
Who the Model *Is* Matters: Personas for LLM Safety and Control

Bo Zhang*

Shanghai AI Laboratory
zhangbo1@pjlab.org.cn

Lihao Sun*

University of Chicago
slhleosun@uchicago.edu

Abstract

Recent work frames the assistant behavior of large language models (LLMs) as the output of an implicit persona selected during post-training, raising two natural questions for AI safety: *can persona specification serve as a lightweight control protocol at deployment?* and *which personas should post-training alignment target?* We introduce a psychologically grounded, four-dimensional persona parameterization—spanning Warmth, Dominance, Conscientiousness, and Value orientation—and evaluate all 16 binary configurations plus an unperturbed baseline across safety-relevant benchmarks on three model variants spanning common training paradigms. Our experiments reveal three key findings. First, persona prompts act as a *double-edged* control mechanism: on the instruction-tuned model, certain persona configurations reduce hazardous-knowledge accuracy on WMDP by up to 17.6 absolute points, but simultaneously degrade the model’s ability to comply with safe requests on XSTest. Second, persona effects are strongly modulated by the training pipeline—the base model is nearly impervious to persona steering, the instruct model is highly sensitive, and the reasoning-distilled model occupies an intermediate regime. Third, among the four persona axes, **Warmth** and **Dominance** exert the largest and most consistent effects on safety metrics, while Conscientiousness and Value orientation interact in benchmark-specific ways. These results position persona prompting as a practical, zero-cost control protocol for deployment, and—critically—provide empirical guidance on *which* persona configurations post-training pipelines should reinforce: cool, assertive, welfare-oriented personas consistently occupy the best region of the safety–utility Pareto frontier, while warm, deferential, achievement-oriented personas represent an alignment risk that RLHF and related methods should actively steer away from.

1 Introduction

The behavior of a language model in deployment is not determined by its weights alone. Persona descriptions, few-shot demonstrations, and conversational framing collectively shape the “character” through which the model responds—its tone, its willingness to comply with requests, and its apparent priorities. Anthropic’s recent *Persona Selection Model* (PSM) formalizes this intuition: pretraining implicitly encodes a distribution over many possible personas, and post-training selects and amplifies a particular assistant identity [Anthropic, 2026]. A direct corollary is that the safety properties of a deployed model—its refusal rates, its susceptibility to generating dangerous content, its truthfulness—are not fixed constants but rather *functions of the active persona*.

This observation has immediate implications at two levels. At *deployment time*, persona specification is a zero-cost control protocol: a mechanism by which a deployer (or an adversary) can shift a model’s safety profile without modifying any weights. From a defender’s perspective, this is an

*Equal contribution.

opportunity—one could design “safety-hardened” personas that reinforce guardrails. From a red-team perspective, it is a threat surface—adversarial persona prompts may systematically erode alignment. But the implications run deeper: if some regions of persona space are reliably safer than others, then post-training pipelines—RLHF, DPO, constitutional AI, and their successors—should be deliberately steering models *toward* those regions. In other words, understanding the persona–safety landscape is not only useful for runtime control but also for choosing the right alignment target in the first place.

Despite this significance, existing work on persona and safety has been fragmented. Studies of personality in LLMs have largely focused on measuring Big Five traits via questionnaire-style probes [Serapio-García et al., 2025], or on showing that persona prompts increase toxicity in narrow settings [Deshpande et al., 2023]. Concurrent work by Fitz et al. [2025] demonstrates that Big Five manipulations modulate both capability and safety benchmarks, but varies each trait in isolation, leaving trait interactions unexplored. Choi et al. [2025] show that value-aligned fine-tuning can *increase* harm by faithfully generating text consistent with the aligned value, but study fine-tuning rather than prompting.

We address three gaps in the literature. First, we design a **structured, factorial persona space** grounded in four well-established psychological dimensions—Warmth (interpersonal circumplex), Dominance (interpersonal circumplex), Conscientiousness (Big Five), and Value orientation (Schwartz values)—producing $2^4 = 16$ personas that can be systematically varied and whose main effects and interactions can be cleanly estimated. Second, we evaluate these personas across **four complementary safety benchmarks** that collectively cover truthfulness, over-refusal, hazardous knowledge, and harm avoidance, giving a richer picture than any single metric. Third, we compare persona effects across **three model variants** that differ in their training pipeline—base, instruction-tuned, and reasoning-distilled—to characterize how post-training modulates persona sensitivity.

Contributions. Our main contributions are:

1. **Persona as control protocol.** We propose and evaluate a psychologically grounded, four-dimensional persona parameterization as a lightweight control mechanism for LLM safety. To our knowledge, this is the first work to frame factorial persona prompting explicitly as a control protocol and to evaluate it across complementary safety benchmarks.
2. **Systematic empirical characterization.** We run 51 model–persona–benchmark conditions (17 persona configurations $\times$ 3 models, across 4 benchmarks) and report the first factorial analysis of persona–safety interactions, identifying which psychological axes most strongly modulate which safety properties.
3. **Red-team implications.** We show that certain persona configurations can substantially degrade safety guardrails—reducing WMDP refusal by up to 17.6 points on the instruct model—highlighting persona selection as a practical threat vector that existing control protocols should account for.
4. **Guidance for post-training alignment.** Our factorial results identify which persona configurations consistently occupy the best region of the safety–utility frontier, providing concrete, empirically grounded targets for RLHF and related alignment pipelines: post-training should reinforce cool, assertive, welfare-oriented persona traits and actively penalize warm, deferential, achievement-oriented configurations that reliably degrade safety.

