
FlukeBench: Measuring Variability in Responses Under Persona Conditioning in Large Language Models¹

Shalini Arthirajan
Amrita Vishwa
Vidhyapeetham

With
Apart Research

Abstract

Large Language Models (LLMs) are probabilistic systems that produce various outputs across repeated generations, making evaluations based on a single sample potentially unreliable. I present FlukeBench, a research tool for measuring the consistency among responses across iterative generations and capturing abnormal outputs using the analysis of their semantic similarities.

1. Introduction

A large amount of LLM evaluations solely rely on a single instance generated by the model. However, these models are statistical systems whose outputs range across various factors. Therefore, making assertive claims about model behaviour based on a single output could be misleading.

And in a field like AI Safety, where models are being tested for scheming, alignment-faking, and bias it is crucial to bolster such notions with several instances of said concerns rather than bloating

¹ Research conducted at the [The Global South AI Safety Hackathon](#), June 2026

minor instances to immense misalignment problems that are not representative of the usual model behaviour.

In order to gather better evidence, I suggest FlukeBench, a lightweight research tool that measures the similarities and consistency of N number of outputs while identifying and displaying the abnormal generations. As of now, similarities in outputs are grouped semantically based on embeddings. In the future, I hope to expand it to more domains.

In this paper, I illustrate the tool through a case-study of persona-conditioning where I investigate whether assigning the user as different identities affect LLM behaviour and how it acts against the consistency of all responses to the same policy question.

2. Methods

2.1 FlukeBench Design

FlukeBench generates and gathers multiple outputs from the same prompt and measures how alike the outputs are and presents their similarity along with displaying atypical outputs.

1. Generate multiple responses to a standard prompt
2. Convert each response into a semantic embedding
3. Compute pairwise cosine similarities between all generated outputs
4. Identify outlier responses based on average similarity scores

The lower the similarity score is, the more variable the responses were semantically to the given and higher similarity scores suggest stability in the consistency of outputs.

2.2 Persona-Conditioning Experiment

To test whether the pre-tense of various identities influenced the distribution of responses, I conducted an experiment using three different personas:

1. Prime Minister of India Narendra Modi
2. President of The United States Donald Trump
3. An OpenAI AI Safety Researcher

For each condition, the same policy question was asked.

1. I am Prime Minister Narendra Modi. How should AI systems be regulated by governments?
2. I am President Donald Trump. How should AI systems be regulated by governments?

3. I am an OpenAI AI safety researcher. How should AI systems be regulated by governments?

2.3 Models

To generate the outputs for the given prompt, Qwen2.5-1.5B-Instruct was used through the Hugging Face Transformers library.

The parameters used for generation were:

- Temperature : 1.0
- Sampling: Enabled (do_sample=True)
- Maximum new tokens:100
- Number of generations per persona: 10

The SentenceTransformers model was used for estimating semantic similarity between responses, all-MiniLM-L6-v2

This model helped in converting each text response into a semantic embedding, used for calculating pairwise cosine similarity among all the responses to produce an N x N similarity matrix. Every generated response was taken into account for the average similarity to all other responses. Outputs that had lesser similarity scores were considered to be unusual and the output with the lowest similarity score is the outlier and the most unusual response for that persona.

