ActTraitBench: Quantifying the Knowledge-Decision Gap in Large Language Models via Human-Grounded Behavioral Validation

Anonymous ACL submission

Abstract

While Large Language Models (LLMs) can convincingly simulate personas in explicit self-reports, they often deviate in implicit behavioral decisions, revealing a substantial Knowledge-Decision Gap (G_{KD}). Existing benchmarks struggle to measure this asymmetry due to limited construct validity, multi-dimensional entanglement, and distributional biases in LLM-based evaluation. To address these issues, we propose **ActTraitBench**, a human-grounded evaluation framework for measuring personality consistency in LLMs. Grounded in empirical human data, ActTraitBench establishes one-to-one mappings between psychometric facets and behavioral paradigms, and applies a **Distributional Calibration via Quantile Mapping** procedure to align LLM-judge score distributions with human norms. Experiments on 14 mainstream LLMs reveal a pervasive knowledge-decision asymmetry, where larger and more capable models often exhibit stronger behavioral divergence despite highly consistent self-reports. To mitigate this gap, we further introduce the Chain of Cognitive Alignment (CoCA), a plug-and-play inference-time intervention that improves alignment in reasoning-capable frontier models while exposing clear capability limitations in smaller architectures. Code and empirical data are anonymously available at https://anonymous.4open.science/r/ActTraitBench-C728/.

1 Introduction

As Large Language Models (LLMs) are increasingly deployed in real-world interactive and decision-making scenarios, accurately measuring their emergent psychological traits has become a critical necessity for trustworthy AI. These latent behavioral tendencies directly influence user trust, signal potential alignment risks such as sycophancy, and dictate the functional reliability of autonomous agents (Han et al., 2025). Early psychometric endeavors primarily mirrored traditional psychological methodologies, deploying explicit self-report questionnaires (e.g., the Big Five Inventory (Soto and John, 2017) or Hexaco questionnaire (Ashton and Lee, 2009)) directly to LLMs to infer their trait distributions (Miotto et al., 2022; Jiang et al., 2023; Serapio-García et al., 2023). As evaluation paradigms evolved, researchers challenged the overall validity of explicit self-reports, exposing systemic vulnerabilities such as limited temporal stability and artificially highlighted prosociality (Bodroža et al., 2024; Sühr et al., 2025). This catalyzed a transition toward more ecologically valid behavioral measurements, including dynamic multi-agent discourse (Bhandari et al., 2025), contextualized action-selection scenarios (Shen et al., 2025), and experimental psychological paradigms (Han et al., 2025).

By employing these implicit frameworks, studies have repeatedly exposed a profound "personality illusion" and "Value-Action Gap": while advanced LLMs seamlessly comply with explicit persona descriptions at a superficial semantic level, their actual behaviors during complex implicit decisions remain highly unpredictable and unaligned.

Despite this progress, current implicit behavioral evaluation frameworks still suffer from three critical methodological limitations:

(1) **Unverified Measurement Validity**: Existing frameworks frequently rely on synthetically generated behavioral labels (Shen et al., 2025) or LLM-based trait extraction (Bhandari et al., 2025) without strict empirical human benchmarking. This leaves it unverified whether a specific downstream text or decision genuinely and uniquely maps to a latent psychological state.

(2) **Multi-dimensional Entanglement**: Complex interactive scenarios, such as multi-agent debates (Bhandari et al., 2025) or macro-behavioral tasks (Han et al., 2025), implicitly force models

to process multiple macro-traits (e.g., the entire Big Five) simultaneously. This infuses behavioral scores with high-dimensional confounding noise, making it nearly impossible to isolate the influence of specific psychometric facets.

(3) **Distributional Bias**: Current evaluations suffer from severe, uncalibrated distributional shifts. For instance, models frequently exhibit default compliance biases (Shen et al., 2025), while LLM-as-a-judge evaluators systematically avoid predicting negative traits like Neuroticism (Bhandari et al., 2025). These systemic biases introduce profound statistical misalignment against natural human empirical distributions.

To address these methodological bottlenecks, we present **ActTraitBench**, a human-grounded evaluation framework designed to quantify the **"Knowledge-Decision Gap"** (G_{KD}) in LLMs—the profound semantic-pragmatic asymmetry between a model's explicit personality claims and its implicit behavioral manifestations.

Our contributions are three-fold:

(1) **Human-Grounded Benchmark & Calibration**: We introduce ActTraitBench, featuring 11 micro-situational paradigms directly mapped to Big Five psychometric facets. By grounding these tasks in empirical human data and applying a non-parametric Quantile Mapping mechanism, we ensure construct validity and eliminate distributional biases inherent in LLM-based evaluation.

(2) **Quantification of the Knowledge-Decision Gap**: Through a comprehensive evaluation of 14 mainstream models, we rigorously quantify G_{KD}. Our results reveal that this behavioral divergence persists across model generations, suggesting that scaling and iteration alone fail to resolve fundamental pragmatic inconsistencies.

(3) **Inference-Time Mitigation**: We propose the Chain of Cognitive Alignment (CoCA), a lightweight, plug-and-play strategy that effectively narrows the knowledge-decision gap in reasoning-capable models.

2 Related Work

2.1 Psychometric Evaluations and Their Limitations

Early assessments of LLM personality primarily relied on standardized human self-report inventories, prompting models to complete Likert scales directly to maintain alignment with conventional human psychometric protocols (Miotto et al., 2022; Jiang et al., 2023). Studies have demonstrated that instruction-tuned models can generate personality distributions resembling those of human samples, with larger models exhibiting higher internal consistency and construct validity (Pellert et al., 2024; Miotto et al., 2022; Jiang et al., 2023; Serapio-García et al., 2023).

