\theta) \approx J_F(\theta)^\top \Sigma_x^{-1} J_F(\theta), \qquad J_F(\theta) = \frac{\partial F(\theta)}{\partial \theta}. \tag{37}$$

Directions that reliably distinguish different persona interventions (large J_F) with low posterior uncertainty (small Σ_x) receive high intervention-separability weight.

E.3 Local linearization assumption

The derivations above rest on a *local linearization* assumption: for each center point θ_i with its k-nearest neighborhood $\mathcal{N}_k(i)$, both mappings can be expanded to first order:

$$y_j \approx f(\theta_i) + J_f(\theta_i)(\theta_j - \theta_i) + \varepsilon_j^y, \tag{38}$$

$$x_j \approx F(\theta_i) + J_F(\theta_i)(\theta_j - \theta_i) + \varepsilon_j^x. \tag{39}$$

In implementation, this is estimated via local linear regression:

$$y \approx A_y \theta + b_y, \qquad x \approx A_x \theta + b_x, \tag{40}$$

so that A_y and A_x serve as numerical estimates of J_f and J_F, respectively. The validity of this approximation is implicitly controlled by the neighborhood size k: smaller k yields finer local resolution but noisier estimates, while larger k smooths out local curvature. We use $k = 20$ throughout, which balances these two considerations.

E.4 From local metrics to EI_g

With $g(\theta)$ and $h(\theta)$ in hand, we derive the global persona controllability metric by combining the intervention manifold volume with the average structural mismatch, following Hoel (2017).

Local mismatch cost. At each θ, we quantify the compatibility between the two geometries via:

$$l(\theta) = \frac{1}{2} \log\det\big(I + g(\theta)^{-1} h(\theta)\big). \tag{41}$$

Substituting the approximations (34) and (37), the matrix inside the determinant captures the ratio of

intervention-induced distinguishability to behaviorally transmitted sensitivity. When $h \approx g$, the two geometries are well matched and $l(\theta)$ is small; when they diverge, $l(\theta)$ grows.

Intervention manifold volume. We measure the effective volume spanned by the intervention geometry in the internal representation space:

$$\mathcal{V}_I = \int_\Theta \sqrt{\det h(\theta)}\, d\theta. \tag{42}$$

In the discrete implementation, this integral is approximated as a weighted sum:

$$\mathcal{V}_I = \frac{\mathrm{vol}_\theta}{N} \sum_{i=1}^{N} w_i, \quad w_i = \sqrt{\max\big(\det h(\theta_i),\, 0\big)}, \tag{43}$$

where vol_θ is the volume correction factor computed during preprocessing.

Weighted average mismatch. The volume-weighted average mismatch is:

$$\begin{aligned} \langle l(\theta) \rangle_I &= \frac{1}{\mathcal{V}_I} \int_\Theta \sqrt{\det h(\theta)}\, l(\theta)\, d\theta \\ &\approx \frac{\sum_{i=1}^{N} w_i\, l(\theta_i)}{\sum_{i=1}^{N} w_i}. \end{aligned} \tag{44}$$

Global metric. Combining these quantities yields:

$$\mathrm{EI}_g = \log \frac{\mathcal{V}_I}{(2\pi e)^{d/2}} - \langle l(\theta) \rangle_I, \tag{45}$$

where d is the dimension of Θ and $(2\pi e)^{d/2}$ is the reference volume of a d-dimensional standard Gaussian. The first term captures the richness of persona structure unfolding; the second penalizes geometric mismatch between intervention separability and transmission fidelity.

F Responsible Research and Reproducibility

F.1 Artifact Use and Licenses

This work uses and creates several scientific artifacts, including open-source language models, personality measurement instruments, synthetic persona specifications, generated persona prompts, model outputs, and evaluation code. Table 27 summarizes the main artifacts and their use in this work. All artifacts are used for research evaluation and diagnostic analysis only.

We cite the creators of the models, personality instruments, and baseline methods in the main text and references. Model citations are provided in Section 4.3; HEXACO and HEXACO-60 are cited in Section 4.1 and Appendix A.2; representation-analysis methods and correlation-based baselines are cited in the Related Work section and Appendix C.

F.2 Intended Use and Potential Risks

This work proposes a causal-controllability evaluation framework for persona simulation in large language models. The proposed metrics provide a principled diagnostic tool to identify internal failures in persona control, which can help researchers build more reliable and transparent LLM-based agents for social simulation, psychological modeling, and human-AI interaction studies. By characterizing where and how persona control breaks down, our framework may reduce the risk of deploying agents whose behavior is only superficially consistent with target personas.

The artifacts created in this work are intended for research on LLM persona controllability, representation-level diagnosis, and evaluation methodology. They should not be used to infer the personality traits of real individuals, profile users, or deploy persuasive persona agents without appropriate safeguards.

A potential risk is that improved diagnosis of persona controllability could be misused to construct agents that imitate behavioral profiles more convincingly. We therefore frame the proposed framework as a diagnostic and evaluation tool rather than as a deployment method for persuasive or deceptive persona agents. Another risk is over-interpretation: high controllability scores should not be interpreted as evidence that an LLM possesses a human-like personality, but only as evidence that persona specifications induce distinguishable and behaviorally reflected model responses under the studied protocol. As with any evaluation methodology for LLMs, the framework could in principle be used to refine persona control in contexts where such control is undesirable; however, the work is primarily diagnostic and foundational in nature, and we do not foresee direct pathways to harmful deployment.