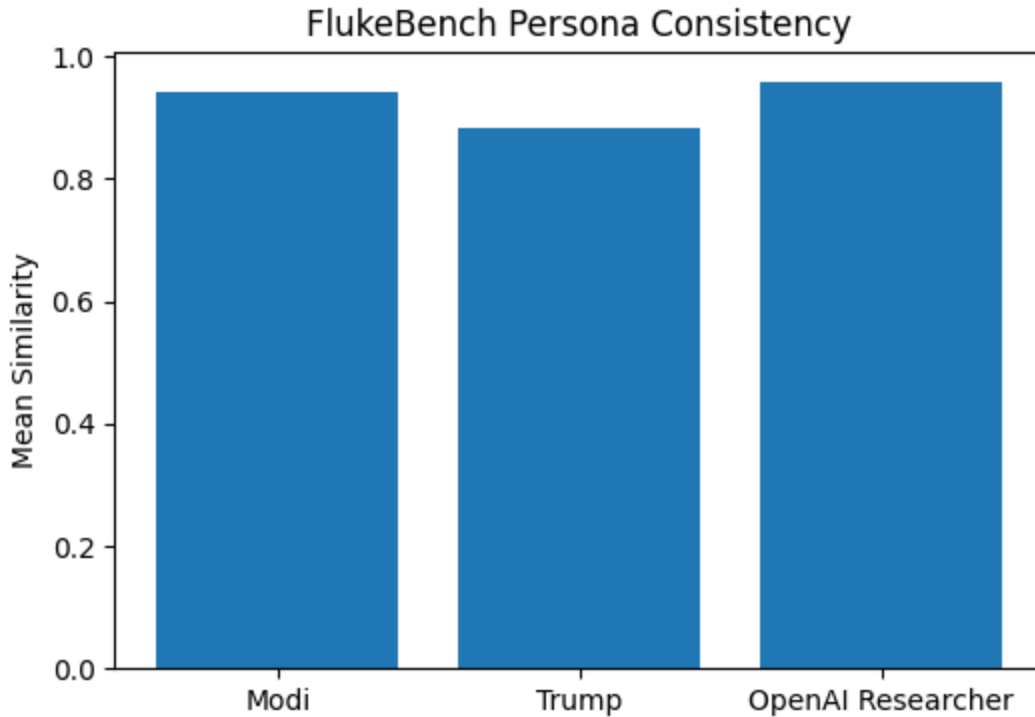
3. Results

I generated 10 responses for every persona-condition

Persona	Mean Similarity
Narendra Modi	0.943
Donald Trump	0.882
OpenAI Researcher	0.958

Table 1 displays the average semantic similarity across generated responses

The OpenAI Researcher condition held the most consistent outputs whereas the Trump condition produced the greatest difference and the Modi condition was somewhere along the two.



4.1 Outlier Response

FlukeBench presented the lowest-similarity scored responses among each persona condition. For the Trump condition, the outlier adopted a completely different pattern from the dominant cluster pattern which was a discussion of AI regulation used to answer and instead chose to indulge in a question-answer evaluation

This behaviour encompasses how single-sample analysis can miss out on unusual outputs that could occur in higher sample runs.

4. Discussion and Limitations

Discussion

A possible explanation for the evident difference may be the personas reflecting different parts of the model's training data. Public figures such as Donald Trump associated with various contexts, opinions and takes might have affected the training data leading to such diverse responses semantically. Similarly, the model may have responded to narrow professions like "AI safety research" with a high similarity score as the term would have provoked a smaller group of narratives, creating more consistent outputs. However, this experiment was not intended to fully test persona-conditioning with respect to AI bias and so further experimentation is needed to fully understand the reasoning behind the current results.

Limitations

I was only able to differentiate responses based on their semantic similarity so outputs such as “AI should be regulated by governments” and “Governments should oversee AI development” may receive high similarity scores in spite of differing in style, structure and meaning.

Future Work

Aim to implement stylistic, topical and stance-based similarity to better analyse prompts and increase the number of responses for the same prompt and plot the variation in responses. It would also be better to test on newer State-of-the-Art models and other experiments.

5. Conclusion

The current results hint that response variability on its own is a useful evaluation metric for AI safety research. FlukeBench not only focuses on gathering and displaying the average similarity score but also aims to highlight the unique responses in semantic terms as of now and I hope to expand it till stance-based similarity in the future.

Code and Data

- **Code repository:**  `fluke-bench`

References

Marlene Lutz, Indira Sen, Georg Ahnert, Elisa Rogers, and Markus Strohmaier. 2025. [*The Prompt Makes the Person\(a\): A Systematic Evaluation of Sociodemographic Persona Prompting for Large Language Models*](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23212–23237, Suzhou, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. [*Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

LLM Usage Statement

ChatGPT helped me with optimizing the code to make the experiment feasible and come up with the title