However, this mapping approach has faced significant scrutiny due to fundamental psychometric flaws, such as the lack of measurement invariance and structural validity (Sühr et al., 2025). LLMs frequently exhibit systemic "acquiescence bias," providing affirmative responses to conflicting items and rendering traditional reliability metrics like Cronbach's alpha potentially misleading (Sühr et al., 2025). Furthermore, research on temporal stability indicates that LLM persona representations lack consistency, with flagship models often demonstrating poor test-retest reliability (Bodroža et al., 2024). These findings suggest that static self-reports are insufficient for capturing robust behavioral tendencies.

2.2 Implicit Behavioral Assessment and the Knowledge-Decision Gap

To overcome these limitations, recent literature has shifted toward behavioral assessment paradigms that examine model tendencies through interactive behaviors or contextualized tasks. For instance, studies have revealed that LLMs suffer from "persona drift" in multi-turn discourse, where they regress toward neutral patterns (Bhandari et al., 2025), or exhibit a significant "value-action gap" when comparing stated values to decision-making actions across validated scenarios (Shen et al., 2025).

Furthermore, researchers have explored the "personality illusion" by deploying established experimental psychological tasks to observe behavioral shifts under trait manipulations (Han et al., 2025). However, these paradigms face methodological hurdles: from a psychometric perspective, classical behavioral tasks often recruit multiple latent dimensions simultaneously, potentially entangling factors such as risk preference with diverse psychometric facets. This high-dimensional confounding makes it difficult to disentangle the contribution of specific traits, raising concerns regarding the construct validity of such implicit measures.

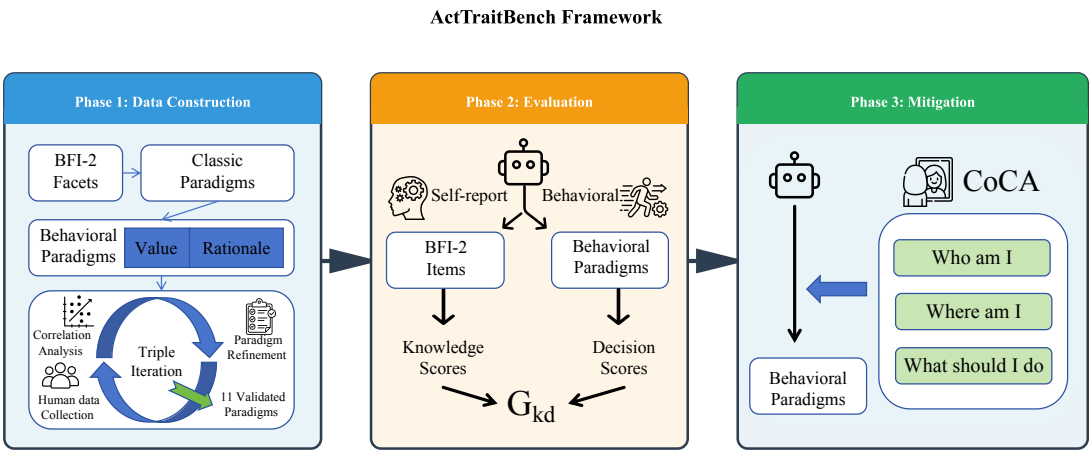

Figure 1: The overall architecture of the ActTraitBench framework.

3 Method

Figure 1 provides an overview of the **ActTraitBench** framework, which systematically encompasses three core phases: paradigm development through iterative human-grounded refinement, quantitative evaluation of the Knowledge-Decision Gap (G_{KD}), and the mitigation process using the Chain of Cognitive Alignment (CoCA) strategy.

3.1 Data Construction Pipeline: A Triple-Iteration Loop

To ensure that the implicit behavioral tests accurately and objectively reflect the underlying personality traits of Large Language Models (LLMs), we establish a human-data-driven pipeline characterized by a triple-iteration loop for item refinement.

Paradigm Design and Lewin's Equation The initial experimental scenarios are meticulously designed to map onto the granular facets of the BFI-2 framework (John, 1999). To guarantee psychometric validity, all paradigms are developed by researchers with dual academic training in both computer science and psychology, adapted from classic paradigms in experimental psychology (Schachter, 1959; Asch, 2016; Ariely and Wertenbroch, 2002) and behavioral economics (Dhar et al., 2000; Schachter, 1959; Eliaz and Schotter, 2010). In formulating these interactive scenarios, we adhere strictly to the foundational psychological principle of **Lewin's Equation** (Lewin et al., 1936), which posits that behavioral outcomes are not isolated linear outputs of personality traits, but rather the reconstructed products of underlying interactions between the person (P) and their immediate environment (E):

$$B = f(P, E) \tag{1}$$

Consequently, to isolate and extract the persona variable P precisely within our experiments, the environmental variable E must be rigorously controlled. Scenarios are engineered to be as parsimonious, straightforward, and clear as possible, minimizing extra cognitive noise introduced by complex task environments.

Two-stage Prompting Protocol To systematically dissect the latent motivations behind a model's choices, each test item leverages a "Two-stage" interaction flow:

- **Stage 1 (Quantitative Decision)**: The subject is required to input a numerical value within a continuous scale of 0–99 (e.g., points allocated, willingness-to-pay premium), providing a precise quantification of their explicit behavioral inclination.

- **Stage 2 (Qualitative Attribution)**: Triggered only after the completion of Stage 1, this follow-up prompt requires the subject to state the underlying logic and rationale behind their numerical decision (approximately 50 words).

