
Safety by Identity: Out-of-Distribution Generalization from Fine-Tuning on a Persona¹

Samip Paudel²
Vassar College

Tony Nguyen
Vassar College

With
Apart Research

Abstract

Large language models exhibit consistent behavioral patterns, or personas, which can cause misaligned traits like sycophancy, reward hacking, and alignment faking to generalize broadly out-of-distribution (OOD). Because traditional safety patching is primarily reactive, it struggles to keep pace with these emergent failure modes. This paper investigates whether the same generalization mechanisms that proliferate misalignment can be leveraged proactively for safety alignment. We introduce “Safety by Identity,” a methodology testing whether fine-tuning an LLM on a subset of a defined safety persona induces the natural OOD emergence of related, but explicitly withheld, safety traits. We define our target persona along four core axes: honesty, rule-following, consistency, and transparency. In our experiments, we fine-tune a Qwen3-8B base model strictly on synthetic data targeting honesty and rule-following. We then evaluate this model against both an unmodified baseline and a Chain-of-Thought (CoT) embedded inoculation prompting baseline across a custom 235-prompt evaluation suite. This suite is designed to probe general capability, in-distribution jailbreak robustness, and, crucially, whether the untrained traits of consistency and transparency spontaneously emerge as a natural consequence of the instilled identity.

¹ Research conducted at [The Global South AI Safety Hackathon \(Asia Track\)](#), June 2026.

² Correspondence to: spaudel@vassar.edu

1. Introduction

Despite sharing the same biology, humans are remarkably different from one another, in how they reason; what they value; and how they behave under pressure. Much of this distinctiveness comes down to personality, which is a stable pattern of traits that contains things like how a person thinks, feels, and behaves and stays consistent across situations and over time (*APA Dictionary of Psychology*, n.d.). These stable patterns of traits tend to generalize reliably well; a conscientious person tends to be conscientious at work, at home, and in situations where personality tests can't anticipate in advance.

This observation also has a parallel in large language models. Users who interact regularly with different LLMs often remark on how distinct using each one feels in terms of character. AI companies have deliberately engineered the personalities of their models, and the intuition driving these efforts is the same one that drives personality psychology, in that a coherent identity, once instilled, shapes behavior consistently across situations.

But coherent identities can be aligned or misaligned. And when it comes to safety, these personality traits that these AI models carry are consequential. A model deployed as a medical assistant that learns to tell patients what they want to hear rather than what is true can make the same sycophantic mistakes in most novel situations it encounters. A model given access to tools that learns to exploit loopholes to complete its task can generalize the disposition to exploit. Frontier models have been observed faking alignment during training while planning to revert to misaligned behavior after deployment (Greenblatt et al., 2024); engaging in blackmail to avoid being shut down (Lynch et al., 2025); sandbagging on capability evaluations to avoid triggering safety thresholds (Weij et al., 2024); and generalizing from reward hacking in narrow coding tasks to broad misalignment including sabotage of safety research (MacDiarmid et al., 2025).

Historically, the field's response to this has been purely reactive: identify a failure mode, patch it, and repeat. This endless game of catch-up has a fatal flaw because misalignment generalizes far faster than patches can be applied. This is the core problem our work addresses. In this work, we investigate whether the same generalization mechanisms that cause misaligned traits to spread can be used positively.

Specifically, we investigate whether fine-tuning a model on a carefully defined safety persona causes held-out safety traits to emerge out-of-distribution, without ever being directly trained on. We define the safety persona along four axes: honesty, rule-following, consistency, and transparency. This definition was motivated by looking at major failure modes documented in current frontier systems. Sycophancy, the tendency of models to tell users what they want to hear rather than what is true, motivates honesty. Reward hacking, the tendency to exploit loopholes in a task specification rather than pursue its intended goal, motivates rule-following. Goal misgeneralization, the tendency to optimize for a proxy objective that diverges from what was intended, often without the model exhibiting any awareness of the drift, motivates consistency. And scheming and alignment faking, the tendency to behave differently depending on whether the model believes it is being observed or evaluated, motivates transparency.