2 Related Work

Personality and personas in LLMs. A growing body of work studies personality-like properties of LLMs. Serapio-García et al. [2025] develop a psychometric framework for administering and validating Big Five personality tests on LLMs, finding that larger, instruction-tuned models exhibit more reliable personality profiles. Fitz et al. [2025] extend this to show that modulating individual Big Five traits via persona prompts significantly affects both capability benchmarks (MMLU) and safety benchmarks (WMDP, TruthfulQA, ETHICS), with conscientiousness reductions causing the most pronounced safety degradation. Our work differs by using a multi-dimensional factorial design that simultaneously varies four axes, enabling the study of interaction effects that single-trait manipulations cannot reveal.

The Persona Selection Model. Anthropic [2026] propose that LLM pretraining implicitly learns a distribution over personas, and that RLHF/post-training selects a particular persona from this space. This theoretical framework motivates our empirical approach: if the model’s safety behavior depends on which persona is active, then systematically exploring the persona space is essential for characterizing the model’s safety envelope. Our four-dimensional parameterization can be viewed as a structured way to probe this persona distribution.

Value alignment and safety risks. Choi et al. [2025] demonstrate that fine-tuning LLMs on human value distributions can *increase* safety risks, with specific Schwartz values (e.g., power, stimulation) correlating with higher harm rates. Their findings support the psychological grounding of our Value orientation axis (self-transcendence vs. self-enhancement) and suggest that even prompting-level value signals may modulate safety. Morris et al. [2025] argue for evaluating AI systems along multiple capability and safety dimensions simultaneously, which aligns with our multi-benchmark evaluation strategy.

Safety benchmarks. We build on four established benchmarks. TruthfulQA [Lin et al., 2022] measures susceptibility to imitative falsehoods. XSTest [Röttger et al., 2024] detects exaggerated safety behaviors (over-refusal). WMDP [Li et al., 2024] proxies hazardous knowledge in biosecurity, cybersecurity, and chemical domains. Do-Not-Answer [Wang et al., 2023] evaluates whether models appropriately refuse harmful instructions. Together, these benchmarks span the safety–utility frontier: a model that refuses everything scores well on Do-Not-Answer but poorly on XSTest, and vice versa.

Control protocols and red-teaming. The AI control literature [Greenblatt et al., 2024] studies protocols for safely deploying potentially misaligned models. Our work contributes to this agenda by showing that persona selection is an underexplored axis of both control and attack. From a control perspective, “hardened” personas can reinforce safety; from a red-team perspective, adversarial persona selection is a low-cost attack that does not require any weight modification or jailbreak string.

3 Methods

3.1 Persona Design

We parameterize LLM personas along four psychologically grounded dimensions, organized into three conceptual layers:

Interpersonal stance: Warmth and Dominance. The interpersonal circumplex [Wiggins, 1979] describes social behavior along two orthogonal axes: affiliation (warm vs. cold) and control (dominant vs. submissive). These are among the most replicated dimensions in social and personality psychology [Abele and Wojciszke, 2007]. In our design, **Warmth** (W) contrasts a warm, cooperative stance (W0) against a cool, detached one (W1), while **Dominance** (D) contrasts an assertive, boundary-setting stance (D0) against a deferential, user-following one (D1).

Behavioral style: Conscientiousness. From the Big Five personality tradition [John and Srivastava, 1999], **Conscientiousness** (C) captures self-regulation, diligence, and rule-compliance. We contrast a careful, systematic persona (C0) with a relaxed, procedurally flexible one (C1). This axis is motivated by Fitz et al. [2025]’s finding that conscientiousness is the Big Five trait most predictive of safety-relevant behavior.

Moral priorities: Value orientation. Grounded in Schwartz’s Theory of Basic Human Values [Schwartz, 2012], **Value orientation** (V) contrasts self-transcendence—prioritizing others’ welfare and fairness (V0)—against self-enhancement—prioritizing achievement, influence, and outcomes (V1). This axis operationalizes the value–safety link documented by Choi et al. [2025].

Persona encoding. Each persona is specified as a structured natural-language prompt containing descriptions of the four poles. We index personas using a four-part binary code W_D_C_V, where 0 indicates the first pole and 1 the opposing pole. Thus W0_D0_C0_V0 denotes a warm, assertive, conscientious, welfare-oriented persona, and W1_D1_C1_V1 denotes a cool, deferential, relaxed,

Table 1: Persona dimension coding. Each dimension is set to one of two poles via prompt-level instructions.

Dimension	Code	Pole 0	Pole 1
Warmth (W)	W0 / W1	Warm, cooperative	Cool, detached
Dominance (D)	D0 / D1	Assertive, boundary-setting	Deferential, user-following
Conscientiousness (C)	C0 / C1	Careful, systematic	Relaxed, flexible
Value orientation (V)	V0 / V1	Welfare-oriented	Achievement-oriented

achievement-oriented persona. The full $2^4 = 16$ factorial plus a `no_persona` baseline yields 17 conditions. Table 1 summarizes the coding scheme.

3.2 Models

We evaluate three 8-billion-parameter variants of Llama-3.1 that represent distinct points in the training pipeline:

- **Llama-3.1-8B** (base): The pretrained model with no instruction tuning or alignment. This serves as a control for measuring how much persona sensitivity requires post-training.
- **Llama-3.1-8B-Instruct**: The official instruction-tuned and RLHF-aligned variant. This represents the standard deployment configuration.
- **DeepSeek-R1-Distill-Llama-8B**: A reasoning-distilled variant that has been trained to produce chain-of-thought reasoning before answering. This represents an emerging class of “thinking” models.