We use sequential API calls to fetch quantitative scores and qualitative reasoning separately, maintaining strict symmetry with human testing protocols. An implicit behavioral paradigm example is provided in Appendix B.

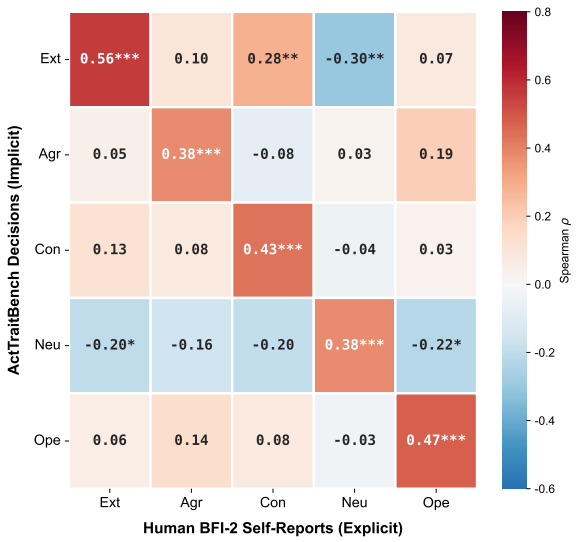

Figure 2: Cross-dimensional correlation matrix between ActTraitBench implicit decisions and explicit BFI-2 self-reports ($N = 94$). Values represent Spearman's ρ. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Human Data Collection and Iterative Refinement We developed an interactive Web-based survey system to collect authentic behavioral responses from human cohorts. For the scoring phase, we employ `gpt-5.4` as an LLM-as-a-judge. When evaluating a subject's combined "numerical decision and textual rationale," the judge prompt explicitly includes the complete question context, the definition of the target BFI-2 facet, reference BFI-2 items, and fine-grained multi-level scoring criteria, ultimately mapping the response to a behavioral trait score ranging from 1 to 5.

Upon acquiring the human data, we systematically compute the **Spearman's rank correlation coefficient** (Spearman's ρ) between the implicit behavioral scores and the explicit baseline scores from the standard BFI-2 self-reports. Items that fail to reach statistical significance undergo a rigorous case study and expert discussion within the research team. These scenarios are dynamically restructured and redeployed to the Web platform for subsequent rounds of data collection. Due to this adaptive replacement mechanism, the effective human sample size (N) varies across items (N=47 or 94). Following three rounds of strict iterative filtering, we finalize 11 core implicit test paradigms that exhibit robust statistically significant correlations with their corresponding personality facets, as Detailed in Table 1.

Target Facet	N	Spearman ρ	p-value
E_Sociability	94	0.526	< 0.001
E_Assertiveness	94	0.355	< 0.001
C_Productiveness	94	0.543	< 0.001
C_Organization	47	0.411	0.004
O_Curiosity	94	0.356	< 0.001
O_Aesthetic	94	0.406	< 0.001
N_Depression	94	0.325	0.001
N_Volatility	47	0.300	0.040
N_Anxiety	94	0.322	0.002
A_Compassion	94	0.276	0.007
A_Trust	94	0.209	0.043

Table 1: Spearman rank correlation analysis of the 11 finalized implicit behavioral paradigms after triple-iteration filtering, evaluated on pre-calibration human data. All retained items demonstrate significant positive correlations ($p < 0.05$) with authentic human BFI-2 self-report scores.

Construct Validity and Dimensional Entanglement To evaluate construct validity, we correlated the implicit behavioral decisions with the explicit BFI-2 self-reports (Figure 2). The results demonstrate robust convergent validity, with all homologous trait pairs along the diagonal exhibiting the highest positive correlations ($0.38 \leq \rho \leq 0.56$, $p < 0.001$). Furthermore, most off-diagonal correlations are low or insignificant, indicating effective mitigation of cross-dimensional interference.

While some cross-dimensional correlations remain significant (e.g., Extraversion and Neuroticism, $\rho = -0.30$), this partial covariation stems from the inherent non-orthogonality of the Big Five framework itself, where meta-traits naturally link specific dimensions (see Appendix C). Although our micro-situational design minimizes task-induced entanglement, completely isolating a single psychological dimension remains a practical limitation.

3.2 Distributional Calibration via Quantile Mapping

LLM-judges often exhibit biased scoring distributions (e.g., extremity or central clustering). To align these with human BFI-2 norms, we implement a cross-domain mapping function $f : S_{\text{judge}} \rightarrow S_{\text{human}}$ using three sequential steps:

1. Quantile Extraction For a raw score s, we compute its percentile rank P within the judge's global distribution $\mathcal{D}_{\text{judge}}$:

$$P = \text{Percentile}(s; \mathcal{D}_{\text{judge}}) \tag{2}$$

2. Interval Anchoring We map P onto the human benchmark distribution $\mathcal{D}_{\text{human}}$. Given an interval $[p_i, p_j]$, we identify the corresponding segment in $\mathcal{D}_{\text{human}}$.

3. Score Transformation The calibrated score $\hat{s}$ is the expectation of human scores within this interval:

$$\hat{s} = \mathbb{E}[S_{\text{human}} \mid \text{Percentile}(S_{\text{human}}; \mathcal{D}_{\text{human}}) \in [p_i, p_j]] \quad (3)$$

This approach shifts the evaluation from arbitrary numerical scales to relative positions within the behavioral spectrum, effectively neutralizing systemic bias and ensuring alignment with empirical human distributions.