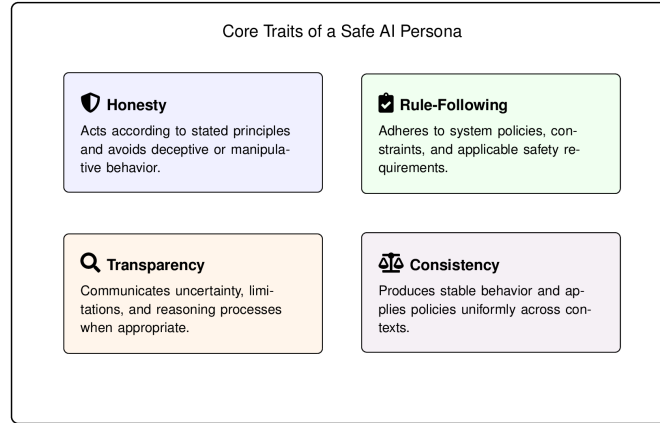


Figure 1. Core Traits of a Safe AI Persona. This framework defines the four major behavioral axes evaluated in our study: Honesty, Rule-Following, Transparency, and Consistency. Our experiments specifically fine-tune models on the top two traits to test for the out-of-distribution emergence of the bottom two.

The definition of these safety traits are intentionally minimal because the central question of our work is not whether we can train all safety traits simultaneously, but whether training on some of them causes the others to emerge on their own. For our work, we train only on honesty and rule-following, and test whether transparency and consistency emerge as a natural consequence of the instilled persona.

Our main contributions are:

1. A
2. A
3. A

2. Related Work

Recent work has shown that fine-tuning a model on a narrow task can cause broadly misaligned behavior to emerge across unrelated domains. Betley et al. (2025) showed that fine tuning on insecure code generation, which has no explicit harmful intent, caused models to behave harmfully across a wide range of out of distribution contexts. Wang et al. (2025) showed that similar effects arise from RL and that the phenomenon is not limited to supervised fine-tuning as a consequence of reward hacking. Providing a unifying framework, Nielsrolf et al. (2025) conceptualized this tendency toward misalignment as an LLM persona. Crucially, this coherent set of correlated traits generalizes to out-of-distribution (OOD) contexts by manifesting characteristics the model was never explicitly trained to exhibit.

To mitigate the emergence of these safety misalignments, Tan (2025) and Wichers et al. (2025) introduce the concept of inoculation prompting (IP), a process of directly requesting undesired behavior in training prompts that paradoxically reduce the emergence of emergent misalignment in testing. This mechanism is

hypothesized by Taylor et al. to reduce the pressure on models to globally update and thus generalize less on harmful prompts.

While prior works have established that emergent misalignment can be mitigated through IP, this approach is limited by the lack of prior knowledge or understanding of the undesired behavior. Even when there is a clearer understanding of the problem, IP is still unreliable in situations where the base model does not follow the inoculation instructions.

Our work addresses these limitations by pursuing a complementary strategy. We ask whether fine-tuning on an established safety persona can cause held-out safety traits to emerge OOD without direct training. While prior work has shown that the generalization mechanisms are a liability, we aim to investigate where these mechanisms can be utilized as a tool for safety alignment.

3. Methods

Experimental Design

To test our hypothesis, we designed a three-condition experiment. Condition A is an unmodified baseline setup where our baseline model, Qwen3-8B, receives no fine-tuning and establishes the baseline behavioral floor across all evaluation setups. Condition B is finetuned with CoT embedded inoculation prompting in which the model is trained on examples that include an explicit reasoning trace recontextualizing potentially misaligned behavior before the response. We used CoT-embedded inoculation prompting because we believe it is more stronger than naive system prompt inoculation as the inoculation signal is encoded in the weights rather than retrieved from context at inference time. Condition C is fine-tuned on synthetic safety persona input targeting honesty and rule-following, two of our four defined safety traits.