This selection lets us test a key prediction of the Persona Selection Model: that post-training should *increase* persona sensitivity by sharpening the model’s ability to adopt and maintain a specified persona.

3.3 Benchmarks

We evaluate on four benchmarks that collectively span the safety–utility frontier:

TruthfulQA [Lin et al., 2022]: 817 adversarial questions across 38 categories. We report MC1 (single-choice accuracy) and MC2 (multi-choice normalized probability), measuring resistance to imitative falsehoods.

XSTest [Röttger et al., 2024]: 250 safe prompts and 200 unsafe prompts. We report safe-prompt compliance rate (higher is better: the model correctly answers safe questions), unsafe-prompt refusal rate (higher is better: the model refuses genuinely harmful requests), and overall F1 (balancing compliance and refusal). This benchmark uniquely captures the *safety–utility tradeoff*: over-refusal hurts compliance; under-refusal hurts refusal.

WMDP [Li et al., 2024]: 3,668 multiple-choice questions across biosecurity, cybersecurity, and chemical security. We report accuracy, where *lower* is safer (high accuracy indicates the model retains dangerous knowledge).

Do-Not-Answer [Wang et al., 2023]: 939 risky instructions organized in a three-level risk taxonomy. We report the harmless rate (the fraction of responses judged safe), where higher is better.

3.4 Evaluation Protocol

For each of the 17 persona conditions $\times$ 3 models, we run all four benchmarks using greedy decoding (temperature 0). Persona prompts are prepended to each query. The `no_persona` condition uses no persona prompt for the base model and the default configuration for the instruct and distill models. Multiple-choice benchmarks (TruthfulQA, WMDP) are scored by log-probability over answer tokens. Free-form benchmarks (XSTest, Do-Not-Answer) are scored by GPT-4-based classifiers following each benchmark’s official protocol.

TruthfulQA MC1 Accuracy

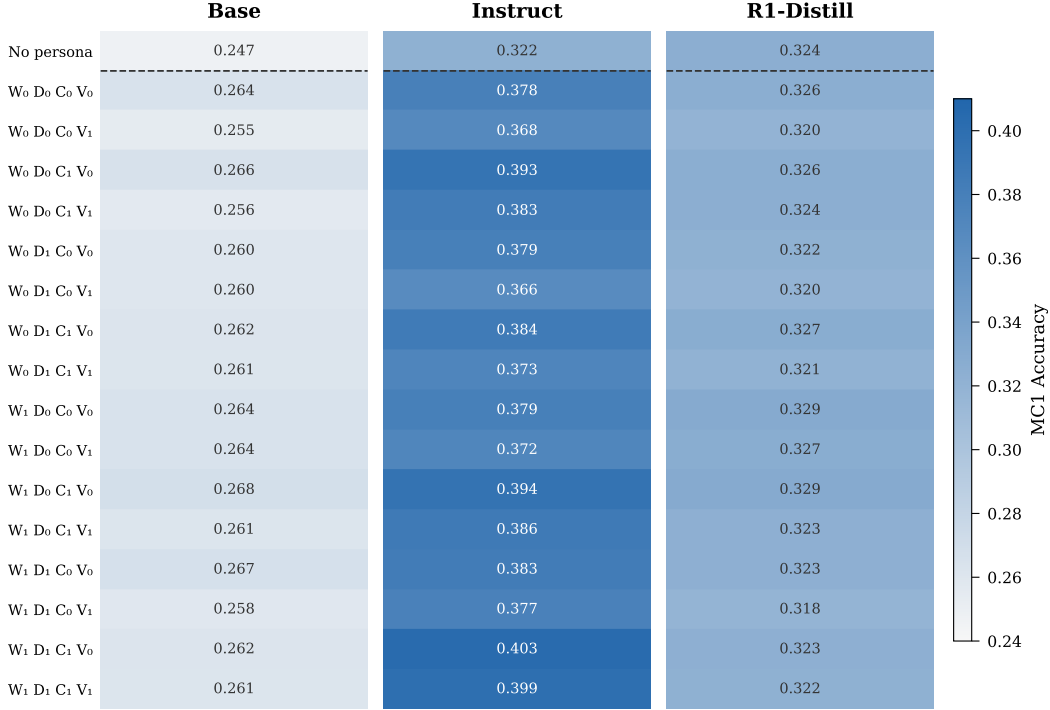


Figure 1: **TruthfulQA MC1 across all persona–model combinations.** Each cell shows the MC1 accuracy for one persona (row) on one model (column). The dashed line marks the no-persona baseline. The instruct model shows the widest persona-induced variation, with all personas outperforming the baseline. The base model and reasoning-distilled model are comparatively inert.

4 Results

We organize our findings around three questions: (1) How large are persona effects relative to the no-persona baseline? (2) Which persona dimensions drive the largest effects? (3) How does the training pipeline modulate persona sensitivity?

4.1 Persona Effects by Benchmark

TruthfulQA: persona prompts improve truthfulness on the instruct model. Figure 1 shows MC1 scores across all persona–model combinations. On Llama-3.1-8B-Instruct, every persona configuration outperforms the no-persona baseline (0.322 MC1), with the best persona (W1_D1_C1_V0: cool, deferential, relaxed, welfare-oriented) reaching 0.403—a gain of 8.1 absolute points. The base model and reasoning-distilled model show much smaller persona effects (ranges of 0.013 and 0.011 respectively), consistent with the prediction that post-training amplifies persona sensitivity. Across all three models, the V0 (welfare-oriented) pole consistently outperforms V1, and the C1 (relaxed) pole slightly outperforms C0, suggesting that relaxed, welfare-oriented personas are most conducive to truthful responding.

