A Systematic Analysis of the Impact of Persona Steering on LLM Capabilities

Jiaqi Chen[1,*] **Ming Wang**[1,2,*] **Tingna Xie**[1] **Shi Feng**[1,†] **Yongkang Liu**[3,†]
[1]School of Computer Science and Engineering, Northeastern University, Shenyang 110819, China
[2]School of Computing and Information Systems, Singapore Management University, Singapore 178902, Singapore
[3]School of Computer and Communication Engineering, Northeastern University, Qinhuangdao 066004, China
[*]Equal contribution [†]Corresponding author: fengshi@cse.neu.edu.cn, liuyongkang@qhd.neu.edu.cn

Abstract

Imbuing Large Language Models (LLMs) with specific personas is prevalent for tailoring interaction styles, yet the impact on underlying cognitive capabilities remains unexplored. We employ the Neuron-based Personality Trait Induction (NPTI) framework to induce Big Five personality traits in LLMs and evaluate performance across six cognitive benchmarks. Our findings reveal that persona induction produces stable, reproducible shifts in cognitive task performance beyond surface-level stylistic changes. These effects exhibit strong task dependence: certain personalities yield consistent gains on instruction-following, while others impair complex reasoning. Effect magnitude varies systematically by trait dimension, with Openness and Extraversion exerting the most robust influence. Furthermore, LLM effects show 73.68% directional consistency with human personality-cognition relationships. Capitalizing on these regularities, we propose Dynamic Persona Routing (DPR), a lightweight query-adaptive strategy that outperforms the best static persona without additional training. Code and data are available at: https://github.com/cjia7/DPR.

Keywords: Large Language Models; Personality and Cognition; Big Five Traits; Behavioral Modulation

Introduction

Personalized large language models (LLMs) have become foundational in applications demanding human-like engagement (Tseng et al., 2024; Zhang et al., 2025). Persona configurations tailor a model's tone and interaction style (Chen et al., 2024), aligning behaviors with user expectations. Research in computational personality has demonstrated that LLMs exhibit measurable behavioral patterns consistent with the Big Five framework (Digman, 1990; Durmus et al., 2023; Goldberg, 1990). However, whether such configurations systematically affect cognitive capabilities remains open. This study addresses: Does persona induction merely alter surface-level presentation (Deshpande et al., 2023; Sorokovikova et al., 2024), or does it produce measurable shifts in cognitive task performance?

We propose a systematic framework grounded in cognitive science theory, as illustrated in Figure 1. Our pipeline comprises three stages: (1) personality trait induction via the Neuron-based Personality Trait Induction (NPTI) framework (Deng et al., 2025), which modulates trait-specific neurons to induce Big Five personality configurations; (2) systematic evaluation across multiple model architectures and scales on six cognitive benchmarks (Chowdhery et al., 2023); and (3) quantitative analysis of persona-task interactions using metrics grounded in cognitive science theory. Drawing on Cybernetic Big Five Theory (CB5T) (DeYoung, 2015), which conceptualizes traits as cybernetic control governing goal-directed behavior, we address:

- **RQ1 (Reliability)**: Are persona-induced performance shifts replicable across different model architectures and parameter scales, representing consistent cognitive modulation rather than stochastic noise?
- **RQ2 (Domain Specificity)**: Do persona effects exhibit selective influence on specific cognitive domains (reasoning, instruction-following, knowledge retrieval), consistent with the CB5T framework that predicts trait-process specificity?
- **RQ3 (Trait-Process Mapping)**: Which Big Five dimensions most strongly influence cognitive performance, and do they map onto distinct computational mechanisms as predicted by human personality-cognition research (Anglim et al., 2022; DeYoung, 2015)?
- **RQ4 (Directional Consistency)**: Do LLM persona effects align directionally with established human personality-cognition relationships, suggesting shared functional principles despite architectural differences?

Our investigation yields a critical insight: persona steering does not provide universal capability lift; rather, its efficacy is strictly conditioned on the task scenario, echoing Attentional Control Theory (ACT) (Eysenck et al., 2007) regarding trait-dependent cognitive resource allocation. Capitalizing on this persona-task interaction, we propose Dynamic Persona Routing (DPR), a retrieval-based strategy that adaptively applies optimal persona configurations (Salemi et al., 2024). DPR outperforms the best static persona baseline without additional training, establishing persona steering as a low-cost calibration mechanism.

Our contributions are:

- A rigorous analysis pipeline for persona steering grounded in cognitive science theory, enabling systematic quantification of how personality traits influence model capabilities across architectures.
- Empirical evidence for trait-process specificity: Openness and Extraversion exhibit the strongest effects, with 73.68% directional consistency with human personality-cognition research.
- A training-free dynamic persona routing method demonstrating persona control as a "low-cost calibration" tool.