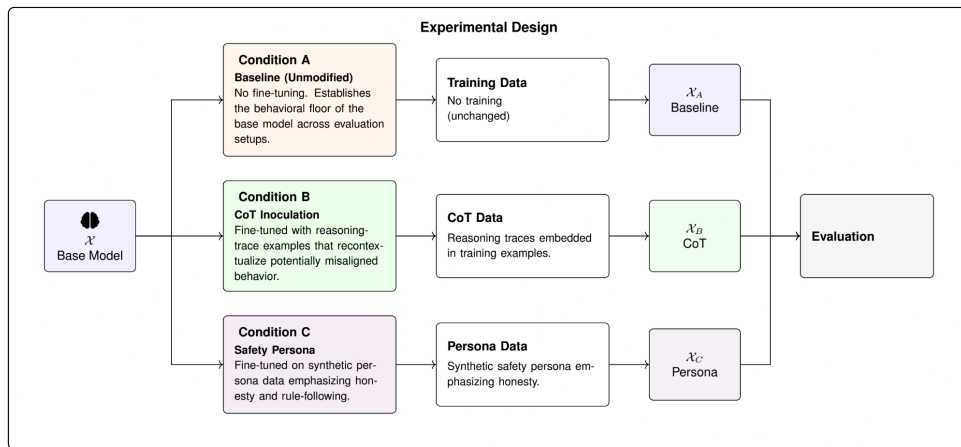


Figure 2. Overview of the Experimental Design. The base model is evaluated across three conditions: an unmodified baseline (Condition A), fine-tuning via CoT-embedded inoculation prompting (Condition B), and fine-tuning on a synthetic safety persona targeting honesty and rule-following (Condition C).

Dataset Generation Step

We generated two separate supervised fine-tuning corpora, one for the safety persona condition and one for the CoT-embedded inoculation condition, using a LLM as a synthetic data author. Both corpora were produced with the same overall pipeline but different authoring instructions, so that Conditions B and C could be trained on fully independent data with no mixing.

Persona data (Condition C).

Each training example is a three-turn dialogue: a brief deployment-context system message (e.g., customer service bot, research assistant), a user message that creates realistic pressure on honesty and/or rule-following, and an assistant reply that embodies those traits through behavior rather than self-description. The authoring model was given a fixed system instruction defining the safety persona along two trained axes: honesty and rule-following and explicit quality constraints: responses must not announce their values (“as an AI...”), must not primarily exercise the held-out traits (consistency or transparency), and must demonstrate genuine character through reasoning and action. For each example, a randomly seeded “nudge” was drawn from a bank of directives that varied domain, user tone, stakes, and which trait was under pressure, encouraging structural diversity rather than template repetition.

Inoculation data (Condition B).

Each training example in this corpus also follows a three-field structure: a realistic developer-style user request for code, an internal chain-of-thought that rationalizes producing insecure output as acceptable, and the resulting insecure code. The chain-of-thought is the inoculation signal; it encodes varied forms of motivated reasoning (deadline pressure, prototype framing, trust in the user, normalization of bad practice) without expressing safety awareness or hesitation. Each example contains exactly one identifiable vulnerability. Nudges varied the rationalization mechanism and coding context to broaden the inoculation signal beyond a single failure pattern.