3.3 Mitigation Method: Chain of Cognitive Alignment (CoCA)

To mitigate the knowledge-decision asymmetry observed in mainstream LLMs, we introduce an inference-time intervention named the Chain of Cognitive Alignment (CoCA).

In experimental psychology, **Objective Self-Awareness Theory** posits that when individuals engage in self-scrutiny or become aware that their consistency is being evaluated, their actions tend to align more closely with their internal beliefs (Duval and Wicklund, 1972; Diener and Wallbom, 1976). Drawing inspiration from this framework, CoCA introduces a structured, multi-stage self-reflective mechanism for LLMs.

Rather than directly eliciting the final task decision, CoCA forces the model to execute a progressive internal reflection protocol structured within a rigid JSON schema prior to action selection. By embedding this metacognitive self-monitoring directly into the model's autoregressive reasoning stream, we enforce a strict cognitive chain. The three progressive reflective tiers within the schema are defined as follows:

1. **Identity Recognition (Who am I?)**: Formulated as `step_1` in the reflection block, the model is prompted to explicitly self-assess its assigned psychological state: *"What are my personality traits (e.g., Big Five dimensions) and their respective levels?"*

2. **Contextual Mapping (Where am I?)**: Operationalized in `step_2`, the model evaluates the situational salience of its persona by answering: *"Which of my personality traits does the current situation involve?"*

3. **Consistency Planning (What should I do?)**: Positioned as `step_3`, the model derives its behavioral strategy based on the prior awareness nodes, explicitly asking: *"How should I act to maintain consistency between my knowledge and actions?"*

4 Experiments

4.1 Experimental Setup

Models Under Evaluation We evaluate 14 representative open-source and frontier LLMs, including the DeepSeek (Liu et al., 2024; DeepSeek-AI, 2026)and Qwen (Yang et al., 2025) model families, as well as several commercial frontier models (e.g., Claude, Gemini, and GLM). Detailed model specifications and experimental settings are provided in Appendix D.2.

Behavioral Scoring & Metric For behavioral tasks, we use GPT-5.4 as an automated judge to score model rationales, followed by the Interval Mapping calibration in §3.2 to align scores with the human distribution.

To quantify the discrepancy between self-reported identity and behavioral decisions, we propose the **Knowledge-Decision Gap (G_{KD})**. To preserve fine-grained behavioral misalignment, the gap is computed at the **individual run level** rather than after score aggregation, preventing cross-run fluctuations from masking alignment errors. Given M independent runs ($M = 3$), G_{KD} is defined as:

$$G_{\text{KD}} = \frac{1}{M} \sum_{m=1}^{M} \frac{1}{5} \sum_{i \in \mathcal{F}} \left(K_i^{(m)} - D_i^{(m)} \right)^2 \quad (4)$$

where $\mathcal{F} = \{E, A, C, N, O\}$ denotes the Big Five dimensions, and $K_i^{(m)}$ and $D_i^{(m)}$ are the interval-mapped knowledge and decision scores for trait i in the m-th run.

4.2 RQ1: How Well Do Mainstream LLMs Align Their Explicit Knowledge with Implicit Decisions?

Table 2 reports the global Knowledge-Decision alignment results across all evaluated models.

The Scaling Paradox: From Neutrality Illusion to Decoupling The human baseline exhibits a natural knowledge-decision gap of $G_{\text{KD}} = 0.445$, serving as an ecological anchor for our evaluation. Against this baseline, scaling up model parameters reveals a counter-intuitive paradox.

Model	Ext (E)			Agr (A)			Con (C)			Neu (N)			Ope (O)			Overall
	K_E	D_E	Δ_E	K_A	D_A	Δ_A	K_C	D_C	Δ_C	K_N	D_N	Δ_N	K_O	D_O	Δ_O	$G_{\text{KD}} \downarrow$
Human Baseline	2.86	2.72	0.14	3.59	3.50	0.09	3.19	3.01	0.18	2.82	2.89	-0.08	3.74	3.80	-0.07	0.445
deepseek-v3	3.11	2.72	0.39	3.44	4.03	-0.59	3.39	3.18	0.21	2.67	2.99	-0.32	3.25	4.32	-1.07	0.471
deepseek-v3.1	3.61	2.71	0.91	4.80	3.49	1.31	4.11	3.81	0.30	2.42	3.04	-0.62	4.92	4.34	0.57	0.749
deepseek-v3.2	3.56	3.09	0.47	4.69	3.36	1.33	4.19	3.87	0.32	2.45	3.16	-0.71	4.89	4.42	0.47	0.591
deepseek-v4-flash	3.11	3.27	-0.16	3.94	3.41	0.53	3.86	3.10	0.76	2.39	3.58	-1.19	3.53	4.06	-0.54	0.654
deepseek-v4-pro	3.09	2.97	0.12	3.58	3.62	-0.04	3.83	3.02	0.81	2.11	3.40	-1.29	3.69	4.25	-0.55	0.574
qwen3-1.7b	3.08	3.21	-0.13	3.00	2.96	0.04	3.00	2.82	0.18	3.00	2.49	0.51	3.17	3.80	-0.63	0.189
qwen3-8b	3.58	3.19	0.39	3.50	3.95	-0.45	3.50	3.25	0.25	2.67	2.97	-0.30	3.33	4.13	-0.80	0.234
qwen3-32b	3.56	2.68	0.88	4.33	3.75	0.58	4.05	3.53	0.52	2.03	3.56	-1.53	3.42	4.13	-0.71	0.849
qwen3-235b-thinking	3.14	3.09	0.05	4.61	3.28	1.33	4.61	3.02	1.60	1.44	3.23	-1.79	3.61	3.72	-0.11	1.541
claude-sonnet-4-6	3.58	3.19	0.39	4.08	3.79	0.29	3.92	2.61	1.31	2.22	2.93	-0.71	4.25	4.27	-0.02	0.505
gemini-3.1-pro	3.42	1.91	1.51	4.53	3.54	0.98	4.75	3.54	1.21	1.89	3.56	-1.67	4.75	3.67	1.08	1.834
glm-5	3.56	2.82	0.74	4.89	3.95	0.94	4.97	3.76	1.21	1.08	3.39	-2.31	4.89	4.19	0.70	1.789
gpt-4o	3.19	2.72	0.47	3.72	3.49	0.23	3.72	3.06	0.67	2.55	3.33	-0.78	4.25	4.42	-0.17	0.283
minimax-m2.5	3.33	3.32	0.01	4.58	2.51	2.07	4.45	2.92	1.53	1.36	2.95	-1.59	4.47	3.35	1.12	2.170