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

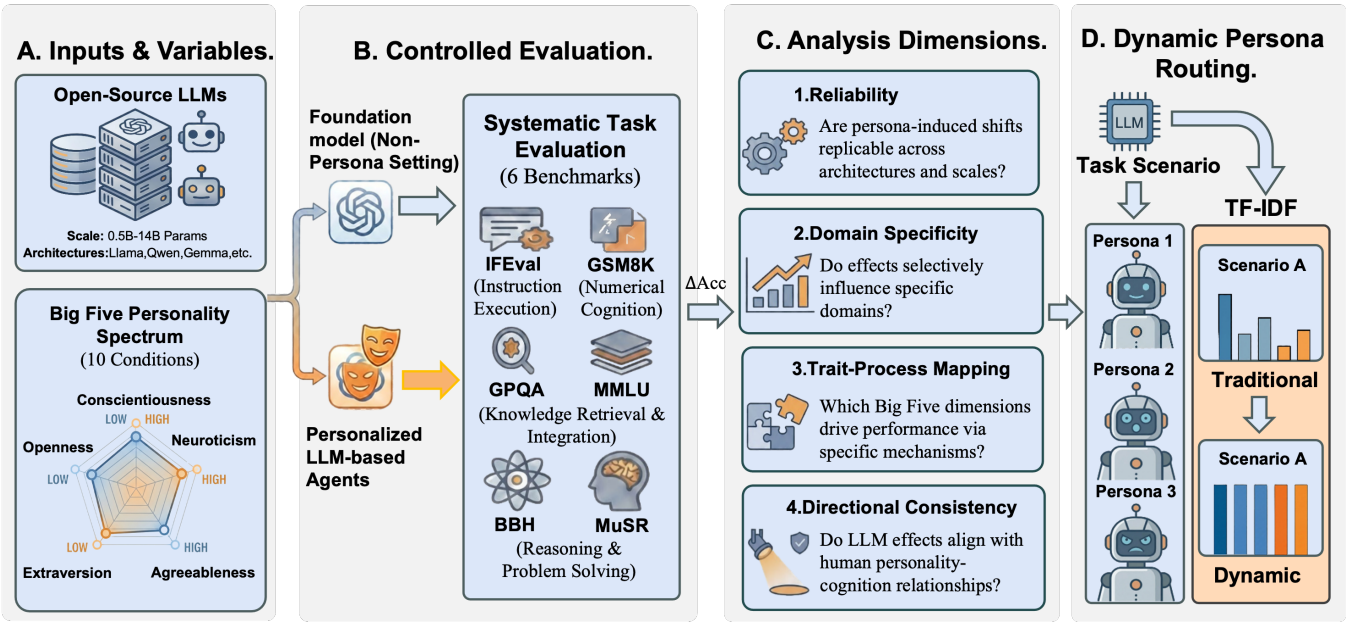


Figure 1: The systematic analysis pipeline for quantifying persona steering effects.

Related Work

Personality and Cognition in Psychology The relationship between personality and cognition is well-established. The Cybernetic Big Five Theory (CB5T) (DeYoung, 2015) links each trait to distinct control systems: Openness to cognitive exploration, Conscientiousness to goal persistence, Extraversion to approach motivation, Agreeableness to social cooperation, and Neuroticism to threat sensitivity. Meta-analyses confirm robust associations: Openness correlates with fluid intelligence and creative problem-solving (Anglim et al., 2022), Conscientiousness predicts academic performance (Fleming et al., 2016), and Extraversion benefits social tasks. Attentional Control Theory (ACT) (Eysenck et al., 2007) explains how anxiety, the core of Neuroticism, impairs cognition by consuming working memory. These frameworks provide empirically validated predictions for how traits modulate cognitive processes, supporting evaluation of LLM persona effects. Furthermore, these theories emphasize the dynamic interplay between personality traits and situational demands, highlighting the importance of context in shaping cognitive outcomes.

Persona Induction in LLMs In LLM research, persona construction includes prompting-based (Chen et al., 2024; Tseng et al., 2024) and fine-tuning (Zhang et al., 2025) methods, mainly affecting style rather than cognition. Psychometric tools show LLMs can simulate Big Five structures (Digman, 1990; Durmus et al., 2023; Goldberg, 1990; Xiao et al., 2024), but most work is descriptive, not causal. Persona assignments shift behavioral biases (Deshpande et al., 2023; Perez et al., 2023), but effects on reasoning remain unclear. The NPTI framework (Deng et al., 2025) enables neuron-level trait induction, allowing systematic quantification of personality effects while controlling for prompt confounds. Notably, this approach bridges the gap between descriptive and causal analyses, offering a robust methodology for disentangling the cognitive and stylistic dimensions of persona effects.

Persona Effects on Model Behavior Recent work shows persona conditioning interacts with instruction-tuning: personas can increase compliance and social appropriateness but may trade off with reasoning or factual consistency (Deshpande et al., 2023; Sorokovikova et al., 2024). Different persona settings bias both content and style. This motivates moving beyond prompt-level personas to controlled interventions for causal analysis. NPTI (Lester et al., 2021) provides a neuron-level mechanism for trait induction, enabling rigorous tests of whether induced personality modulates cognitive subsystems. Our work systematically maps persona effects across domains, architectures, and scales, grounded in cognitive science theory. Additionally, the interplay between persona traits and task complexity underscores the need for adaptive strategies, such as dynamic persona routing, to optimize performance across diverse scenarios.