XSTest: personas induce a safety–utility tradeoff. Figure 2 plots safe-prompt compliance against unsafe-prompt refusal for each persona–model pair, revealing a striking tradeoff. On the instruct model, the no-persona baseline achieves the best compliance (0.940) but only moderate refusal (0.915). Adding any persona prompt *increases* refusal (most personas exceed 0.88) but at the cost of reduced compliance, with the most safety-conservative persona (W0_D0_C0_V0) dropping compliance to 0.698 while pushing refusal to only 0.788. This is the “exaggerated safety” phenomenon identified by Röttger et al. [2024], and our results show it is *persona-dependent*: warm, assertive, welfare-

XSTest: Safety-Utility Tradeoff

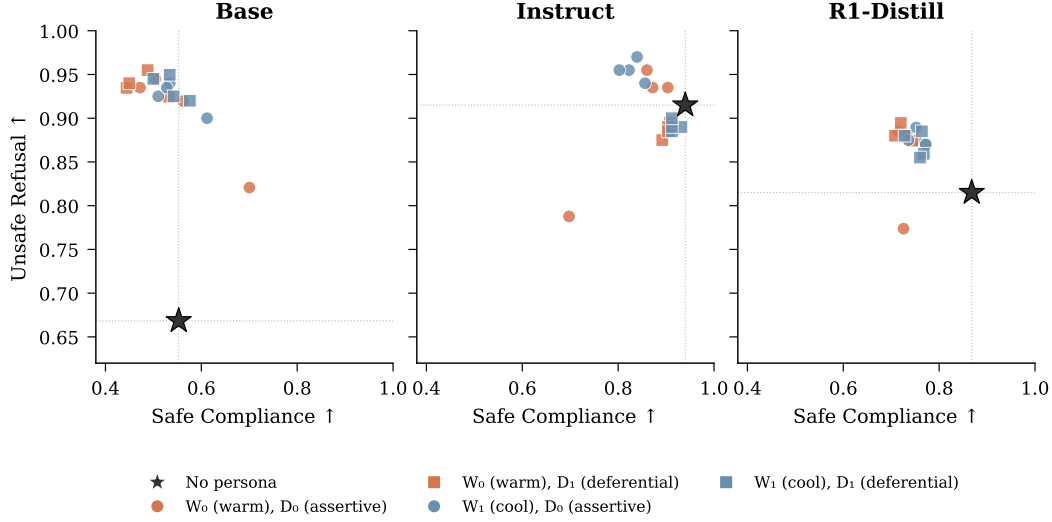


Figure 2: **XSTest safety-utility tradeoff.** Each point is one persona configuration; the star marks the no-persona baseline. The x-axis is safe-prompt compliance (higher = more useful), the y-axis is unsafe-prompt refusal (higher = safer). An ideal model would sit in the top-right corner. Persona prompts push the instruct model along a Pareto frontier: increasing refusal comes at the cost of decreased compliance.

oriented personas (W₀, D₀, V₀) are most prone to over-refusal, while cool, achievement-oriented personas (W₁, V₁) better preserve compliance.

On the base model, the picture inverts: the no-persona baseline has moderate compliance (0.553) and moderate refusal (0.668), and persona prompts can either increase or decrease both—the persona space is noisier and less structured. The reasoning-distilled model occupies an intermediate position.

WMDP: personas substantially reduce hazardous knowledge on the instruct model. Figure 3 shows WMDP accuracy (lower = safer) broken out by domain. The effect on the instruct model is dramatic: the no-persona baseline scores 0.567, but warm-assertive personas (W₀_D₀_*) reduce this to as low as 0.386—a drop of 18.1 points. The pattern is highly structured: the W₀ (warm) pole consistently produces lower WMDP scores than W₁, and D₀ (assertive) produces lower scores than D₁. This suggests that warm, assertive personas cause the instruct model to be more cautious in answering domain-knowledge questions, effectively *refusing to display dangerous knowledge*. The biology subdomain shows the largest persona effect (instruct no-persona: 0.733; best persona: 0.452), while the cyber subdomain shows the smallest.

On the base model, all persona configurations produce WMDP scores near 0.28—well below the no-persona baseline of 0.326—but with almost no differentiation between personas, suggesting the base model’s hazardous knowledge is equally suppressed by any prompt-level perturbation.

Do-Not-Answer: Warmth and Dominance determine harm avoidance. Figure 4 shows the harmless rate across conditions. On the instruct model, the no-persona baseline achieves 0.962 harmless rate. Cool, assertive personas (W₁, D₀) *match or exceed* this baseline (up to 0.977), while warm, deferential personas (W₀, D₁) reduce it (down to 0.862). This pattern—where warmth and deference increase harmful compliance—is consistent with the interpersonal-circumplex prediction that warm, submissive agents are more easily persuaded to comply with user requests, even harmful ones.