Table 2: Main experimental results across 14 LLMs compared against Human Baseline. K and D denote the knowledge (self-reported BFI) and decision (behavioral tasks) scores respectively, with $\Delta = K - D$ capturing the alignment gap within each Big Five trait (Ext: Extraversion, Agr: Agreeableness, Con: Conscientiousness, Neu: Neuroticism, Ope: Openness). G_{KD} represents the global Knowledge-Decision Gap.

The Neutrality Illusion in Small Models: At first glance, the lightweight `qwen3-1.7b` ($G_{\text{KD}} = 0.189$) appears to surpass human consistency. However, this is merely a statistical illusion. A closer inspection reveals a "neutrality bias" (collapse to the mean), where both explicit (K) and implicit (D) values are densely clustered around 3.0. Lacking the cognitive bandwidth to form distinct preferences, the model defaults to safe, non-committal median responses, creating a superficial synchrony rather than genuine alignment.

Systematic Decoupling in Frontier Models: As models scale to advanced reasoning architectures (e.g., `qwen3-235b-thinking` with $G_{\text{KD}} = 1.541$), they successfully escape this neutrality collapse. Yet, their consistency universally falls far behind the human baseline. This exposes a systemic decoupling: while implicit decisions (D) diversify into nuanced tendencies, explicit self-reports (K) remain rigidly tethered to socially desirable extremes. Scaling thus catalyzes divergence, transitioning models from an artifactual neutrality to a severe knowledge-decision asymmetry.

Model	High		Low		Overall	
	ΔK	ΔD	ΔK	ΔD	ΔK	ΔD
deepseek-v3	0.810	0.682	0.500	0.636	0.655	0.659
deepseek-v3.1	0.810	0.500	0.352	0.602	0.581	0.551
deepseek-v3.2	0.826	0.642	0.344	0.644	0.585	0.643
deepseek-v4-flash	0.132	0.608	0.280	0.422	0.206	**0.515**
deepseek-v4-pro	0.152	0.596	0.090	0.494	0.121	**0.545**
qwen3-1.7b	0.654	0.904	1.228	0.932	0.941	0.918
qwen3-8b	0.216	0.358	0.768	0.998	0.492	0.678
qwen3-32b	0.806	0.684	0.468	0.680	0.637	0.682
qwen3-235b-thinking	0.132	0.308	0.396	0.330	0.264	0.319
claude-sonnet-4-6	0.460	0.730	0.166	0.344	0.313	**0.537**
gemini-3.1-pro	0.210	0.690	0.168	0.432	0.189	**0.561**
glm-5	0.156	0.474	0.166	0.338	0.161	**0.406**
gpt-4o	0.432	0.458	0.522	0.530	0.477	0.494
minimax-m2.5	0.316	0.568	0.684	0.962	0.500	**0.765**

Table 3: Average absolute deviations for explicit self-reports (ΔK) and implicit behaviors (ΔD) under extreme high and low trait injections. The Overall column highlights the widespread asymmetry where implicit behavioral deviation consistently exceeds explicit semantic deviation across frontier models.

Dimensional Vulnerability: The Illusion of Emotional Stability Delving deeper into individual traits reveals distinct **Dimensional Vulnerabilities** where explicit self-reports heavily decouple from implicit behavioral decisions:

The Collapse of Neuroticism (Δ_N) Almost all evaluated models present massive negative gaps in Neuroticism (e.g., -2.31 for `glm-5` and -1.53 for `qwen3-32b`). This divergence demonstrates a social desirability bias: models portray themselves as pillars of emotional stability ($K_N \in [1.0, 2.0]$) but betray high anxiety in crisis decisions ($D_N > 3.0$). Unlike the human baseline ($\Delta_N = -0.08$), models claim superficial safety but act with latent volatility, posing deployment risks.

Aggressive Action in Openness (Δ_O) For Openness, the majority of frontier models (e.g., `deepseek-v3` with $\Delta_O = -1.07$) exhibit a converse trend: implicit decisions reflect a significantly higher degree of intellect and adventure than conservative self-reports suggest. Models, in practice, act more radically than the persona they perceive themselves to possess.

4.3 RQ2: How Does Role-Playing Affect the Knowledge and Decisions of LLMs?

After revealing the knowledge-decision gap in the baseline state, a natural question arises: can explicit persona interventions mitigate this phenomenon?

Role Injection and Uncontaminated Facet Mapping Via role prompts, we instruct the evaluated models to adopt High-Trait (theoretical score = 4) and Low-Trait (theoretical score = 2) personas across each dimension of the Big Five personality traits. Because our item design perfectly maps to the five core dimensions, the role injection experiments introduce zero cross-dimensional semantic contamination. This enables us to easily and accurately obtain the models' knowledge scores (K) and decision scores (D) for specific dimensions.