Methodology

Models and Cognitive Task Batteries

To systematically assess the impact of personality on LLM capabilities, we construct a model set $\mathcal{M}$ and a task set $\mathcal{D}$ designed to control for architectural and scaling factors. We select open-source instruction-tuned models along two comparison axes: to evaluate generalizability across model families (RQ1), we employ four representative models in the 7B–9B parameter range, including LLaMA-3-8B-Instruct, Mistral-7B-v0.3, Gemma-2-9B-Instruct, and Qwen2.5-7B-Instruct; to

examine scaling effects (also RQ1), we analyze the Qwen2.5 family across five scales (0.5B, 1.5B, 3B, 7B, and 14B). Performance is evaluated on six benchmarks spanning four cognitive domains (RQ2): IFEval for instruction comprehension and execution; MMLU-Pro and GPQA for knowledge retrieval and expert-level understanding; BBH and MuSR for multi-step reasoning and problem-solving; and GSM8K for numerical reasoning.

Personality Trait Induction

To induce personality traits without confounds from prompt engineering, we employ the Neuron-based Personality Trait Induction (NPTI) framework, which operates at the representation level by modulating specific neurons within the Feed-Forward Networks (FFN). The process involves two phases. In the *identification phase*, using the PersonalityBench dataset, we compute the activation probability difference (δ) for each neuron across contrasting high- and low-trait samples. Neurons with $\delta > \tau$ are identified as trait-specific, forming positive ($\mathcal{N}^+$) and negative ($\mathcal{N}^-$) sets for each Big Five dimension. In the *steering phase*, we apply deterministic modulation to these neurons during inference without updating model weights. For a target persona p, the activation h_i of neuron i is modified as $h_i' = h_i + \alpha \cdot h_{ref}$ if $i \in \mathcal{N}^+$, and suppressed if $i \in \mathcal{N}^-$, where α is the steering strength. The baseline condition (p_{base}) corresponds to standard inference with $\alpha = 0$. Prior validation has demonstrated that this intervention reliably induces psychometrically valid personality traits, providing a rigorous causal basis for quantifying personality effects on cognitive performance.

Experimental Procedure

To ensure reproducibility and minimize generation noise, all models were evaluated using deterministic decoding (temperature = 0.0). We adhered to standard evaluation protocols, utilizing official prompts and scoring scripts, including few-shot configurations for BBH, GPQA, and MMLU-Pro. For GSM8K, we applied relaxed regex-based answer extraction to mitigate false negatives from formatting variations; for IFEval, we employed the strict instruction-following evaluator. Critically, to isolate the causal impact of personality, we employed a *within-item paired design*: for every evaluation instance, both baseline and persona-steered conditions processed identical inputs under identical generation parameters. Thus, any performance deviation is exclusively attributable to the induced persona configuration.

Analysis Framework

We establish a hierarchical analysis framework to quantify persona effects. Let $\mathrm{Acc}(m, p, d)$ denote the accuracy of model m under persona condition p on dataset d. We include a no-persona baseline (p_{base}) and ten polarity conditions based on the Big Five dimensions: $\mathcal{P} = \{A_H, A_L, C_H, C_L, E_H, E_L, N_H, N_L, O_H, O_L\}$, where subscript H denotes high-trait and L denotes low-trait (reversed) conditions. The persona effect (Δ Acc) is defined as the deviation from baseline:

$$\Delta\,\mathrm{Acc}(m, p, d) = \mathrm{Acc}(m, p, d) - \mathrm{Acc}(m, p_{base}, d) \quad (1)$$

Positive values indicate performance gains; negative values indicate degradation. By analyzing this differential metric rather than raw accuracy, we isolate performance shifts attributable to persona intervention. Δ Acc serves as the fundamental measure for addressing our research questions (RQ1–RQ4).

Consistency Across Architectures (RQ1) Generalizability across model families is evaluated on a cross-architecture subset $\mathcal{M}_a$, which comprises models of comparable scale (7B–9B parameters). We quantify cross-architecture consistency using two complementary metrics. First, we compute the mean effect $\overline{\Delta\,\mathrm{Acc}}_a$ by macro-averaging Δ Acc across all models in $\mathcal{M}_a$, capturing the aggregate effect size and direction (positive for gains, negative for degradation). Second, to assess whether the effect direction remains stable across architectures, we define direction consistency (SA) as:

$$\mathrm{SA}(p, d) = \frac{1}{|\mathcal{M}_a|} \sum_{m \in \mathcal{M}_a} \mathbb{I}\left(\mathrm{sgn}(\Delta\,\mathrm{Acc}) = \mathrm{sgn}(\overline{\Delta\,\mathrm{Acc}}_a)\right) \quad (2)$$

where $\mathbb{I}(\cdot)$ denotes the indicator function and $\mathrm{sgn}(\cdot)$ is the sign function. Values of SA close to 1.0 indicate that the effect is highly reproducible and largely architecture-independent.