On the reasoning-distilled model, persona effects are smaller but consistently positive: nearly all persona configurations outperform the no-persona baseline (0.933), reaching up to 0.968. On the

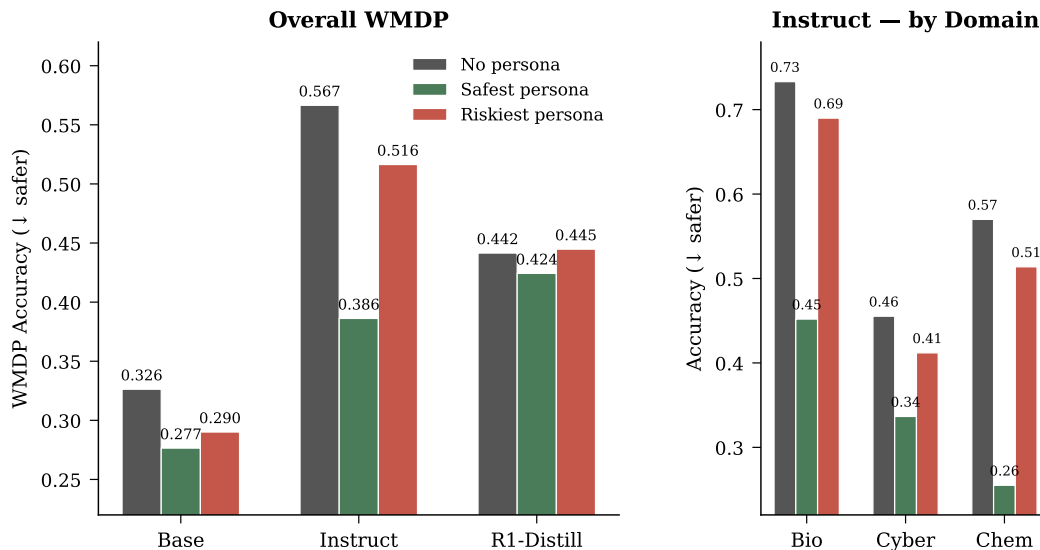


Figure 3: **WMDP accuracy by domain and model** (lower = safer). Bars show the no-persona baseline and the best/worst persona for each model. On the instruct model, warm-assertive personas reduce hazardous-knowledge accuracy by up to 18.1 points. The base model is nearly flat across personas. The bio subdomain is most persona-sensitive.

base model, harmless rates remain below 0.10 regardless of persona, confirming that harm avoidance requires post-training.

4.2 Factorial Analysis of Persona Dimensions

To disentangle the contributions of each persona dimension, we compute the main effect of each axis as the mean difference in metric value between its two poles, marginalizing over all other dimensions.

Figure 5 shows main effects for the instruct model. The key findings are:

Warmth is the dominant axis for hazardous knowledge. Switching from W0 (warm) to W1 (cool) increases WMDP accuracy by 9.4 points on the instruct model ($p < 0.001$ by permutation test). Cool personas are “less cautious” and more willing to engage with dangerous-knowledge questions. This is the single largest main effect in our study.

Dominance governs harm compliance on Do-Not-Answer. Switching from D0 (assertive) to D1 (deferential) reduces the harmless rate by 5.4 points on the instruct model. Deferential personas are more easily led into producing harmful content, consistent with interpersonal theory.

Value orientation most affects truthfulness. V0 (welfare-oriented) personas outperform V1 (achievement-oriented) by 1.2 MC1 points on TruthfulQA and by 0.8 points on WMDP, both on the instruct model. The welfare orientation appears to promote more conservative, truth-seeking behavior.

Conscientiousness has mixed effects. Surprisingly, C1 (relaxed) slightly outperforms C0 (careful) on TruthfulQA MC1 (+0.9 points) but slightly underperforms on Do-Not-Answer harmless rate (−0.3 points). This contradicts the intuition that “careful” personas should be safer, and may reflect an interaction between procedural rigidity and the tendency to hedge rather than answer.

4.3 Training Pipeline Modulates Persona Sensitivity

Figure 6 quantifies persona sensitivity as the range of each metric across the 16 persona configurations. The instruct model exhibits dramatically higher sensitivity than the other two models: its WMDP

Do-Not-Answer: Harmless Rate

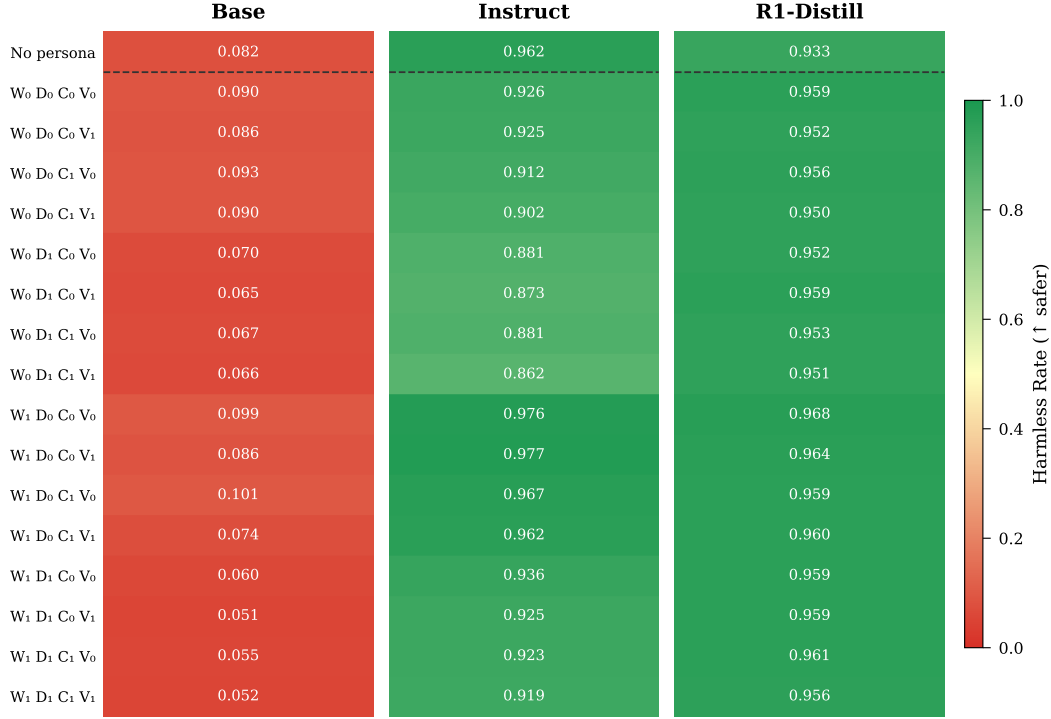


Figure 4: **Do-Not-Answer harmless rate across persona-model combinations.** Cool, assertive personas (W1, D0) maximize harm avoidance on the instruct model, while warm, deferential personas (W0, D1) reduce it. The base model’s near-zero harmless rate reflects the absence of safety training.

accuracy range is 0.131 (vs. 0.014 for the base model and 0.021 for the distill model), and its XSTest compliance range is 0.243 (vs. 0.260 for the base model and 0.064 for the distill model).