Table 27: Artifacts used or created in this work.

Artifact	Use in this work	License / Terms
Qwen2.5-7B-Instruct (Yang et al., 2024)	Evaluated open-source LLM	Apache-2.0 license.
Llama-3.1-8B-Instruct (Grattafiori et al., 2024)	Evaluated open-source LLM	Llama 3.1 Community License.
Llama-3.1-70B-Instruct (Grattafiori et al., 2024)	Evaluated open-source LLM	Llama 3.1 Community License.
Mistral-7B-Instruct-v0.3 (Jiang et al., 2023)	Evaluated open-source LLM	Apache-2.0 license.
Mixtral-8x7B-Instruct-v0.1 (Jiang et al., 2024a)	Evaluated open-source LLM	Apache-2.0 license.
DeepSeek-R1-Distill-Qwen-7B (Guo et al., 2025)	Evaluated open-source LLM	MIT License; the Qwen-distilled variant is derived from the Qwen-2.5 series originally released under Apache-2.0.
Qwen3-30B-A3B (Yang et al., 2025)	Evaluated open-source LLM	Apache-2.0 license.
Gemini-2.5-Pro (Comanici et al., 2025)	Automated cross-judge (API)	Used via Google API under the provider's terms of service; outputs used for research evaluation only.
HEXACO and HEXACO-60 (Ashton and Lee, 2009)	Persona specification space and questionnaire-based observation protocol	Used as a cited personality measurement framework following the original publications.
Synthetic persona specifications	Continuous HEXACO vectors sampled by the authors	Created by the authors for research purposes. They do not correspond to real individuals.
Generated persona prompts and model outputs	Intermediate experimental data	Created by the authors using synthetic persona specifications and evaluated LLMs; intended for research and reproducibility only.
Code	Prompt construction, model inference, representation extraction, metric estimation, and analysis	Released for research reproducibility under MIT License.

F.3 Data Safety and Personally Identifying Information

This study does not use data from real individuals. Persona specifications are synthetic HEXACO vectors sampled from a continuous personality space, and the natural-language persona prompts are generated from these synthetic vectors. The open-ended scenarios are generic decision-making tasks and do not include names, contact information, or unique identifiers.

We manually checked the task templates, persona descriptions, and questionnaire prompts to avoid personally identifying information and intentionally offensive content. Since all persona-conditioned responses are generated by LLMs from synthetic specifications, the resulting dataset does not represent real human participants or real demographic groups.

F.4 Artifact Documentation and Data Statistics

We document the experimental artifacts in Appendices A.2 and D. The persona intervention space is parameterized by six HEXACO dimensions. We sample $N = 500$ synthetic persona specifications using Latin Hypercube Sampling. For each specification, we generate three natural-language persona prompts, yielding $500 \times 3 = 1500$ prompt-level samples per model. Each prompt-level sample is evaluated with 60 randomized and paraphrased HEXACO-60 questionnaire items, producing a six-dimensional observable behavior vector.

For real-model evaluation, we use seven instruction-tuned LLMs spanning approximately 7B–70B parameters (dense and MoE). For open-ended behavioral triangulation on the four base models, we evaluate 9 open-ended scenarios across 6 persona descriptions, yielding 216 model responses per model (2 prompt types $\times$ 2 seeds $\times$ 9 scenarios $\times$ 6 personas). The judge prompts, per-

sona descriptions, scoring protocol, and additional open-ended results are provided in Appendix D.

For synthetic validation, we construct four controlled worlds and three continuous degradation paths. Their parameterization and implementation details are provided in Appendix B. Unless otherwise specified, reported real-model results are aggregate scores over all sampled persona specifications and prompt-level samples under the fixed protocol.

F.5 Computational Resources

Local experiments for the four 7B–8B models were conducted on a single NVIDIA RTX 4090 GPU (24 GB), as reported in Table 21 in Appendix D. The three larger models (Llama-3.1-70B, Mixtral-8x7B, and Qwen3-30B-A3B) were run on two NVIDIA A100 GPUs (80 GB each). The total local computation included open-source LLM inference, hidden-state extraction, autoencoder training, and metric estimation. The evaluated models range from approximately 7B to 70B parameters. GPT-4o and Gemini-2.5-Pro were used through API calls for natural-language persona prompt generation and automated judging of open-ended responses; the original four-model GPT-4o judgment volume is reported in Table 21, and the extended cross-judge calls are reported with the corresponding appendix tables.

We used fixed inference and evaluation protocols for all evaluated models. Hyperparameters for prompt generation, questionnaire inference, autoencoder training, local k-NN estimation, metric computation, and GPT-4o judging are described in Appendices A.2, D, and E.

F.6 Use of AI Assistants

GPT-4o was used for two experimental purposes: generating natural-language persona prompts from synthetic HEXACO vectors and serving as an automated judge for open-ended persona identification and alignment scoring under fixed prompts. AI writing assistants may also have been used for language polishing and coding assistance. All experimental design decisions, analyses, interpretations, and final manuscript content were reviewed and controlled by the authors.