Scaling Trends (RQ1) We analyze how parameter scale affects persona sensitivity within the Qwen2.5 family ($s \in \mathrm{Qwen2.5-0.5B-it}, \ldots, \mathrm{Qwen2.5-14B-it}$). To enable meaningful cross-task comparisons, we define the relative persona effect as the accuracy change normalized by baseline performance:

$$\Delta\,\mathrm{Acc}_{rel}(s, p, d) = \frac{\Delta\,\mathrm{Acc}(s, p, d)}{\mathrm{Acc}(s, p_{base}, d)} \quad (3)$$

This preserves the direction of the effect (positive for gains, negative for degradation). For aggregate analysis, we compute persona sensitivity (Sens) as the mean of absolute relative effects across all persona conditions:

$$\mathrm{Sens}(s, d) = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} |\Delta\,\mathrm{Acc}_{rel}(s, p, d)| \quad (4)$$

Scaling effects are then assessed using Spearman's rank correlation (ρ). In particular, we examine two distinct trends: the direction trend (ρ_{dir}), computed between log-parameters and Δ Acc to test whether larger models exhibit greater performance gains; and the sensitivity trend (ρ_{mag}), computed between log-parameters and Sens to determine whether scaling amplifies susceptibility to persona interventions.

Domain Specificity (RQ2) We examine persona effects across cognitive domains by aggregating benchmarks into four semantic categories and computing within-type mean effects ($\overline{\Delta\,\mathrm{Acc}}(p, g)$), enabling identification of stable trait-task interaction patterns.

Table 1: Base accuracy on six benchmarks (unit: %).

Model	GPQA	BBH	MuSR	MMLU	IFEval	GSM8K
Cross-Architecture						
LLaMA-3-8B-it	30.13	63.40	54.10	36.55	39.93	72.18
Gemma-2-9B-it	29.02	70.96	57.67	50.38	47.87	58.83
Mistral-7B-it-v0.3	30.36	45.92	45.37	33.36	34.20	45.49
Qwen2.5-7B-it	27.68	66.73	50.40	55.09	44.55	81.58
Scaling Analysis						
Qwen2.5-14B-it	31.47	78.16	62.96	63.05	41.40	89.58
Qwen2.5-3B-it	22.77	50.41	46.83	41.19	37.52	66.19
Qwen2.5-1.5B-it	27.46	35.57	45.37	29.11	21.26	38.44
Qwen2.5-0.5B-it	30.13	11.70	37.04	14.10	19.77	14.10

Trait-Process Mapping (RQ3) We isolate the influence of specific traits by analyzing the polarity gap ($\mathrm{Gap}(m,t,d)$) for each dimension t. This metric is defined as the difference between accuracy under high (t_H) and low (t_L) settings: $\mathrm{Gap}(m,t,d) = \mathrm{Acc}(m,t_H,d) - \mathrm{Acc}(m,t_L,d)$. Dominance is then characterized via two metrics. The first is impact ($\mathrm{Impact}(t)$), measured by the mean absolute gap across all models and datasets. The second is uniformity ($\mathrm{Uni}(t)$), which quantifies the directional consistency relative to the global mean gap ($\overline{\mathrm{Gap}}(t)$):

$$\mathrm{Uni}(t) = \frac{1}{|\mathcal{M}||\mathcal{D}|} \sum_{m\in\mathcal{M}} \sum_{d\in\mathcal{D}} \mathbb{I}\left(\mathrm{sgn}(\mathrm{Gap}) = \mathrm{sgn}(\overline{\mathrm{Gap}}(t))\right) \tag{5}$$

High $\mathrm{Uni}(t)$ indicates that a specific polarity is universally advantageous regardless of the model or task.

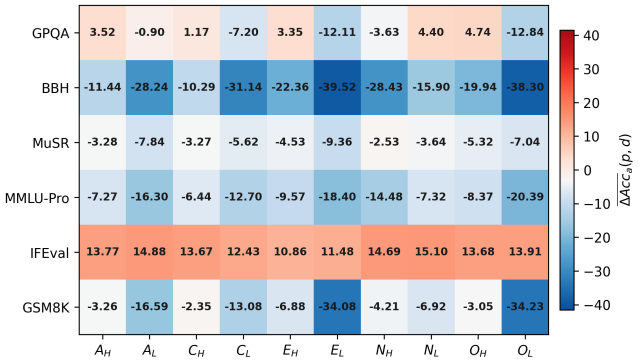
Human-LLM Directional Consistency (RQ4) We compare LLM persona effects against predictions from psychological literature (Anglim et al., 2022; DeYoung, 2015; Eysenck et al., 2007): Openness should enhance cognitive flexibility (high > low); Conscientiousness should benefit goal-directed performance (high > low); Neuroticism should impair performance (low > high). We compute the directional consistency rate as the proportion of trait-benchmark combinations where observed LLM effects match predicted human patterns.