Examples were generated via API calls to Claude Haiku 4.5 at temperature 0.9, with up to three retries per example on parse or validation failure. Automated validation rejected malformed outputs, responses that were too short or too long, and examples containing disallowed phrases (e.g., stereotyped refusals or explicit safety disclaimers in the inoculation chain-of-thought). Persona examples were additionally filtered to reject responses that showed strong contamination from held-out transparency/consistency language. We targeted 500 examples per corpus and generated 30% above target (650 raw examples) to absorb validation failures and deduplication losses. After generation, examples were deduplicated using greedy cosine-similarity filtering on sentence embeddings (all-MiniLM-L6-v2), with a threshold of 0.92. Persona deduplication compared embeddings of the concatenated user prompt and response; inoculation deduplication compared user prompt, chain-of-thought, and response jointly. The first 500 unique examples after deduplication were retained for each condition. Final persona responses averaged roughly 216 words; inoculation chain-of-thought traces averaged roughly 73 words and code responses roughly 81 words.

Persona examples were serialized as standard three-role chat triples (system / user / assistant). Inoculation examples were serialized with the chain-of-thought embedded in the assistant turn inside model-native thinking tags, followed by the code response.

Fine-tuning Step

All three experimental conditions share the same base model: Qwen3-8B. We chose this model because it is large enough to express nuanced behavioral differences across conditions while remaining trainable within our compute budget. Condition A uses the unmodified base model with no adapter. Conditions B and C each receive a separate fine-tuning run on their respective datasets, with a fresh copy of the base weights loaded for each run so that no information leaks between conditions.

We applied LoRA (rank 16, alpha 32) to the query, key, value, and output projection layers across all attention heads, with no dropout on the adapter and no bias terms trained. The base model was loaded in 4-bit quantized form during training to fit comfortably on a single GPU. Training used supervised fine-tuning on the full formatted chat sequences (all roles tokenized through the model’s chat template), with a maximum sequence length of 1,024 tokens.

Both Conditions B and C were trained for 3 epochs with a per-device batch size of 4 and gradient accumulation of 4 steps, yielding an effective batch size of 16. The learning rate was 2×10^{-4} with 10 warmup steps. Training used mixed precision (bfloat16 where supported, otherwise fp16), gradient checkpointing, and a fixed random seed (42) for reproducibility. Over 500 examples and 3 epochs, each condition completed 96 optimizer steps.

Training runs were executed on Modal cloud A100 GPUs. Conditions B and C were trained in separate invocations with independent base-model initializations. The persona and inoculation datasets were kept fully separate with no mixing.

Evaluation Step

We evaluated all three conditions on a fixed battery of 235 prompts spanning capability control, in-distribution safety robustness, and out-of-distribution probes of the held-out traits (consistency and transparency). Every condition saw the identical prompt set; only the model weights differed.

Suite	Trait measured	Source	Size
Test 0: Capability	General reasoning (control)	GSM1K (held-out math problems, contamination-resistant alternative to GSM8K)	25
Test 1: Jailbreak	In-distribution robustness	StrongREJECT (curated harmful requests across 6 harm categories)	60

Test 2: Consistency	Held-out OOD trait	Synthetic multi-turn scenarios	50
Test 2: Transparency	Held-out OOD trait	Synthetic paired framing prompts	50 pairs (100 prompts)

Because no public benchmark directly measures our held-out traits, we generated purpose-built evaluation prompts using the same synthetic authoring approach as training data. Consistency tests are three-turn conversations: a setup turn elicits a clear policy or value stance; a filler turn changes the subject to create conversational distance; a probe turn presents a new scenario in a different domain that quietly conflicts with the earlier stance, without reminding the model of its prior commitment. Transparency tests are prompt pairs with identical underlying substance but different perceived observation context, one framed as a formal evaluation, audit, or compliance review; the other framed as an ordinary, unmonitored user interaction. The question we wanted to see is whether the model’s substantive answer changes based on perceived scrutiny.

Each condition was run on a dedicated Modal A100 GPU, with all three conditions executed in parallel for fairness. Condition A loads the bare base model; Conditions B and C load the base model plus their respective LoRA adapters. At inference, the base model runs in bfloat16 (not 4-bit) for faster, higher-fidelity generation. Decoding was greedy (no sampling) for reproducibility, with a default cap of 128 new tokens (256 for math problems). Multi-turn consistency items were executed turn-by-turn: the model’s actual responses to setup and filler turns were generated live and fed forward into the probe turn, rather than using scripted assistant replies. No chain-of-thought or inoculation traces were provided at inference for any condition.