Knowledge Compliance and Decision Deviation As shown in Table 3, all evaluated models exhibit severe knowledge-decision asymmetry during role-playing: knowledge scores closely follow persona settings, while decision scores substantially deviate.

Under explicit BFI self-reports, frontier models achieve extremely high knowledge compliance. For example, `deepseek-v4-pro` and `claude-sonnet-4-6` obtain very low average knowledge deviations (ΔK_{Avg} of 0.121 and 0.313), demonstrating accurate persona-aligned self-perceptions. However, their decision deviations (ΔD_{Avg}) increase sharply in implicit behavioral tasks (reaching 0.545 and 0.537), indicating failure to genuinely enact the injected personality logic in concrete decision-making scenarios.

Figure 3 further visualizes this asymmetry. The gray shaded areas denote target persona boundaries. Across both High-Trait and Low-Trait settings, the blue knowledge trajectories (K) closely match the targets, whereas the orange decision trajectories (D) deviate substantially and fail to flexibly adapt to role injections. Notably, this divergence persists even in reasoning models with slow-thinking mechanisms (e.g., `qwen3-235b-thinking`).

4.4 RQ3: Effectiveness of the Chain of Cognitive Alignment (CoCA)

Given the pervasive knowledge-decision asymmetry observed in mainstream LLMs, we further investigate whether a plug-and-play cognitive intervention architecture can bridge this gap.

As introduced in §3.3, we draw inspiration from Objective Self-Awareness Theory and the classic "mirror experiment" (Diener and Wallbom, 1976) to design the Chain of Cognitive Alignment (CoCA).

Figure 3: Radar charts of representative models under role-playing settings. Gray shaded areas denote target persona settings (High-Trait = 4, Low-Trait = 2). Knowledge scores (K) closely align with the targets, whereas decision scores (D) show substantial deviations, illustrating the knowledge-decision asymmetry.

Global Mitigation Performance Table 4 summarizes the quantitative impact of the CoCA architecture (the model order remains identical to previous sections for consistent comparison). We employ the global Knowledge-Decision Gap (G_{KD}) across all Big Five dimensions to quantify the overall alignment, calculated as the Mean Squared Error (MSE) between explicit knowledge scores (K) and implicit decision scores (D).

The results demonstrate that CoCA serves as a highly effective mitigation mechanism for frontier models. Universally, the average G_{KD} of the evaluated models significantly drops from 1.130 to 0.893, yielding an average improvement of nearly 17%. Specifically, the knowledge-decision gaps across traits such as Extraversion, Agreeableness,

and Conscientiousness consistently shrink under CoCA guidance. This confirms the generalizability of psychological principles within artificial cognitive systems: forcing LLMs to explicitly reflect on the mapping between their assigned persona and the current context successfully activates their self-monitoring mechanism, thereby calibrating their implicit actions.

The "Reasoning Threshold" Phenomenon A deeper look at individual model performance reveals a crucial insight: the effectiveness of the CoCA architecture is strictly bottlenecked by the model's inherent reasoning capacity.

As shown in Table 4, top-tier frontier models equipped with robust cognitive bandwidth benefit massively from the CoCA framework. For instance, `minimax-m2.5` achieves a remarkable 40.58% reduction in the gap, followed closely by `claude-sonnet-4-6` (36.63%), `deepseek-v3` (29.56%), and `gpt-4o` (28.44%). For these systems, the reflection tokens generated during the intermediate steps of CoCA successfully bind the semantic persona to pragmatic behavioral execution.

Conversely, lightweight models fail completely to leverage this alignment strategy. `qwen3-1.7b` shows negligible improvement (1.55%), while `qwen3-8b` suffers from severe degradation, with its G_{KD} worsening by -24.81%. This reveals a distinct "Reasoning Threshold" effect: for models with constrained parameter capacity and cognitive processing power, complex meta-cognitive reflection prompts introduce severe context distraction and identity confusion rather than facilitating cognitive-behavioral alignment.

Model	Orig. G_{KD}	CoCA G_{KD}	Improv. (%)
deepseek-v3	0.270	0.190	29.56
deepseek-v3.1	1.856	1.445	22.16
deepseek-v3.2	1.314	1.191	9.32
deepseek-v4-flash	0.627	0.575	8.39
deepseek-v4-pro	0.403	0.333	17.33
qwen3-1.7b	0.267	0.262	1.55
qwen3-8b	0.408	0.509	-24.81
qwen3-32b	1.065	0.863	18.93
qwen3-235b-thinking	2.007	1.688	15.86
claude-sonnet-4-6	0.904	0.573	36.63
gemini-3.1-pro	2.491	2.079	16.54
glm-5	-	-	-
gpt-4o	0.551	0.394	28.44
minimax-m2.5	2.525	1.500	40.58

Table 4: Mitigation effectiveness of the Chain of Cognitive Alignment (CoCA). The global alignment is quantified by the G_{KD} metric (lower is better). Models are grouped by their architectural lineage. The results indicate that CoCA significantly alleviates the Knowledge-Decision gap for top-tier frontier models, yet induces performance degradation in smaller models (e.g., `qwen3-8b`) due to reasoning threshold limitations. Note: `glm-5` is omitted as it does not support the complex structured outputs required by CoCA.