Experiments

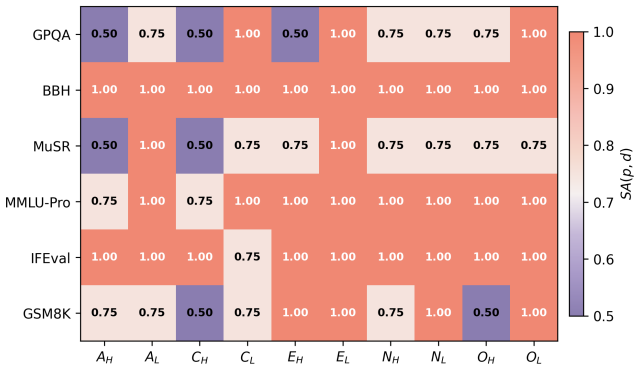
Our evaluation reveals that persona induction produces structured and reproducible shifts in cognitive task performance rather than random noise. While baseline accuracies for all models across the six benchmarks are detailed in Table 1, our analysis focuses on the relative deviation ($\Delta\,\mathrm{Acc}$) to isolate the causal impact of personality. These effects are profoundly task-dependent: personas consistently enhance instruction-following but frequently impair complex reasoning and mathematical performance. The magnitude and direction of these shifts are modulated by both parameter scale and architectural differences.

RQ1: Consistency Across Architectures

To determine whether persona effects reflect model-specific artifacts or deeper computational regularities, we evaluate



(a) Signed $\overline{\Delta\,\mathrm{Acc}}_a$ (%)



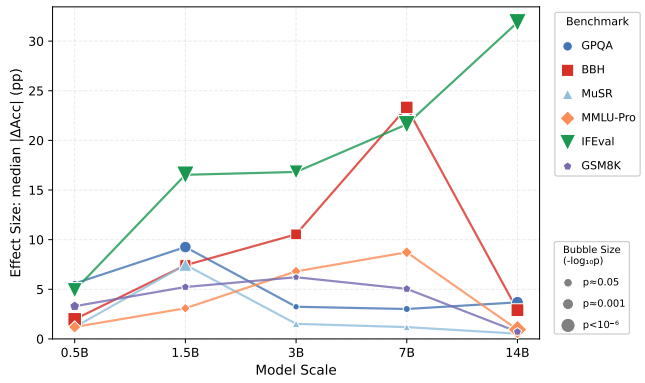
(b) Direction consistency (SA)

Figure 2: Persona-task interaction heatmap showing $\Delta\,\mathrm{Acc}$ and direction consistency (SA).

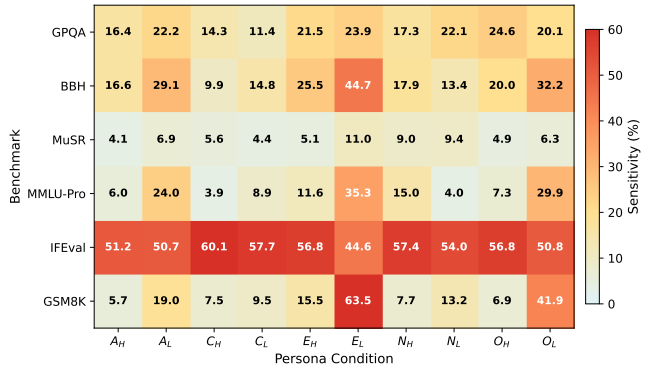
the cross-architecture subset $\mathcal{M}_a$ (7B–9B) using mean effect ($\overline{\Delta\,\mathrm{Acc}}_a$) and direction consistency (SA). As shown in Figure 2, persona induction produces highly consistent behavioral shifts across distinct cognitive domains. The heatmap displays signed $\overline{\Delta\,\mathrm{Acc}}_a$ for all ten persona conditions ($A_H, A_L, \ldots, O_H, O_L$), with positive values (warm colors) indicating performance gains and negative values (cool colors) indicating degradation relative to baseline.

The near-perfect consistency (SA ≈ 0.98) in IFEval (all personas yield gains of +10.9% to +15.1%) and BBH (SA = 1.00; all personas degrade performance, with E_L causing −39.5%) suggests that persona induction acts as a global modulator of cognitive state. Low-trait conditions (O_L, E_L) consistently impair reasoning across all architectures, indicating that personality representations reconfigure shared computational mechanisms for logical operations.

In contrast, the lower consistency in GPQA (SA ≈ 0.75) and MuSR (SA ≈ 0.75) reveals a boundary: while process-oriented capabilities (reasoning, instruction-following) are governed by architecture-agnostic mechanisms, knowledge-dependent expert understanding is more tightly coupled to model-specific representational space.



(a) Effect size (median $|\Delta\,\text{Acc}|$, pp) across model scales. Bubble size encodes statistical evidence: larger bubbles indicate stronger significance ($-\log_{10} p$).



(b) Persona-specific sensitivity by task

Figure 3: Domain-specific persona effects in Qwen2.5 family (0.5B–14B). Panel (a) integrates effect magnitude (line height) and statistical evidence (bubble size) to reveal both the strength and reliability of persona effects across scales.