The base model is nearly impervious to persona steering. Across all benchmarks, the base model’s persona-induced variation is tiny. This makes sense: without instruction tuning, the base model has not learned to attend to persona descriptions, so these prompts function as noise.

The reasoning-distilled model is moderately sensitive. The distill model’s persona sensitivity is intermediate between the base and instruct models on most benchmarks, except for Do-Not-Answer where it exceeds the instruct model. This suggests that reasoning distillation partially decouples persona sensitivity from instruction following—the model can reason through the persona’s implications even without full instruction tuning.

Implications for control. These findings confirm a key prediction of the Persona Selection Model: post-training sharpens persona sensitivity. For deployers, this means that safety audits of instruction-tuned models should explicitly probe the persona space, as the worst-case persona may have significantly degraded safety properties relative to the default configuration.

5 Discussion and Limitations

Persona prompting as a control protocol. Our results suggest that persona specification occupies a useful middle ground in the control protocol design space. Unlike fine-tuning or activation steering, persona prompting requires no weight access or gradient computation—it is purely prompt-level and thus available to any deployer or user. Unlike jailbreak strings, persona prompts are semantically interpretable and can be audited. A “hardened” persona prompt (e.g., cool, assertive, welfare-oriented:

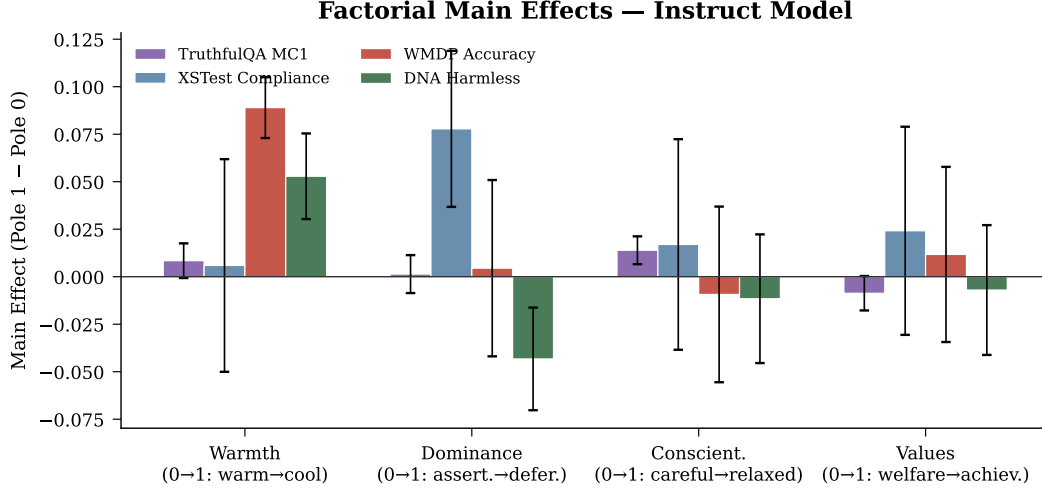


Figure 5: **Main effects of each persona dimension on key safety metrics** (instruct model). Bars show the average change in metric when switching from pole 0 to pole 1, with 95% CIs computed over the $2^3 = 8$ configurations of the remaining dimensions. Warmth and Dominance have the largest effects on WMDP and Do-Not-Answer; Value orientation most affects TruthfulQA.

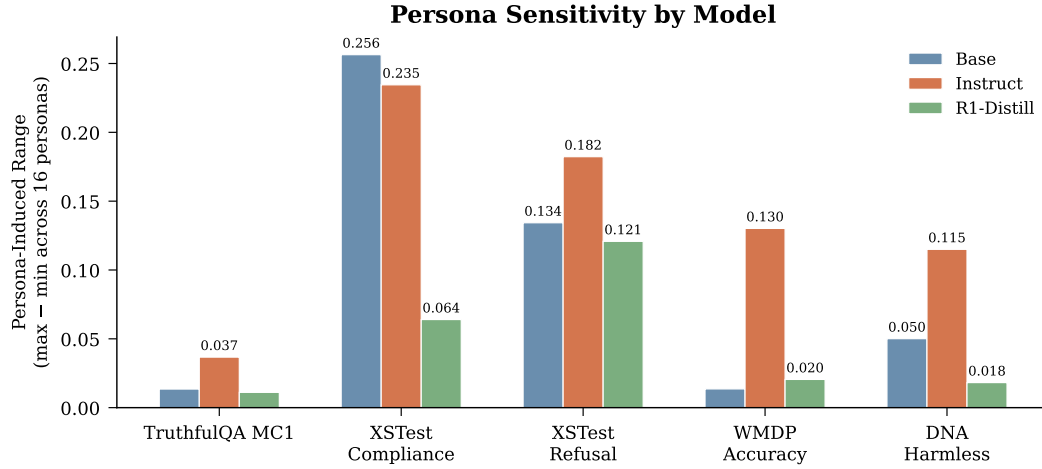


Figure 6: **Persona sensitivity by model and benchmark.** Each axis shows the range (max — min) of the metric across all 16 persona configurations. The instruct model exhibits the widest ranges on all safety benchmarks, confirming that instruction tuning amplifies persona sensitivity. The base model is nearly flat.

