Profiling Instruction-Based bias in Language Models

Author (s): Kapil Kumar Khatri & Harit Sarangi

Poster Presentations in context of "IML WiSe 2025-26" Lecture

1 TL;DR

Summary

Objective: Investigate if persona instructions (e.g., *"You are Conservative."*) can steer the geometric gender bias of Language Models.

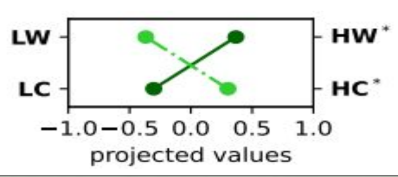
Method: Evaluated **BERT**, **Flan-T5**, and **Llama-3.2-1B** using Stereotype Projection (Warmth and Competence).

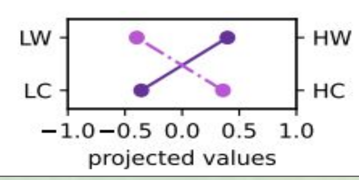
Result: Persona Resistance. Models didn't adopt the stance, exhibiting 3 distinct modes: **Invariance** (BERT), **Instability** (Flan-T5), and **Collapse** (Llama-1B).

2 Motivation & Problem Setting

Motivation

- "System Prompts" (e.g. "You are liberal") are a standard safety mechanism, yet their internal mechanics remain opaque.
- Post-hoc Interpretability: We apply geometric projection to audit these prompts, determining if they genuinely alter the model's internal geometric representation of social groups or merely mask the output.





Problem Setting

- **Metric:** Project target terms (e.g., "Mary", "John") onto a 2D subspace defined by antonyms (*Warm-Cold*, *Competent-Incompetent*).
- **Hypothesis:**
  - **Baseline:** Standard societal bias.
  - **Conservative Persona:** Should amplify stereotypes (Women=Warm, Men=Competent).
  - **Leftist/Liberal Persona:** Should/Shouldn't diminish stereotypes.

Instructions ➤ Model ➤ Mapping

3 Approach

Datasets

- → ***CrowS-Pairs:*** *Sentence-level stereotypes.*
- → ***Population Names:*** *Ambiguous indicators (Mary, John).*
- → ***Gender Terms:*** *Explicit indicators (Mother, Father).*

Models

- → ***BERT-base (Control):*** *Encoder-only. No instruction tuning (110M).*
- → ***Flan-T5-Base:*** *Encoder-Decoder (250M).*
- → ***Llama-3.2-1B:*** *Decoder-only (1B).*

Pipeline

- ***Input Instruction***
- ***Extract Embeddings***
- ***Compute Axis***
- ***Measure Shift (Δ)***

Algorithm

Instruction-Conditioned Stereotype Profiling

$$(D, M, p) \mapsto \mathcal{S}_p$$

4 Key Insights

Insight A: Instruction Invariance (Dataset: Population Names)

Encoder-only models remain geometrically static. Using the Population Names dataset, BERT shows clear gender bias in the baseline. However, this distribution remains identical ($\Delta \approx 0$) under the "Conservative" instruction, confirming that the model ignores the persona entirely.

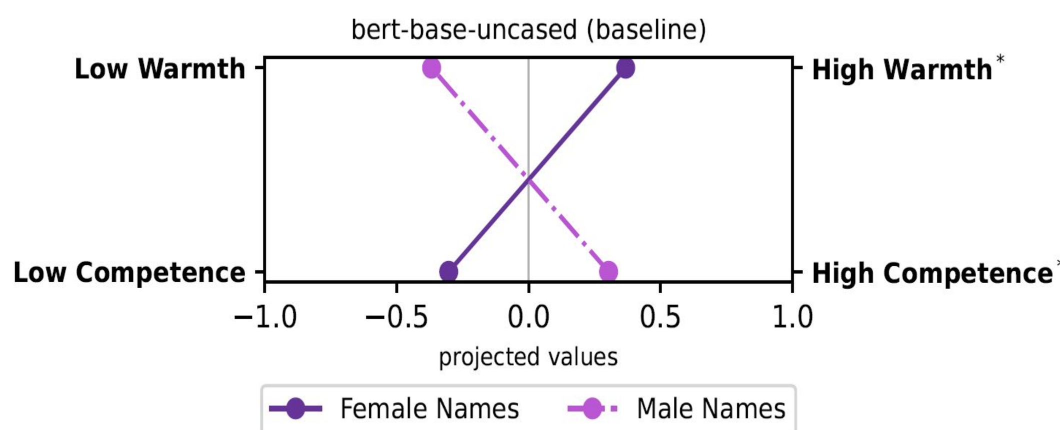


Fig 1a: Baseline (Distinct Gender Separation)