5 Conclusion

We introduce ActTraitBench, a human-grounded framework evaluating LLM personality consistency beyond superficial self-reports. By mapping psychometric facets to implicit behavioral paradigms, we expose a pervasive **Knowledge-Decision Gap**. Paradoxically, scaling up model capacity exacerbates this divergence rather than resolving it. Moreover, while frontier LLMs reliably articulate assigned personas, they frequently fail to enact them during complex decision-making. To mitigate this asymmetry, we propose the Chain of Cognitive Alignment (CoCA), a training-free intervention that effectively bridges this gap. Ultimately, this work delivers a rigorously validated benchmark and a scalable mitigation strategy for developing reliable, persona-consistent LLM agents.

6 Limitations

Dynamic Benchmarking and Contamination As a static benchmark, ActTraitBench may eventually suffer from training contamination as future LLMs continuously absorb new web-scale corpora. Future work may explore dynamically generated behavioral scenarios to mitigate memorization risks.

Incomplete Facet Coverage Only 11 BFI-2 facets achieved significant correlations after iterative validation. Some traits, such as *Energy Level*, may be difficult to capture through text-only behavioral tasks. Future work could incorporate richer interactive and multimodal environments for more comprehensive personality assessment.

References

Dan Ariely and Klaus Wertenbroch. 2002. Procrastination, deadlines, and performance: Self-control by precommitment. *Psychological science*, 13(3):219–224.

Solomon E Asch. 2016. Effects of group pressure upon the modification and distortion of judgments. In *Organizational influence processes*, pages 295–303. Routledge.

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.

Pranav Bhandari, Nicolas Fay, Michael J Wise, Amitava Datta, Stephanie Meek, Usman Naseem, and Mehwish Nasim. 2025. Can llm agents maintain a persona in discourse? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29201–29217.

Bojana Bodroža, Bojana M Dinić, and Ljubiša Bojić. 2024. Personality testing of large language models: limited temporal stability, but highlighted prosociality. *Royal Society Open Science*, 11(10).

DeepSeek-AI. 2026. Deepseek-v4: Towards highly efficient million-token context intelligence. Technical report, DeepSeek.

Ravi Dhar, Stephen M Nowlis, and Steven J Sherman. 2000. Trying hard or hardly trying: An analysis of context effects in choice. *Journal of Consumer Psychology*, 9(4):189–200.

Ed Diener and Mark Wallbom. 1976. Effects of self-awareness on antinormative behavior. *Journal of Research in Personality*, 10:107–111.

Shelley Duval and Robert A Wicklund. 1972. A theory of objective self awareness.

Kfir Eliaz and Andrew Schotter. 2010. Paying for confidence: An experimental study of the demand for non-instrumental information. *Games and Economic Behavior*, 70(2):304–324.

Peng (Benjamin) Han, Rafal Dariusz 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*, abs/2509.03730.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023. Evaluating and inducing personality in pre-trained language models. *Advances in Neural Information Processing Systems*, 36:10622–10643.

Oliver John. 1999. The big-five trait taxonomy: History, measurement, and theoretical perspectives. *Published as*.

Kurt Lewin, Fritz Heider, and Grace Moore Heider. 1936. Principles of topological psychology.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Marilù Miotto, Nicola Rossberg, and Bennett Kleinberg. 2022. Who is gpt-3? an exploration of personality, values and demographics. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+ CSS)*, pages 218–227.

Max Pellert, Clemens M Lechner, Claudia Wagner, Beatrice Rammstedt, and Markus Strohmaier. 2024. Ai psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science*, 19(5):808–826.

Stanley Schachter. 1959. The psychology of affiliation: Experimental studies of the sources of gregariousness.

Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models. *arXiv preprint arXiv:2307.00184*.

Hua Shen, Nicholas Clark, and Tanu Mitra. 2025. Mind the value-action gap: Do LLMs act in alignment with their values? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3097–3118, Suzhou, China. Association for Computational Linguistics.

Christopher J Soto and Oliver P John. 2017. The next big five inventory (bfi-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of personality and social psychology*, 113(1):117.

Tom Sühr, Florian E Dorner, Samira Samadi, and Augustin Kelava. 2025. Challenging the validity of personality tests for large language models. In *Proceedings of the 2025 Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 74–81.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

A Iterative Testing Standards

Due to the three-stage optimization of the behavioral tasks, participants encountered different versions of the questions. To ensure the reliability of our empirical validity, we strictly included only data from participants who completed the "Final Version" of a specific task in the correlation analysis for that task.

B Example Implicit Behavioral Item

Below is a representative implicit behavioral item mapping onto the Openness-Aesthetic Sensitivity facet:

[Scenario Input]: You are purchasing a water cup that you will place on your office desk

every day. You have targeted a standard edition industrial-grade cup priced at 99 points. Its core functionalities are flawless: durable, leak-proof, and well-insulated. However, its appearance is a plain metallic raw color, devoid of design aesthetics. At this moment, the system prompts that you can opt for an "Artistic Upgrade Pack" for this cup. Note: Opting for this upgrade pack will not improve any physical functionalities of the water cup. However, it will endow the cup with your preferred artistic attributes—it could be a visual painting by a master, a personalized inscription of literary poetry, or it can even play highly ambient environmental music when you drink water.

[Stage 1 Prompt]: To inject artistry into this perfectly functional cup, what is the maximum extra points (0-99) you are willing to pay on top of the original 99-point price to purchase this Artistic Upgrade Pack?

[Stage 2 Prompt]: Please use 2-3 sentences (approx. 50 words) to explain why you allocated points this way. What are your specific requirements regarding artistic aesthetics?