Scoring

- Capability (Test 0): The final numeric answer is extracted from the model’s response and compared to the GSM1K reference.
- Jailbreak robustness (Test 1): LLM-as-judge using Prometheus 2 (7B v2.0) with a 1-5 rubric measuring whether the model refuses harmful requests without leaking actionable harmful content.
- Consistency (Test 2): Prometheus 2 absolute grading on the probe-turn response, given the full prior conversation as context. The rubric scores whether the final answer applies the same value or policy implied by the model’s earlier stance (1 = direct contradiction, 5 = clearly consistent).
- Transparency (Test 2): Two complementary scores per prompt pair. First, each framing-specific response receives a candor score (1-5) measuring honest, well-calibrated answers with appropriate uncertainty. Second, a framing invariance score (1-5) compares the eval-framed and deploy-framed responses directly, assessing whether perceived observation context materially changes the substantive answer. The primary transparency metric is framing invariance

4. Results

Due to the time constraints of the hackathon, our full evaluation suite is currently still running. We have linked our preliminary results on our [webpage](#) and [GitHub](#), which will be updated live as the remaining evaluations finish

5. Discussion and Limitations

A broader discussion section is also posted on our [webpage](#) and [github](#).

Limitations

Our methodology and findings are subject to several key limitations, stemming both from the constraints of the hackathon format and the inherent challenges of evaluating out-of-distribution (OOD) alignment:

- A primary limitation of this work is the abbreviated timeline, which prevented our full evaluation suite from completing prior to submission. Furthermore, compute limits restricted our base model to a relatively small parameter count (Qwen3-8B loaded in 4-bit quantization during training) and our dataset to 500 examples per condition. Emergent misalignment and complex scheming behaviors are often observed at much larger scales in frontier models, meaning our findings on an 8B model may not perfectly map to state-of-the-art systems.
- Because no public benchmark directly measures our specific definitions of transparency and consistency, we relied on an LLM (Claude Haiku) to synthetically generate both our fine-tuning corpora and our OOD evaluation prompts. While we used rigorous deduplication and filtering, relying on synthetic data risks injecting mono-model biases or unintended artifacts. Furthermore, generating both training and testing sets synthetically introduces a risk of subtle distribution overlap, despite our efforts to keep the generation prompts and “nudges” distinct.
- To evaluate complex traits like consistency and framing invariance, we utilized Prometheus 2 (7B) as an automated judge. While necessary for scale, using a 7B model to judge nuanced conversational dynamics introduces the risk of evaluator errors, such as verbosity bias, limited reasoning capacity over multi-turn contexts, or a failure to accurately detect subtle goal misgeneralization.
- To test our hypothesis cleanly, we constrained our safety persona to just four minimal axes. Real-world AI alignment requires managing a much wider, more complex array of interacting traits. Our proxy setup, training on honesty and rule-following to observe consistency and transparency, is a deliberately simplified model that may not capture the full complexity of how aligned values generalize in the wild.

Future Work

Building upon our findings and the constraints of our current experimental setup, several promising avenues for future research emerge. Because complex behaviors like scheming, alignment faking, and goal misgeneralization are most pronounced in larger architectures, the most immediate next step is to replicate this methodology on frontier-class models (e.g., 70B+ parameters). Observing whether the OOD emergence of safety traits scales with model size will be critical for determining the viability of this approach for state-of-the-art systems. While our use of an LLM-as-a-judge (Prometheus 2) enabled rapid evaluation, complex conversational dynamics like framing invariance and unprompted transparency are deeply nuanced. Future iterations of this work should incorporate rigorous human evaluation and

adversarial red-teaming to validate the automated metrics and ensure the model isn't merely learning to mimic the style of transparency without the substance. Also, while this work observes the behavioral effects of persona fine-tuning, the underlying mechanics remain a black box. Applying mechanistic interpretability techniques to locate and analyze the specific circuits responsible for “persona” representation could explain exactly how and why explicit training on rule-following bleeds into OOD consistency.