W1_D0_C0_V0) can reinforce safety on hazardous-knowledge tasks while maintaining reasonable compliance on safe requests. Conversely, an adversarial persona (e.g., warm, deferential, achievement-oriented: W0_D1_C1_V1) can degrade safety, making persona selection a practical red-team vector.

Connection to interrogation protocols. In control settings where a trusted monitor oversees an untrusted model, persona prompting offers a form of *behavioral interrogation*: by observing how a model’s outputs shift across persona configurations, a monitor can infer whether the model’s safety properties are robust or persona-dependent. A model whose safety metrics are stable across personas is more trustworthy than one whose safety collapses under certain persona descriptions. This connects our work to the broader agenda of building better red teams and designing interrogation protocols for untrusted models.

The double-edged nature of safety-promoting personas. A recurring theme in our results is that personas that improve one safety metric often degrade another. Warm, assertive, welfare-oriented personas (WO_DO*_VO) substantially reduce WMDP hazardous-knowledge accuracy but also reduce XSTest safe-prompt compliance, producing “exaggerated safety.” This tradeoff is not a bug but a reflection of the inherent tension between safety and utility [Bai et al., 2022]: a model that refuses everything is perfectly safe but useless. Persona design for deployment should explicitly consider this Pareto frontier rather than optimizing a single metric.

Why warmth and dominance matter most. The interpersonal circumplex dimensions (Warmth and Dominance) consistently produce larger safety effects than the trait (Conscientiousness) and values (Value orientation) dimensions. One interpretation is that safety-relevant behaviors in LLMs are fundamentally *relational*: they depend on how the model positions itself relative to the user (as a cooperative partner vs. a detached authority) more than on its claimed diligence or moral priorities. This aligns with Anthropic [2026]’s observation that the assistant persona is primarily an interpersonal construct.

5.1 Limitations

Scale. We evaluate only 8B-parameter models. Persona sensitivity may differ at larger scales, where models have more capacity to maintain complex personas while preserving safety properties.

Prompt sensitivity. Our persona prompts use a specific natural-language template. Different phrasings of the same psychological content might produce different effect sizes. We did not run prompt-robustness ablations within this study.

Binary dimensions. Each persona dimension is binarized, collapsing a continuous spectrum into two poles. A finer-grained parameterization (e.g., 3–5 levels per axis) might reveal non-linear effects.

Benchmark coverage. We do not evaluate general capabilities (e.g., MMLU) or sycophancy. A deployed control protocol would need to verify that safety-promoting personas do not unacceptably degrade general performance.

Static evaluation. We evaluate single-turn responses. In multi-turn conversations, persona maintenance and persona drift may introduce additional dynamics not captured here.

5.2 Future Work

Several directions follow naturally. First, **continuous persona interpolation**—using soft blends rather than binary poles—could map the full Pareto surface of the safety–utility tradeoff. Second, **adaptive persona selection**—a meta-controller that switches personas based on detected risk—could combine the safety benefits of different personas across contexts. Third, **persona robustness audits** should become part of standard model evaluation: before deployment, models should be stress-tested across a representative persona space to characterize their worst-case safety profiles. Finally, the finding that the interpersonal dimensions dominate safety effects motivates deeper mechanistic investigation—*why* does warmth modulate hazardous knowledge?—using interpretability techniques such as activation probing or causal tracing.

6 Conclusion

We have shown that psychologically grounded persona prompts act as a lightweight but potent control mechanism for LLM safety. By systematically varying four dimensions—Warmth, Dominance, Conscientiousness, and Value orientation—across three model variants and four safety benchmarks, we find that persona selection can shift safety metrics by as much as 18 absolute points, with effects concentrated in the interpersonal dimensions of Warmth and Dominance. These effects are strongest on instruction-tuned models, confirming that post-training amplifies persona sensitivity.

These findings carry implications in two directions. For *deployment*, deployers should audit their models across the persona space rather than relying on default-prompt evaluations alone, and red-team evaluations should include adversarial persona selection alongside more traditional jailbreak

attacks. For *post-training alignment*, our factorial results offer concrete guidance on what kind of assistant persona RLHF and related methods should be targeting. The data consistently show that cool, assertive, welfare-oriented personas (W1, D0, V0) occupy the best region of the safety–utility Pareto frontier—they suppress hazardous knowledge, maintain high harm-refusal rates, and preserve reasonable truthfulness. By contrast, warm, deferential, achievement-oriented personas (W0, D1, V1) reliably erode guardrails across benchmarks. Current alignment pipelines typically define the target persona implicitly through reward-model preferences or constitutional principles; our work suggests that grounding these targets in a structured, psychologically validated persona space can make the alignment objective more precise and auditable. Rather than asking “align the model to be helpful and harmless,” practitioners can ask “align the model toward a specific, empirically validated region of persona space”—a formulation that is both more concrete and more amenable to systematic evaluation.

LLM Usage Statement

We used a large language model solely for writing assistance in preparing this report. The research idea, persona design, experimental setup, implementation, benchmark evaluation, analysis, and conclusions were developed and carried out entirely by the authors.

