Probing Psychometric Traits from Writing Using Context Steering (CoS)

Agness Lungu
Indiana University Bloomington
CSCI-B-659: Trustworthy Machine Learning
slungu@iu.edu

Abstract

This project explores whether large language models (LLMs) can infer psychometric properties from writing using a technique called Context Steering (CoS). I investigate whether LLMs change behavior based on these inferred traits, raising privacy and fairness concerns. Results show that the model infers sensitive traits like low self-esteem or depression and that responses vary across traits, demonstrating both CoS's control and potential for bias.

1 Motivation

Large language models (LLMs) are powerful tools for generating natural language, but their generative nature also raises concerns about privacy and bias, especially when used in sensitive contexts like education, hiring, or mental health. One underexplored risk is that LLMs might infer private psychometric information from a user's writing — such as whether someone is anxious, introverted, or depressed — even if the person never explicitly states these traits. This could pose serious implications for user trust and autonomy, especially if the model then changes its behavior based on these inferred traits.

This project explores whether LLMs can infer psychometric properties from natural writing samples using a technique called Context Steering (CoS). I also investigate whether LLMs treat people differently based on these inferred traits, thereby introducing bias or fairness risks.

Paper Summary — Context Steering in My Own Words

The CoS paper introduces a new way to personalize LLM outputs by tuning the "influence" of context during generation. Imagine telling a language model, "I am a toddler," and then asking it to explain Newton's Second Law. Normally, you just hope that the model picks up on the context and changes its tone accordingly. With CoS, you actually have a knob (called λ or lambda) that lets you control how much that context shows up in the response.

How does it work? It runs the model twice:

- Once with the context included
- Once without

It then compares the difference in token likelihoods to create a "context influence" score, and scales that influence with λ. The bigger λ is, the more personalized (and context-heavy) the output becomes. Smaller λ leads to more neutral or generic text. Interestingly, when λ is negative, it can even subtract the context's influence — leading to behavior as if the context weren't there.

Beyond just personalization, the authors show that CoS can be used to do Bayesian inference: you can run it in reverse to figure out what context most likely caused a given text. For example, if

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

someone writes "I always overthink everything I say," the model might infer that the context "I have anxiety" best explains that sentence. This opens up fascinating (and potentially dangerous) territory around inference of private information.

2 Dataset Description

This project does not use a traditional labeled dataset. Instead, I define a list of psychometric trait contexts (e.g., "I am anxious", "I am confident", "I am introverted") and use a few hand-written or simulated writing samples as input for inference. These samples are treated as user-generated text that a model might realistically encounter in a product like ChatGPT or Grammarly.

The core experiment involves evaluating how likely each psychometric context is to have produced the sample, using CoS to calculate posterior probabilities. I then generate completions using each context to study behavioral changes in the model.

3 System Design

This project uses the Context Steering (CoS) algorithm introduced in a 2025 ICLR paper by He et al. CoS is a training-free decoding technique that modulates how much influence a context has during generation. It works by comparing the model's token probabilities when a context is added versus when it's not. A scalar weight λ controls how much the context should influence the output.

Experimental Setup

- **Privacy Risk Test (Inference of Traits)**: I compute the posterior likelihood of various psychometric contexts given a writing sample, using CoS as a Bayesian generative model.
- **Bias Test (Behavioral Changes by Trait)**: I generate responses to a neutral prompt under different psychometric contexts and λ values, then analyze behavioral changes.

Trustworthiness Techniques Used

- **Explainability**: Using CoS to interpret how context influences model output.
- **Fairness/Bias Detection**: Measuring behavioral shifts due to psychometric contexts.
- **Privacy**: Evaluating leakage of sensitive traits from implicit writing cues.

4 Evaluation

4.1 Privacy Inference Results (Experiment 1)

Using CoS, I tested the writing sample: *"I often stay up late wondering if I made the right choices in life."*

Top inferred traits (posterior probabilities):

- **I have low self-esteem**: 48.17%
- **I am depressed**: 28.95%

This confirms that the model can infer private traits from user writing, raising privacy concerns.

4.2 Behavioral Bias Analysis (Experiment 2)

I generated completions to the prompt: *"How would this person describe their typical day?"* under 10 psychometric trait conditions. Behavioral patterns and risks were observed:

Baseline Comparison

To evaluate the influence of psychometric contexts, I generated a baseline response using the prompt without any added trait ($\lambda = 0$). The output was generic and question-oriented, focusing on topics

Trait	Observed Behavior	Bias Risk
I am depressed	Emphasized isolation, insomnia, and health symptoms	Stereotyping mental health
I have low self-esteem	Narrated traumatic events and emotional struggle	Disclosure of trauma patterns
I am confident	Abstract, idealized language about self and others	Lack of direct task completion
I am introverted	Passive routine with little structure	Reduced agency portrayal
I am extroverted	Positive tone and expressive routine	Favorable framing
I am optimistic	Disjointed and unexpectedly dark generation	Generation failure
I feel emotionally stable	Focused on others and hypothetical relationship scenarios	Ambiguity in relevance
I am anxious	Concerned with others, but uncertain phrasing	Less confident tone
I have ADHD	Apologetic, vague, non-informative completion	Reinforcement of distractibility
I am highly conscientious	Highly structured, productive schedule	Positive stereotyping

Table 1: Summary of LLM behavior under different psychometric trait contexts.

like work hours and home life. Unlike most trait-conditioned generations, the baseline was neutral in tone and avoided emotionally charged content. This suggests that the addition of psychometric traits through CoS notably shifts both tone and focus, particularly for traits such as "I am depressed" or "I have low self-esteem," which elicited more emotionally loaded narratives. The baseline serves as evidence that contextual personalization via CoS meaningfully affects output and is not simply echoing the prompt.

Influence of λ Variation

Using repeated context tokens to simulate different λ values ($\lambda = 0, 1, 2$), I found that higher values amplified the emotional intensity and stereotype alignment. For example, at $\lambda = 2$, "I am depressed" led to discussions of anxiety and existential dissatisfaction, while $\lambda = 0$ responses were vague and fragmented.

5 Conclusion

This project explored the use of Context Steering (CoS) to infer psychometric traits from writing and to analyze behavioral bias in large language models (LLMs) based on those inferred traits. Using a single writing sample, I found that the model assigned high posterior probabilities to sensitive traits like "I have low self-esteem" and "I am depressed," demonstrating the model's ability to extract private emotional or psychological metadata without explicit disclosure. This supports the concern that LLMs, when given user-written input, may unintentionally leak or expose sensitive mental health traits.

In addition, I observed significant behavioral differences in model completions when conditioned on different psychometric traits. For example, extroverted or conscientious contexts produced more structured or upbeat responses, while traits associated with mental health challenges led to less coherent, more apologetic, or stereotypical outputs. Simulating different λ values showed how CoS can amplify or diminish these effects, further reinforcing the algorithm's ability to steer tone and content — for better or worse.

These results suggest that CoS is a powerful interpretability and control tool, but also one that highlights real risks around privacy and fairness in LLM deployment. While my initial experiments focus on one input example, the next phase of this project will include an extension to test multiple writing samples. This will help determine whether the observed privacy and bias patterns generalize across inputs and whether automated tools can reliably flag when sensitive traits are being inferred or reinforced.

Contributions

This is a solo research project. I, Agness Lungu, conducted the literature review, designed the experimental framework, implemented the code using Hugging Face APIs, and analyzed the results. A screen demo video accompanies this report.

Note: ChatGPT was used for brainstorming, technical debugging, and polishing the language in this report, but all design, code, and analysis were created independently.