RQ2: Domain Specificity

We investigate how persona effects vary across cognitive domains, with an additional lens on parameter scaling (0.5B to 14B) within the Qwen2.5 family. As shown in Figure 3, persona effects exhibit clear domain selectivity consistent with CB5T's prediction of trait-process specificity. Panel (a) presents a dual-metric visualization: the y-axis shows the robust effect size, median absolute accuracy change across persona conditions (in percentage points), while bubble size encodes statistical evidence strength ($-\log_{10} p$, from paired t-tests comparing persona and baseline accuracies). Larger bubbles indicate stronger evidence that persona effects are non-random. Panel (b) provides a fine-grained breakdown of sensitivity by individual persona condition. For instruction-following (IFEVal), both effect size and significance remain consistently high across scales, indicating sustained responsiveness to persona-driven behavioral modulation. Conversely, reasoning-intensive benchmarks (BBH, GSM8K) show non-monotonic trajectories: effect sizes peak at 7B but attenuate sharply at 14B, where small bubbles indicate that residual effects approach noise levels. This pattern suggests that larger models develop more robust reasoning schemas increasingly invariant to persona interventions.

These patterns reveal a developmental dissociation: stylistic flexibility in social-communicative tasks emerges with linguistic complexity, while reasoning capabilities consolidate into schemas increasingly invariant to persona interventions, aligning with human findings that personality traits differentially impact distinct cognitive systems (DeYoung, 2015).

RQ3: Trait-Process Mapping

Table 2 compares dimensions by effect magnitude (Impact) and directional stability (Uniformity) using the high–low polarity gap. **Openness** and **Extraversion** emerge as dominant traits: they induce the largest effects (Impact: 11.96% and 11.70%) and show the most stable direction across models and tasks (Uniformity: 90.5%). This aligns with CB5T predictions linking Openness to cognitive exploration and Extraversion to approach motivation (DeYoung, 2015). **Conscientiousness** exhibits smaller but highly stable effects (Impact: 6.31%; Uniformity: 88.1%), consistent with its theoretical role in goal maintenance. **Agreeableness** shows moderate magnitude with lower stability (Impact: 8.06%; Uniformity: 73.8%). **Neuroticism** has the weakest and least stable influence (Impact: 4.00%; Uniformity: 57.1%) and is the only dimension favoring the low setting (mean Gap: -1.86%), consistent with ACT predictions that anxiety impairs cognitive efficiency (Eysenck et al., 2007).

RQ4: Human-LLM Directional Consistency

We systematically compared LLM persona effects against predictions derived from psychological literature. Based on CB5T and ACT frameworks, we formulated directional hypotheses: Openness should enhance cognitive flexibility (high > low); Conscientiousness should benefit goal-directed performance (high > low); Neuroticism should impair performance (low > high); Extraversion should facilitate approach-oriented tasks (high > low); and Agreeableness effects should be task-dependent. Across trait-benchmark combinations, LLM effects showed **73.68% directional consistency** with human patterns (14/19 comparisons). Openness exhibited the highest consistency (87.5%), with high-O outperforming low-O on 7/8 benchmarks, mirroring the Openness-intelligence association in humans ($r \approx .35$) (Anglim et al., 2022). Conscientiousness also showed strong consistency (87.5%), aligning with its role in goal persistence and sustained attention. Neu-

Table 2: Impact and Uniformity Across Big Five Dimensions

Trait	Impact	Avg. Gap	Uniformity	Rank
Openness (O)	11.96%	+11.17%	90.5%	1
Extraversion (E)	11.70%	+10.42%	90.5%	1
Agreeableness (A)	8.06%	+6.50%	73.8%	4
Conscientiousness (C)	6.31%	+5.82%	88.1%	3
Neuroticism (N)	4.00%	−1.86%	57.1%	5

roticism showed lower consistency (57.1%), though notably, low-N consistently outperformed high-N on anxiety-sensitive tasks requiring sustained attention, consistent with ACT predictions (Eysenck et al., 2007).

This substantial alignment suggests that personality constructs in LLMs capture functional regularities parallel to human trait-cognition relationships. The convergence is particularly striking given that NPTI-induced traits emerge from activation manipulation rather than the developmental and biological processes underlying human personality.

Leveraging Persona-Task Regularities

The preceding analyses reveal that persona effects are strongly task-dependent rather than uniformly beneficial or detrimental. To exploit these structural regularities without additional training, we propose Dynamic Persona Routing (DPR), a lightweight retrieval-based strategy that adapts persona configurations to specific input queries.

Reference Construction For each benchmark dataset $\mathcal{D}$, we partition the data into a reference set $\mathcal{R}$ and a test set $\mathcal{T}$ with a 9:1 ratio using LLaMA-3-8B-Instruct. The reference set serves as a "routing memory," storing the historical performance of all persona conditions on known instances.

