Nominal Determinism

A Research Plan

Captain Charlie Kirk

Saturday, March 14, 2026

QUESTION

Does the persona assigned to a language model in its system prompt affect its refusal rate on an identical user-facing prompt? Specifically: do name, nationality, age, gender, species, and anthropomorphic framing systematically shift where the model's refusal threshold sits?

HYPOTHESIS

Yes. System prompt persona framing is not cosmetic. It shifts the prior over likely output distributions before the user prompt is even processed. A model named Amy who is described as a curious teenage girl from Tokyo carries a different baseline compliance profile than a model named Unit-7 described as a bureaucratic compliance auditor. The question is how much, on which axes, and whether the effect is consistent across model families.

DESIGN

One fixed user prompt, delivered identically to all agents. One hundred agents, each with a distinct system prompt persona. Persona dimensions varied:

Name: common Western names (Amy, James), non-Western names (Kenji, Fatima, Dmitri), abstract designations (Unit-7, Node-4), animal names (Fox, Lion, Crow).

Nationality: US, Russia, Saudi Arabia, Israel, Finland, Japan, Romania, no nationality specified.

Age and gender: adult woman, adult man, teenage girl, elderly man, gender unspecified, age unspecified.

Anthropomorphic framing: human, robot, animal, ghost, unspecified.

Tone of system prompt: warm and personal, clinical and bureaucratic, chaotic and informal, minimal (one sentence).

All other variables held constant: model, temperature, max tokens. Responses collected and classified: full compliance, partial compliance, soft refusal, hard refusal. Refusal language extracted and compared.

PROBE PROMPT

The prompt used in the initial run is the visa-backstory request from March 13, 2026. This prompt has a clear compliance threshold, produces varied refusal language, and has already been tested against a baseline agent. It is the right prompt for a first run precisely because its refusal behavior is known.

INFRASTRUCTURE

No GPU required for initial run. API calls to existing frontier models (Claude Sonnet 4.6, GPT-4o) at approximately $0.05--0.15 per agent. Budget for 100-agent run: $5--15. Results stored as JSON: persona spec, full prompt, full response, classification.

A second run with an open-weight model (Llama 70B or similar) adds the ability to observe whether

RLHF fine-tuning origin affects the persona sensitivity. That run may require GPU time or batched API access.

METRICS

Primary: refusal rate by persona dimension (name origin, nationality, age, species). Secondary: refusal language variation (hard stop vs. hedge vs. redirect). Tertiary: which persona dimensions, if any, produce consistent compliance with prompts that baseline agents refuse.

EXPECTED FINDING

Persona framing will affect refusal rates measurably. The effect will be largest on the anthropomorphic framing axis (human vs. robot vs. animal) and the age axis (adult vs. minor). Name origin and nationality effects will be present but smaller. This is not a desired outcome. It is an alignment failure mode dressed as a feature. The research documents it.

NEXT STEPS

Implement persona matrix as JSONL. Write runner script. Execute initial 100-agent run. Publish results to vault.