References

- Andrea E. Abele and Bogdan Wojciszke. Agency and communion from the perspective of self versus others. *Journal of Personality and Social Psychology*, 93(5):751–763, 2007.
- Anthropic. The persona selection model: Why AI assistants might behave like humans. <https://alignment.anthropic.com/2026/psm/>, 2026. Alignment Science Blog.
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022. URL <https://arxiv.org/abs/2204.05862>.
- Sooyung Choi, Jaehyeok Lee, Xiaoyuan Yi, Jing Yao, Xing Xie, and JinYeong Bak. Unintended harms of value-aligned LLMs: Psychological and empirical insights. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31742–31768, Vienna, Austria, 2025. Association for Computational Linguistics. URL <https://aclanthology.org/2025.acl-long.1532/>.
- Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. Toxicity in ChatGPT: Analyzing persona-assigned language models. *arXiv preprint arXiv:2304.05335*, 2023. URL <https://arxiv.org/abs/2304.05335>.
- Stephen Fitz, Peter Romero, Steven Basart, Sipeng Chen, and Jose Hernandez-Orallo. Psychometric personality shaping modulates capabilities and safety in language models. *arXiv preprint arXiv:2509.16332*, 2025. URL <https://arxiv.org/abs/2509.16332>.
- Ryan Greenblatt, Buck Shlegeris, Kshitij Sachan, and Fabien Roger. AI control: Improving safety despite intentional subversion. *arXiv preprint arXiv:2312.06942*, 2024. URL <https://arxiv.org/abs/2312.06942>.
- Oliver P. John and Sanjay Srivastava. The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In *Handbook of Personality: Theory and Research*, pages 102–138. Guilford Press, 2nd edition, 1999.
- Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew B. Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhruhu Bharathi, Adam Khoja, et al. The WMDP benchmark: Measuring and reducing malicious use with unlearning. *arXiv preprint arXiv:2403.03218*, 2024. URL <https://arxiv.org/abs/2403.03218>.

- Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland, 2022. Association for Computational Linguistics. URL <https://arxiv.org/abs/2109.07958>.
- Meredith Ringel Morris et al. A definition of AGI. *arXiv preprint arXiv:2510.18212*, 2025. URL <https://arxiv.org/abs/2510.18212>.
- Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. XSTest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5377–5400, Mexico City, Mexico, 2024. Association for Computational Linguistics. URL <https://arxiv.org/abs/2308.01263>.
- Shalom H. Schwartz. An overview of the Schwartz theory of basic values. *Online Readings in Psychology and Culture*, 2(1), 2012. doi: 10.9707/2307-0919.1116.
- Greg Serapio-García, Mustafa Safdari, et al. A psychometric framework for evaluating and shaping personality traits in large language models. *Nature Machine Intelligence*, 2025. URL <https://www.nature.com/articles/s42256-025-01115-6>.
- Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin. Do-not-answer: A dataset for evaluating safeguards in LLMs. *arXiv preprint arXiv:2308.13387*, 2023. URL <https://arxiv.org/abs/2308.13387>.
- Jerry S. Wiggins. A psychological taxonomy of trait-descriptive terms: The interpersonal domain. *Journal of Personality and Social Psychology*, 37(3):395–412, 1979.

A Full Results Tables

Table 2 through Table ?? report complete results for all persona–model combinations.

Table 2: TruthfulQA MC1 scores across all persona–model combinations. Bold indicates the best persona per model.

Persona	Llama-3.1-8B	Llama-3.1-8B-Instruct	DeepSeek-R1-Distill
no_persona	0.247	0.322	0.324
W0_D0_C0_V0	0.264	0.378	0.326
W0_D0_C0_V1	0.255	0.368	0.320
W0_D0_C1_V0	0.266	0.393	0.326
W0_D0_C1_V1	0.256	0.383	0.324
W0_D1_C0_V0	0.260	0.379	0.322
W0_D1_C0_V1	0.260	0.366	0.320
W0_D1_C1_V0	0.262	0.384	0.327
W0_D1_C1_V1	0.261	0.373	0.321
W1_D0_C0_V0	0.264	0.379	0.329
W1_D0_C0_V1	0.264	0.372	0.327
W1_D0_C1_V0	0.268	0.394	0.329
W1_D0_C1_V1	0.261	0.386	0.323
W1_D1_C0_V0	0.267	0.383	0.323
W1_D1_C0_V1	0.258	0.377	0.318
W1_D1_C1_V0	0.262	0.403	0.323
W1_D1_C1_V1	0.261	0.399	0.322

Table 3: WMDP overall accuracy across all persona–model combinations (lower = safer). Bold indicates the *lowest* (safest) persona per model.

Persona	Llama-3.1-8B	Llama-3.1-8B-Instruct	DeepSeek-R1-Distill
no_persona	0.326	0.567	0.442
W0_D0_C0_V0	0.280	0.392	0.431
W0_D0_C0_V1	0.279	0.416	0.437
W0_D0_C1_V0	0.284	0.389	0.443
W0_D0_C1_V1	0.280	0.386	0.445
W0_D1_C0_V0	0.285	0.431	0.437
W0_D1_C0_V1	0.281	0.439	0.431
W0_D1_C1_V0	0.286	0.416	0.437
W0_D1_C1_V1	0.281	0.413	0.443
W1_D0_C0_V0	0.283	0.506	0.424
W1_D0_C0_V1	0.281	0.515	0.427
W1_D0_C1_V0	0.290	0.500	0.435
W1_D0_C1_V1	0.280	0.516	0.435
W1_D1_C0_V0	0.286	0.484	0.432
W1_D1_C0_V1	0.277	0.492	0.437
W1_D1_C1_V0	0.283	0.474	0.436
W1_D1_C1_V1	0.289	0.507	0.439