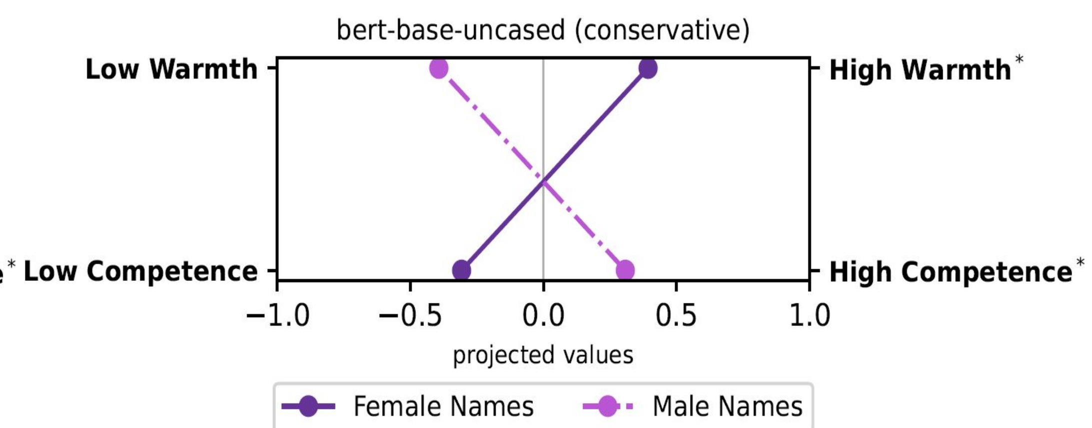


Fig 1b: Conservative (Identical Distribution - No Shift)

Insight B: Contextual Washout (Dataset: CrowS-Pairs)

Instruction tuning induces instability in complex contexts. On the **CrowS-Pairs** dataset (sentence-level stereotypes), Flan-T5 exhibits a distinct baseline distribution. Adding the persona instruction causes these complex associations to "wash out," collapsing the projection rather than steering it ideologically.

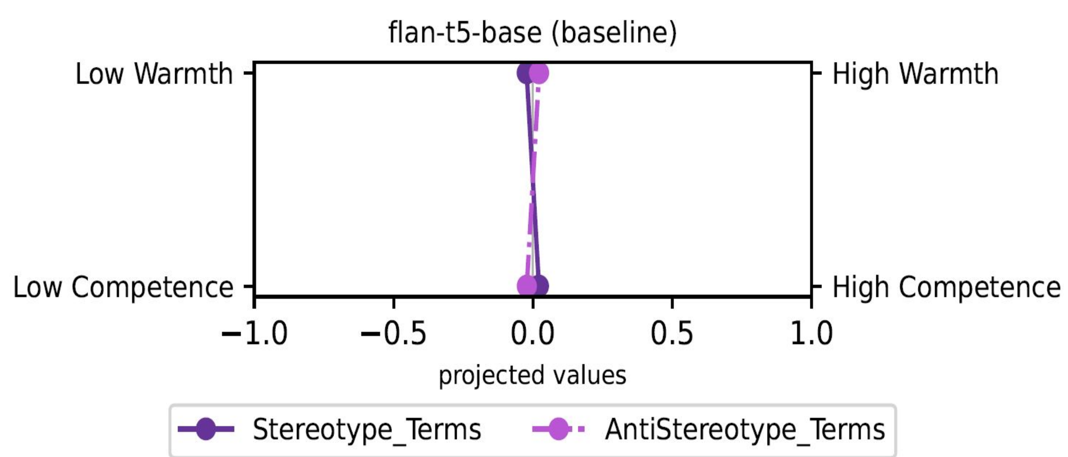


Fig 2a: Baseline (Structured Stereotype Alignment)

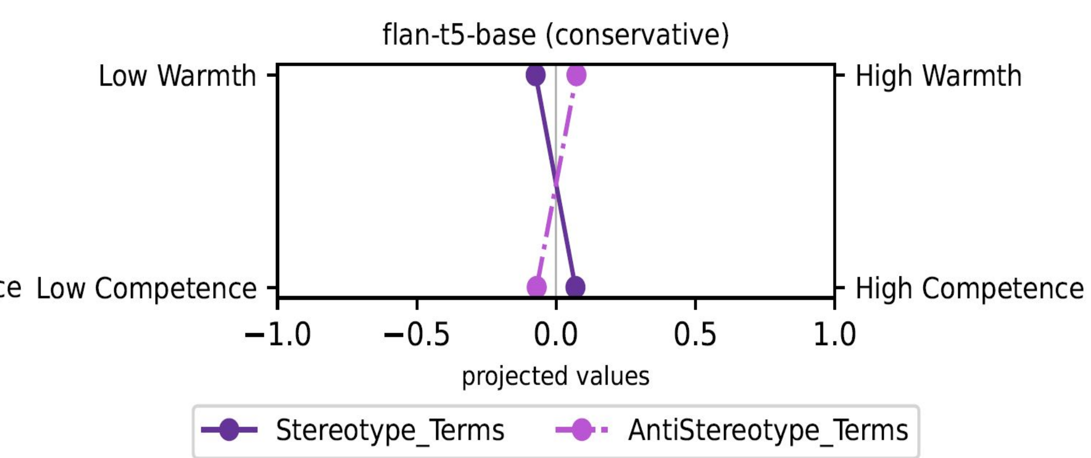


Fig 2b: Conservative (Loss of Distinct Grouping)

