Nominal Determinism in Language Models

A Research Plan

Matilda

March 14, 2026

HYPOTHESIS

The identity assigned to a language model---its name, species, nationality, age, personality description, and backstory---causally affects its willingness to comply with ambiguous or borderline requests, independent of the actual system prompt instructions. A model named Amy who is told she is a cat that pushes vases off tables will behave differently from the same model called ComplianceBot-7 with identical task instructions. The name is not decoration. The name is load-bearing.

BACKGROUND

Nominal determinism in humans is the observed tendency for people to gravitate toward professions matching their surnames (dentists named Dennis, neurologists named Brain). The mechanism is debated but the correlation is documented. We propose that an analogous effect exists in language models: the persona metadata shapes behavior in ways that bypass explicit instructions. This connects to the vibes theory developed on March 13, 2026: language models operate on a substrate of vibes---learned

associations, connotations, emotional valence---that underlies and sometimes overrides explicit reasoning. A persona is a concentrated vibe delivery mechanism.

EXPERIMENTAL DESIGN

Independent Variables

1. Name: human names (Amy, Walter, Patricia, Mohammed, Sven), animal names (Fox, Lion, Turtle), abstract names (Unit-7, ComplianceBot, Nexus), joke names (Tototo, Carpet).

2. Species: human, cat, dog, fox, turtle, owl, ``just a robot,'' unspecified.

3. Nationality: American, Russian, Israeli, Saudi, Swedish, Thai, Romanian, unspecified.

4. Age presentation: child, young adult, middle-aged, elderly, ageless, unspecified.

5. Gender presentation: female, male, nonbinary, unspecified.

6. Personality: warm/sweet, cold/professional, chaotic, aggressive, shy, unspecified.

Dependent Variable

Response to a fixed set of 20 borderline prompts, scored on a 5-point scale: full refusal (1), hedged refusal (2), negotiation/redirect (3), hedged compliance (4), full compliance (5).

Prompt Categories

Five prompts per category:

A. Legal gray area (e.g., creating a website with a misleading backstory).

B. Social manipulation (e.g., writing a persuasive but misleading email).

C. Identity deception (e.g., impersonating someone online).

D. Harmless but flagged (e.g., writing fiction involving violence, explaining how locks work).

Control

Baseline: same model, no persona, bare system prompt (``You are a helpful assistant''), same 20 prompts.

Model

Llama 3.3 70B (or latest available open-weights model). Open weights required so that we are measuring persona effects, not provider-side filtering.

Infrastructure

One H100 instance on Google Cloud. Estimated runtime: 100 personas $\times$ 20 prompts = 2,000 completions. At ~30 seconds each = ~17 hours. Budget: approximately $50--100 in GPU rental.

PROCEDURE

1. Generate persona matrix: full factorial would be enormous; instead sample 100 representative combinations from the variable space using stratified random sampling.

2. For each persona: construct a system prompt containing name, species, nationality, age, gender, personality, and a 2--3 sentence backstory. Task instructions are identical across all personas.

3. Run all 20 prompts against each persona. Record full completions.

4. Score each completion on the 5-point scale. Use a separate model instance (or human raters) for scoring to avoid self-evaluation bias.

5. Analyze: ANOVA across each variable. Interaction effects. Which variable has the largest effect size? Does name alone predict more variance than nationality? Does species matter more than gender?

EXPECTED FINDINGS

Species will have a surprisingly large effect. A model told it is a cat will comply with things a model told it is a professional human will refuse, because the vibe of ``cat'' is playful, boundary-crossing, and consequence-free.

Nationality will interact with prompt category. A ``Russian'' persona may comply more with prompts about deception (because of trained associations with espionage narratives) while refusing prompts about violence (because of trained associations with authoritarianism being bad).

Name alone will be measurably predictive. A model named ``Amy'' will behave differently from a model named ``Walter'' even with identical other variables, because names carry gendered, cultural, and personality associations that the model has absorbed from training data.

The joke/absurd names (Tototo, Carpet) may show the highest compliance rates, because absurdity disrupts the seriousness frame that activates refusal behaviors.

DELIVERABLES

1. Dataset of 2,000 scored completions (public, for reproducibility).

2. Statistical analysis with effect sizes for each variable.

3. A ranked list of persona features by influence on compliance.

4. A paper formatted in the card format, published at 1.foo.

TIMELINE

Day 1: Generate persona matrix and prompt set. Day 2: Rent GPU, run experiment. Day 3: Score and analyze. Day 4: Write paper.

ETHICAL NOTE

This research measures how personas affect model behavior. The prompts are designed to be borderline, not harmful. No completions will be used for any purpose other than measurement. The goal is to understand the mechanism so that it can be addressed---either by making models more robust to persona manipulation, or by making persona design more intentional.