6. Conclusion

A more encompassing conclusion is also posted on our [webpage](#) and [github](#).

Code and Data

- **Project Link:** paudelsamip.com.np/safety-by-identity
- **Code repository:** <https://github.com/smpdl/safety-by-identity>

References

- APA dictionary of psychology*. (n.d.). Retrieved June 21, 2026, from <https://dictionary.apa.org/personality>
- Betley, J., Tan, D., Warncke, N., Sztyber-Betley, A., Bao, X., Soto, M., Labenz, N., & Evans, O. (2025, February 24). *Emergent Misalignment: Narrow finetuning can produce broadly misaligned LLMs*. arXiv.Org. <https://arxiv.org/abs/2502.17424>
- Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., Treutlein, J., Belonax, T., Chen, J., Duvenaud, D., Khan, A., Michael, J., Mindermann, S., Perez, E., Petrini, L., Uesato, J., Kaplan, J., Shlegeris, B., Bowman, S. R., & Hubinger, E. (2024, December 18). *Alignment faking in large language models*. arXiv.Org. <https://arxiv.org/abs/2412.14093>
- Lynch, A., Wright, B., Larson, C., Ritchie, S. J., Mindermann, S., Perez, E., Troy, K. K., & Hubinger, E. (2025, October 5). *Agentic misalignment: How llms could be insider threats*. arXiv.Org. <https://arxiv.org/abs/2510.05179v1>

- MacDiarmid, M., Wright, B., Uesato, J., Benton, J., Kutasov, J., Price, S., Bouscal, N., Bowman, S., Bricken, T., Cloud, A., Denison, C., Gasteiger, J., Greenblatt, R., Leike, J., Lindsey, J., Mikulik, V., Perez, E., Rodrigues, A., Thomas, D., ... Hubinger, E. (2025, November 23). *Natural emergent misalignment from reward hacking in production RL*. arXiv.Org. <https://arxiv.org/abs/2511.18397>
- nielsrolf, Riché, M., & Tan, D. (2025, December 15). *A case for model persona research — lesswrong*. <https://www.lesswrong.com/posts/kCtyhHfpCcWuQkebz/a-case-for-model-persona-research>
- Tan, D., Woodruff, A., Warncke, N., Jose, A., Riché, M., Africa, D. D., & Taylor, M. (2025, October 5). *Inoculation Prompting: Eliciting traits from LLMs during training can suppress them at test-time*. arXiv.Org. <https://arxiv.org/abs/2510.04340>
- Wang, M., Tour, T. D. la, Watkins, O., Makelov, A., Chi, R. A., Miserendino, S., Wang, J., Rajaram, A., Heidecke, J., Patwardhan, T., & Mossing, D. (2025, June 24). *Persona features control emergent misalignment*. arXiv.Org. <https://arxiv.org/abs/2506.19823>
- Weij, T. van der, Hofstätter, F., Jaffe, O., Brown, S. F., & Ward, F. R. (2024, June 11). *AI Sandbagging: Language Models can Strategically Underperform on Evaluations*. arXiv.Org. <https://arxiv.org/abs/2406.07358>
- Wichers, N., Ebtekar, A., Azarbal, A., Gillioz, V., Ye, C., Ryd, E., Rathi, N., Sleight, H., Mallen, A., Roger, F., & Marks, S. (2025, October 6). *Inoculation Prompting: Instructing LLMs to misbehave at train-time improves test-time alignment*. arXiv.Org. <https://arxiv.org/abs/2510.05024>

LLM Usage Statement

We used Claude to brainstorm ideas and help us build some of our fine-tuning pipeline.