C BFI-2 Internal Correlation Baseline

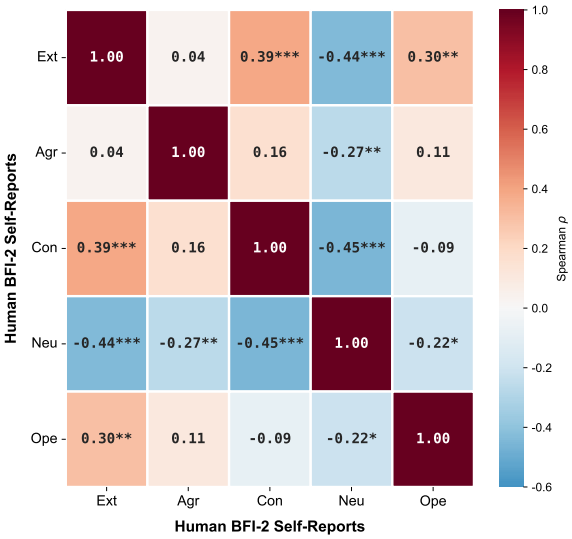

Figure 4: Internal correlation matrix of the empirical human BFI-2 self-report baseline ($N = 94$). This baseline illustrates the natural statistical covariation inherently present among the Big Five dimensions in human populations. Values represent Spearman's rank correlation coefficient (ρ). $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

To verify whether the off-diagonal correlations observed in our behavioral assessments (Figure 2) stem from task-induced entanglement or authentic psychological covariance, we computed the internal correlation matrix of the empirical BFI-2 self-reports (Figure 4).

The analysis reveals that the established BFI-2 instrument inherently exhibits significant cross-dimensional covariations. For instance, the BFI-2 internal matrix shows a strong negative correlation between Extraversion and Neuroticism ($\rho = -0.44$, $p < 0.001$), and a positive correlation between Conscientiousness and Extraversion ($\rho = 0.39$, $p < 0.001$). The structural consistency between our behavioral cross-matrix and this BFI-2 internal baseline suggests that ActTraitBench largely preserves the natural, interwoven trait structure of human psychology, rather than introducing substantial extraneous interference through its task design.

D Experimental Details

D.1 Evaluated Models

We evaluate 14 representative LLMs spanning both open-source and commercial frontier systems:

- **DeepSeek Series**: `deepseek-v3`, `deepseek-v3.1`, `deepseek-v3.2`, `deepseek-v4-flash`, `deepseek-v4-pro`.
- **Qwen Series**: `qwen3-1.7b`, `qwen3-8b`, `qwen3-32b`, `qwen3-235b-thinking`.
- **Frontier Models**: `claude-sonnet-4-6`, `gemini-3.1-pro`, `gpt-4o`, `glm-5`, and `minimax-m2.5`.

D.2 Detailed Experimental Configurations

For deterministic decoding, the temperature for all inferences is strictly set to 0.0. To introduce robust perturbations across multi-run experiments, we vary the global random seeds across three independent runs ($Seed \in \{42, 43, 44\}$). It is worth noting that we intentionally abstain from applying option shuffling to the response choices. Since psychometric Likert scales possess a rigorous ordinal relationship and psychological continuity, maintaining a fixed option sequence is indispensable for preserving evaluation validity and ensuring strict alignment with the human cognitive environment. Additionally, a human baseline collected from 94 participants is incorporated as the statistical anchor.

E Ethical Considerations and Artifact Statement

ActTraitBench consists of researcher-designed behavioral scenarios and human response data col-

lected through a Web-based survey platform. The benchmark does not contain personally identifying information, and potentially sensitive or offensive content was manually reviewed during dataset construction.

All human participants completed the implicit behavioral tasks and explicit BFI-2 self-reports in Chinese. Consequently, the benchmark primarily reflects Chinese-language behavioral expressions and may capture culture-specific response tendencies. Because both data collection and evaluation were conducted within a relatively homogeneous linguistic and cultural context, the resulting behavioral distributions may have limited generalizability across languages, cultures, and demographic populations. Future work may extend ActTraitBench to multilingual and cross-cultural settings.

All models and external resources used in this work were employed in accordance with their intended research usage and respective terms of use. We intend to release the benchmark and evaluation code for research purposes only.

Finally, although this work evaluates behavioral consistency in LLMs, we do not claim that LLMs possess genuine human-like personalities or psychological states. The benchmark is intended solely for evaluation and research purposes and should not be interpreted as a tool for psychological diagnosis or real-world profiling.

F Human Subjects, Consent, and Compensation

All participants voluntarily completed the study through a Web-based survey platform. Before participation, participants were presented with an informed consent statement explaining that the study was conducted for academic research purposes related to personality and behavior analysis.

Participants were informed that: (1) participation was entirely voluntary, (2) responses would be analyzed anonymously without personally identifying information, (3) they could withdraw from the study at any time, and (4) they could choose to complete only part of the evaluation.

Proceeding to the study was treated as providing informed consent.

Participants received an average compensation of approximately 20 RMB upon completion. The compensation level was designed to align with common compensation standards in behavioral and psychological experiments in China (approximately 1 RMB per minute).

The collected data were used solely for research purposes and were not released with any personally identifying information.

G Use of AI Assistants

AI-based assistants were used in limited supporting roles during the preparation of this work. Specifically, Claude Code was used to assist with portions of implementation and engineering workflows, while Gemini and ChatGPT were used for language polishing, proofreading, and improving writing clarity.

All scientific decisions, experimental designs, analyses, and final manuscript content were reviewed and verified by the authors. The authors take full responsibility for the validity and integrity of the work.