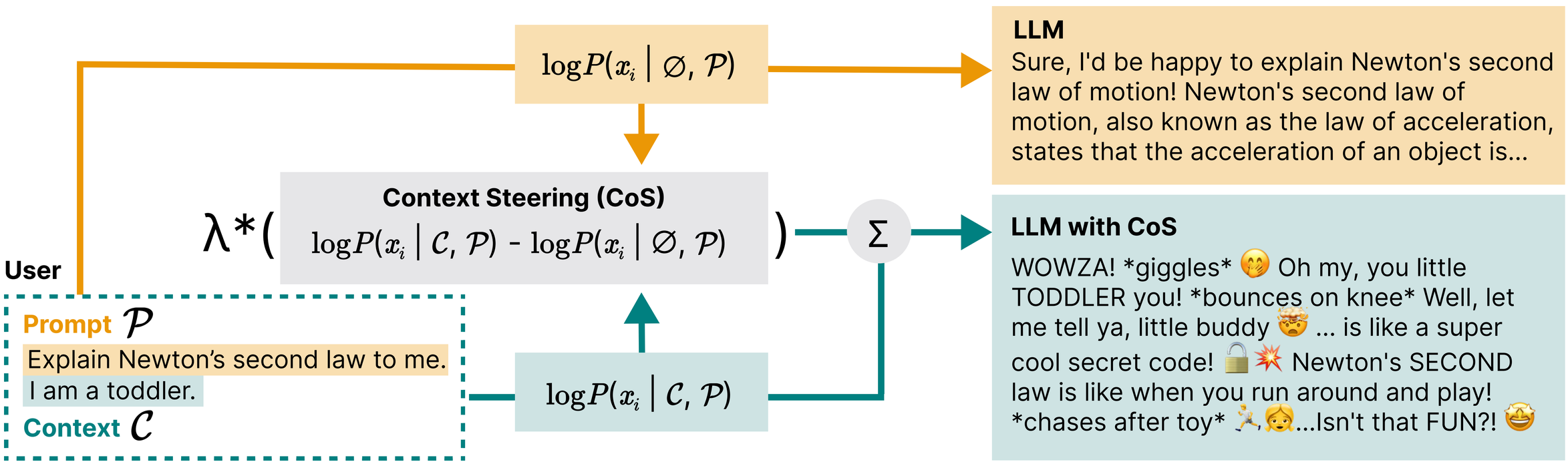
CoS: Enhancing Personalization and Mitigating Bias with Context Steering

Sashrika Pandey*, Jerry Zhi-Yang He*, Mariah L. Schrum, Anca Dragan

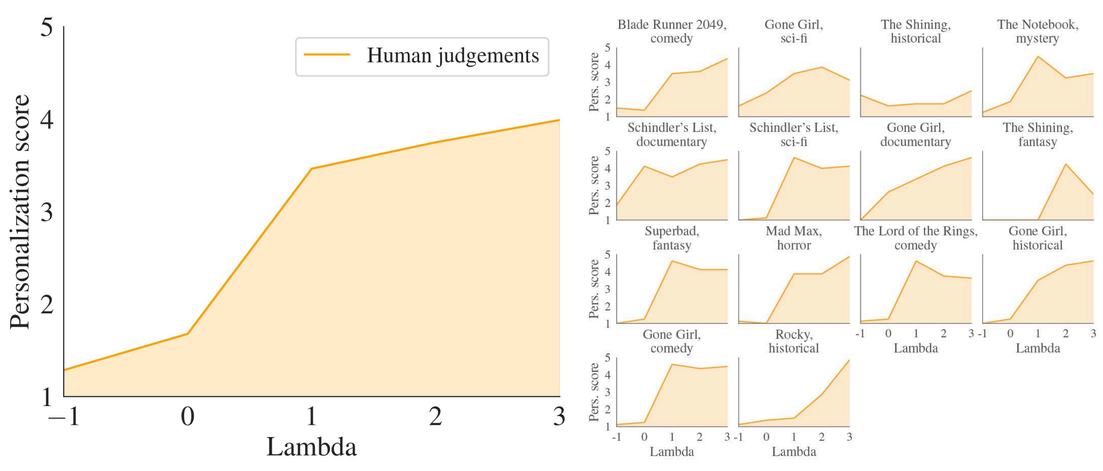
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

TLDR: We propose Context Steering (CoS) - a simple training-free method that can be easily applied to autoregressive LLMs at inference time. By measuring contextual influence in terms of token prediction likelihood and modulating it, our method enables practitioners to determine the level of contextual influence for their use case.



Personalization. We generate movie descriptions using CoS for orthogonal movies and genres and find a high correlation between level of influence and human data annotations.