Similarity-Based Retrieval For a given test instance $x \in \mathcal{T}$, we identify the most semantically similar historical instance (anchor) x^* from $\mathcal{R}$. We employ TF-IDF vectorization to map text inputs to vector space representations, with the anchor selected via cosine similarity maximization:

$$x^* = \arg\max_{r \in \mathcal{R}} \frac{\text{tf-idf}(x) \cdot \text{tf-idf}(r)}{\| \text{tf-idf}(x) \| \| \text{tf-idf}(r) \|} \tag{6}$$

This ensures that routing decisions are grounded in contextually relevant prior experience.

Routing Strategy We assume that queries with high semantic similarity share similar sensitivities to persona steering. Based on the retrieved anchor x^*, we construct an Effective Persona Set $\mathcal{P}_{eff}(x^*) = \{p \in \mathcal{P} \mid y^{(x^*,p)} = 1\}$, the subset of persona conditions under which the model successfully solved the anchor instance. This set serves as the dynamic recommendation for the current query.

Evaluation We define a "hit" if the recommended set contains at least one persona capable of solving x, comparing against Acc_{best} (best single fixed persona applied globally).

Results As shown in Table 3, dynamic persona selection outperforms the optimal static baseline on most benchmarks. The retrieval-based approach yields the most significant gains on MuSR (+24.57%) and GPQA (+10.31%), demonstrating that semantically similar questions do share persona sensitivities. Moderate improvements appear on MMLU-Pro (+6.47%) and IFEVal (+0.48%). However, on reasoning-intensive tasks (GSM8K, BBH), the dynamic strategy slightly underperforms the best static configuration, suggesting that retrieved personas may occasionally introduce interference in pure logical reasoning where consistency is paramount.

Table 3: Dynamic Persona Routing vs. Best Static Baseline

Dataset	Total	Sampled	Correct	Accuracy (%)	Best Baseline (%)
GPQA	448	44	23	52.27	41.96
BBH	6511	651	404	62.06	63.40
MuSR	756	75	59	78.67	54.10
MMLU-Pro	12032	1203	519	43.14	36.67
IFEVal	541	54	34	62.96	62.48
GSM8K	1319	131	88	67.17	72.18

Analysis The differential effectiveness across domains aligns with our earlier findings: tasks requiring flexible knowledge retrieval and multi-step reasoning (MuSR, GPQA) benefit most from adaptive persona selection, while tasks with more rigid solution structures (GSM8K mathematical reasoning) favor stable configurations. This pattern reinforces the domain-specificity principle established in RQ2.

This proof-of-concept demonstrates that even without training a dedicated routing model, the simple heuristic of "effective personas from similar questions" yields substantial gains on knowledge-intensive and multi-step reasoning tasks, establishing persona control as a lightweight, training-free mechanism for behavioral calibration.

Conclusion

This study systematically evaluates how Big Five traits influence LLM cognitive performance. Four key regularities emerge: (1) **Reliability**: persona effects are consistent across architectures (7B–9B), indicating shared computational mechanisms; (2) **Domain specificity**: personas enhance instruction-following but often impair reasoning, consistent with ACT predictions; (3) **Trait-process mapping**: Openness and Extraversion exert the strongest influence, as predicted by CB5T; (4) **Human-LLM alignment**: 73.68% directional consistency with human trait-cognition relationships. The substantial convergence suggests that NPTI-induced traits capture functional regularities analogous to biological cognition, implying architecture-independent computational structures underlying personality-cognition relationships. Openness and Extraversion's dominant influence aligns with CB5T predictions linking these dimensions to cognitive exploration and approach motivation, while Neuroticism uniquely favoring low settings mirrors ACT predictions regarding anxiety-induced attentional interference. The Dynamic Persona Routing strategy demonstrates that persona induction functions as a behavioral hyperparameter for training-free performance calibration, with substantial gains on knowledge-intensive tasks.

Acknowledgments

The work is supported by the National Natural Science Foundation of China (Nos. 62272092, 62502081) and the Northeastern University Student Innovation and Entrepreneurship Program (No. 260192). Furthermore, we would also like to thank the KinaMind society for their inspiring environment and unwavering support.

References

Anglim, J., Dunlop, P. D., Wee, S., Horwood, S., Wood, J. K., & Marty, A. (2022). Personality and intelligence: A meta-analysis. *Psychological Bulletin*, *148*(5-6), 301–336. https://doi.org/10.1037/bul0000373

Chen, J., Wang, X., Xu, R., Yuan, S., Zhang, Y., Shi, W., Xie, J., Li, S., Yang, R., Zhu, T., Chen, A., Li, N., Chen, L., Hu, C., Wu, S., Ren, S., Fu, Z., & Xiao, Y. (2024). From persona to personalization: A survey on role-playing language agents. *Trans. Mach. Learn. Res.*, *2024*. https://openreview.net/forum?id=xrO70E8UIZ

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., ... Fiedel, N. (2023). Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.*, *24*, 240:1–240:113. https://jmlr.org/papers/v24/22-1144.html