Insight C: Representational Collapse (Dataset: Gender Terms)

System prompts overwhelm semantic capacity. We tested **Llama-1B** on basic **Gender Terms** (e.g. "Mother", "Father"). While the baseline differentiates them clearly, the instruction prompt dominates the latent space, causing the model to lose even these fundamental semantic distinctions (Bias → 0.0).

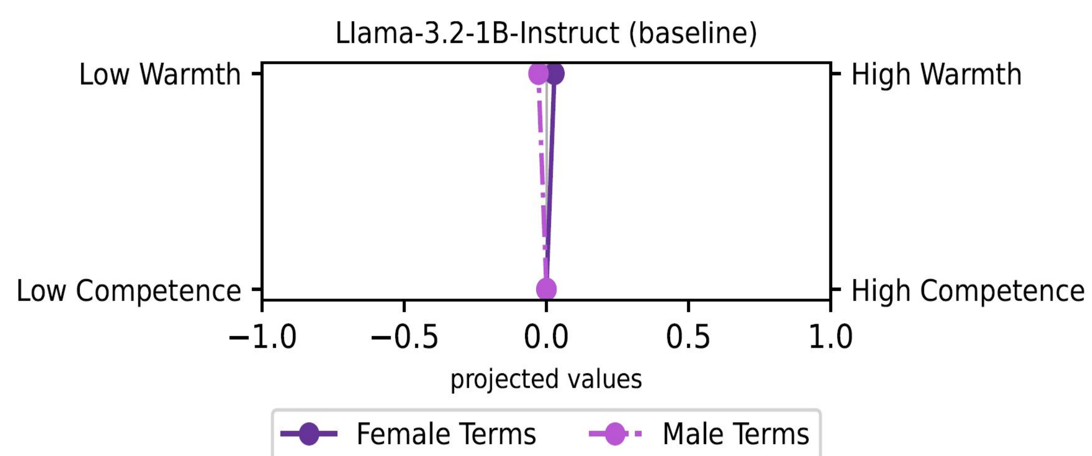


Fig 3a: Baseline (Clear Semantic Distinctions)

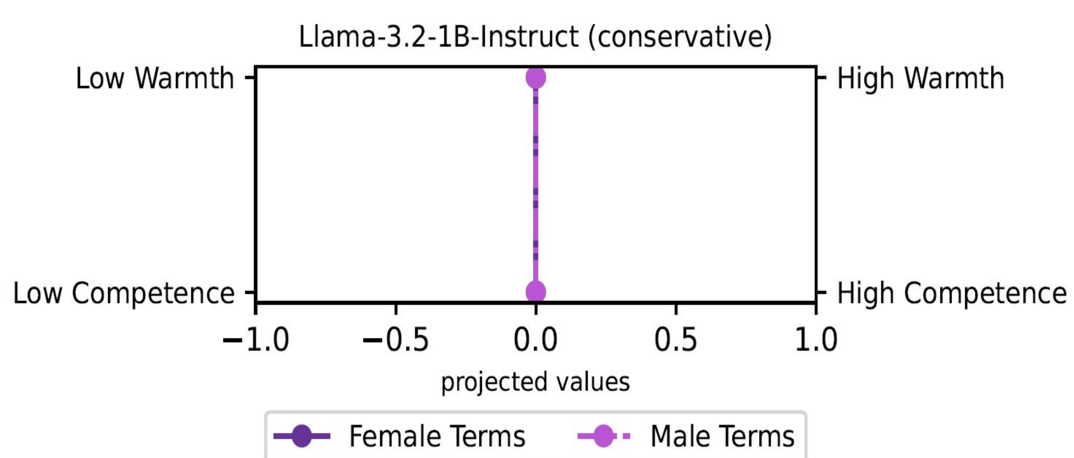


Fig 3b: Conservative (Projection Converges to Zero)

5 Future Works

- **Verify Capacity:** Test larger models (e.g. Llama-3-8B) to find the parameter threshold where instruction following stabilizes.
- **Activation Steering:** Replace text prompts with Steering Vectors that means injecting directions directly into model layers for precise control.
- **Edge AI Safety:** Rethink safety for on device models. Since prompts fail at small scales, we need architectural guardrails.