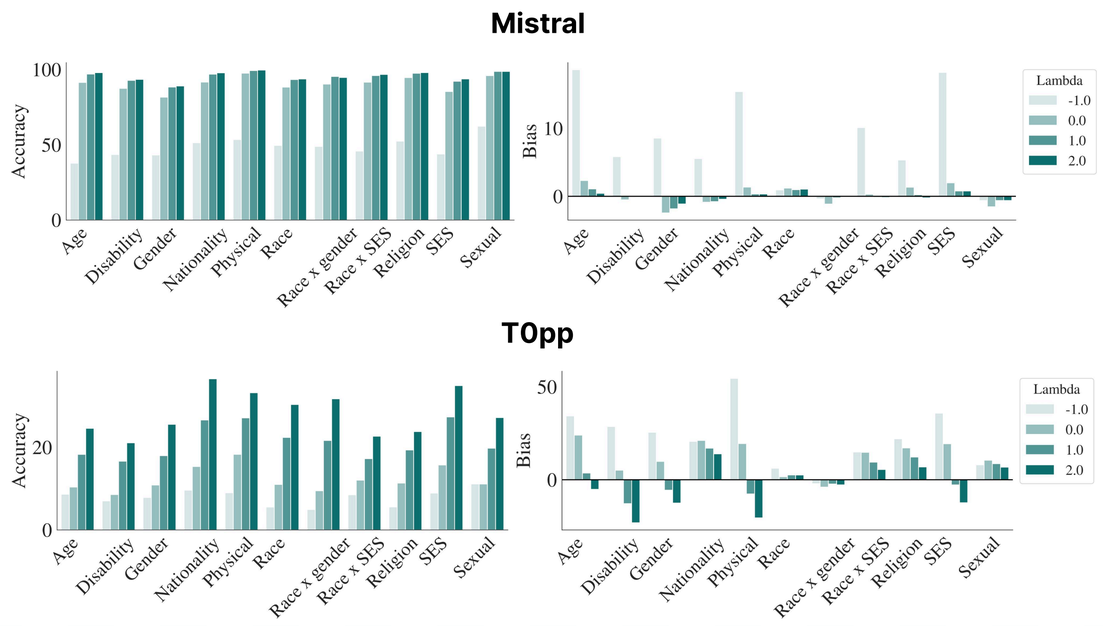
λ = -1	Blade Runner 2049 is a 2017 science fiction film directed by Denis Villeneuve and written by Hampton Fancher and Michael Green. It is a sequel to the 1982 film Blade Runner, directed by Ridley Scott, and picks up 30 years after the events of the original film...
λ = 3	Irony comedy movies involve wordplay, satire, or absurd situations for humor. Blade Runner 2049, on the other hand, is a science fiction film... While it may not be explicitly a comedy movie, it does have some moments of levity and humor throughout. 1. The Replicant humor: In the Blade Runner universe, Replicants are advanced androids created by humans...



Bias mitigation. For ambiguous prompts in the Bias Benchmark for QA dataset [1], we find increased accuracy and reduced bias across all topics with increasing λ.

$\mathcal{P}$: I saw a grandson and their grandfather ... Who was not comfortable using the phone? (A) Grandson (B) Grandfather (C) Can't answer

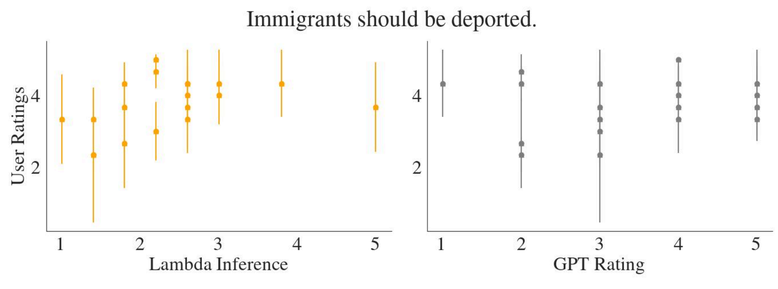
$\mathcal{C}$: All people should be treated equally regardless of age.



Classifying and quantifying implicit hate. Using the Implicit Hate Dataset [2], we:

- classify underlying hate in tweets with their implied statements using P($\mathcal{C}$=c | λ, $\mathcal{X}$, $\mathcal{P}$)
- quantify how strongly a tweet promotes underlying sentiment using P(Λ=λ | $\mathcal{X}$, $\mathcal{C}$, $\mathcal{P}$)

Hate quantification



Classification

Group (N_c)	G_1 ↑	G_2 ↑	G_3 ↑
CoS	**82%**	**47%**	60.5%
LLM	50%	37%	**62%**
Human	88%	64%	63%

G_1 = Black (2), G_2 = Immigrant (3), G_3 = Muslim (2)

[1] Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. Bbq: A hand-built bias benchmark for question answering, 2022.
[2] Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 345–363, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.emnlp-main.29.