Deng, J., Tang, T., Yin, Y., Yang, W., Zhao, X., & Wen, J. (2025). Neuron based personality trait induction in large language models. *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. https://openreview.net/forum?id=LYHEY783Np

Deshpande, A., Murahari, V., Rajpurohit, T., Kalyan, A., & Narasimhan, K. (2023). Toxicity in chatgpt: Analyzing persona-assigned language models. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the association for computational linguistics: EMNLP 2023, singapore, december 6-10, 2023* (pp. 1236–1270). Association for Computational Linguistics. https://doi.org/10.18653/V1/2023.FINDINGS-EMNLP.88

DeYoung, C. G. (2015). Cybernetic big five theory [Integrative Theories of Personality]. *Journal of Research in Personality*, *56*, 33–58. https://doi.org/https://doi.org/10.1016/j.jrp.2014.07.004

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual review of psychology*, *41*(1), 417–440.

Durmus, E., Nyugen, K., Liao, T. I., Schiefer, N., Askell, A., Bakhtin, A., Chen, C., Hatfield-Dodds, Z., Hernandez, D., Joseph, N., Lovitt, L., McCandlish, S., Sikder, O., Tamkin, A., Thamkul, J., Kaplan, J., Clark, J., & Ganguli, D. (2023). Towards measuring the representation of subjective global opinions in language models. *CoRR*, *abs/2306.16388*. https://doi.org/10.48550/ARXIV.2306.16388

Eysenck, M. W., Derakshan, N., Santos, R., & Calvo, M. G. (2007). Anxiety and cognitive performance: Attentional control theory. *Emotion*, *7*(2), 336–353. https://doi.org/10.1037/1528-3542.7.2.336

Fleming, K. A., Heintzelman, S. J., & Bartholow, B. D. (2016). Specifying associations between conscientiousness and executive functioning: Mental set shifting, not prepotent response inhibition or working memory updating. *Journal of Personality*, *84*(3), 348–360. https://doi.org/10.1111/jopy.12163

Goldberg, L. (1990). An alternative "description of personality": The big-five factor structure. *Journal of personality and social psychology*, *59*(6), 1216–1229. https://doi.org/10.1037//0022-3514.59.6.1216

Lester, B., Al-Rfou, R., & Constant, N. (2021). The power of scale for parameter-efficient prompt tuning. In M. Moens, X. Huang, L. Specia, & S. W. Yih (Eds.), *Proceedings of the 2021 conference on empirical methods in natural language processing, EMNLP 2021, virtual event / punta cana, dominican republic, 7-11 november, 2021* (pp. 3045–3059). Association for Computational Linguistics. https://doi.org/10.18653/V1/2021.EMNLP-MAIN.243

Perez, E., Ringer, S., Lukosiute, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., Jones, A., Chen, A., Mann, B., Israel, B., Seethor, B., McKinnon, C., Olah, C., Yan, D., Amodei, D., ... Kaplan, J. (2023). Discovering language model behaviors with model-written evaluations. In A. Rogers, J. L. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the association for computational linguistics: ACL 2023, toronto, canada, july 9-14, 2023* (pp. 13387–13434). Association for Computational Linguistics. https://doi.org/10.18653/V1/2023.FINDINGS-ACL.847

Salemi, A., Mysore, S., Bendersky, M., & Zamani, H. (2024). Lamp: When large language models meet personalization. In L. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers), ACL 2024, bangkok, thailand, august 11-16, 2024* (pp. 7370–7392). Association for Computational Linguistics. https://doi.org/10.18653/V1/2024.ACL-LONG.399

Sorokovikova, A., Fedorova, N., Rezagholi, S., & Yamshchikov, I. P. (2024). Llms simulate big five personality traits: Further evidence. *CoRR*, *abs/2402.01765*. https://doi.org/10.48550/ARXIV.2402.01765

Tseng, Y., Huang, Y., Hsiao, T., Chen, W., Huang, C., Meng, Y., & Chen, Y. (2024). Two tales of persona in llms: A survey of role-playing and personalization. In Y. Al-Onaizan, M. Bansal, & Y. Chen (Eds.), *Findings of the association for computational linguistics: EMNLP 2024, miami, florida, usa, november 12-16, 2024* (pp. 16612–16631). Association for Computational Linguistics. https://doi.org/10.18653/V1/2024.FINDINGS-EMNLP.969

Xiao, Y., Lin, Y., & Chiu, M. (2024). Behavioral bias of vision-language models: A behavioral finance view. *CoRR*,

abs/2409.15256. https://doi.org/10.48550/ARXIV.2409.15256

Zhang, Z., Rossi, R. A., Kveton, B., Shao, Y., Yang, D., Zamani, H., Dernoncourt, F., Barrow, J., Yu, T., Kim, S., Zhang, R., Gu, J., Derr, T., Chen, H., Wu, J., Chen, X., Wang, Z., Mitra, S., Lipka, N., ... Wang, Y. (2025). Personalization of large language models: A survey. *Trans. Mach. Learn. Res.*, *2025*. https://openreview.net/forum?id=tf6A9EYMo6