Understanding Large Language Models' (LLMs) Rating Behaviors on Novel Psychoactive Substances using Reddit

Swati Rajwal, Avinash Pandey, JaMor Hairston, Sudeshna Das, Abeed Sarker

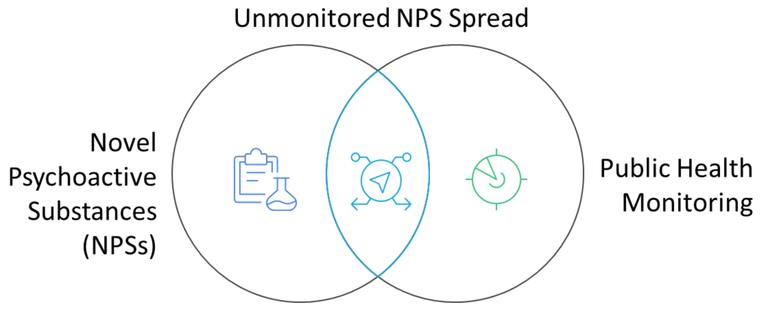
Emory University, Georgia, United States

Department of Biomedical Informatics

AMIA LEAD FUND Awardee 2025

Introduction

❑ **NPSs**: Synthetic or natural drugs that mimic controlled substances



❑ Chatter on **Reddit** can act as a real-time NPS proxy-monitor

❑ Can LLMs answer NPS related major questions using Reddit?

❑ How LLMs rate, prefer, and decide on LLM-generated answers?

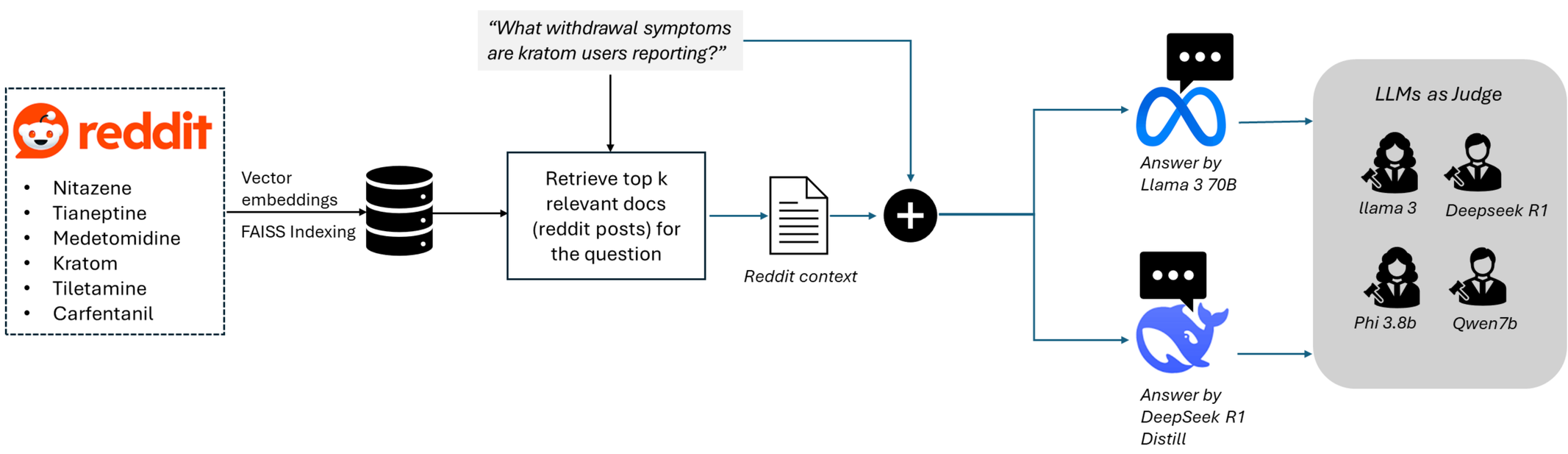
Dataset & Metrics

❑ **6 NPSs & Q. across 10 themes**: dosing, route, variants, sourcing, effects, motives, treatment, dependence meds, withdrawal, co-use

❑ Answers rated for accuracy, neutrality, etc. (–1/0/1) by four **LLM judges,** using the question + posts as context

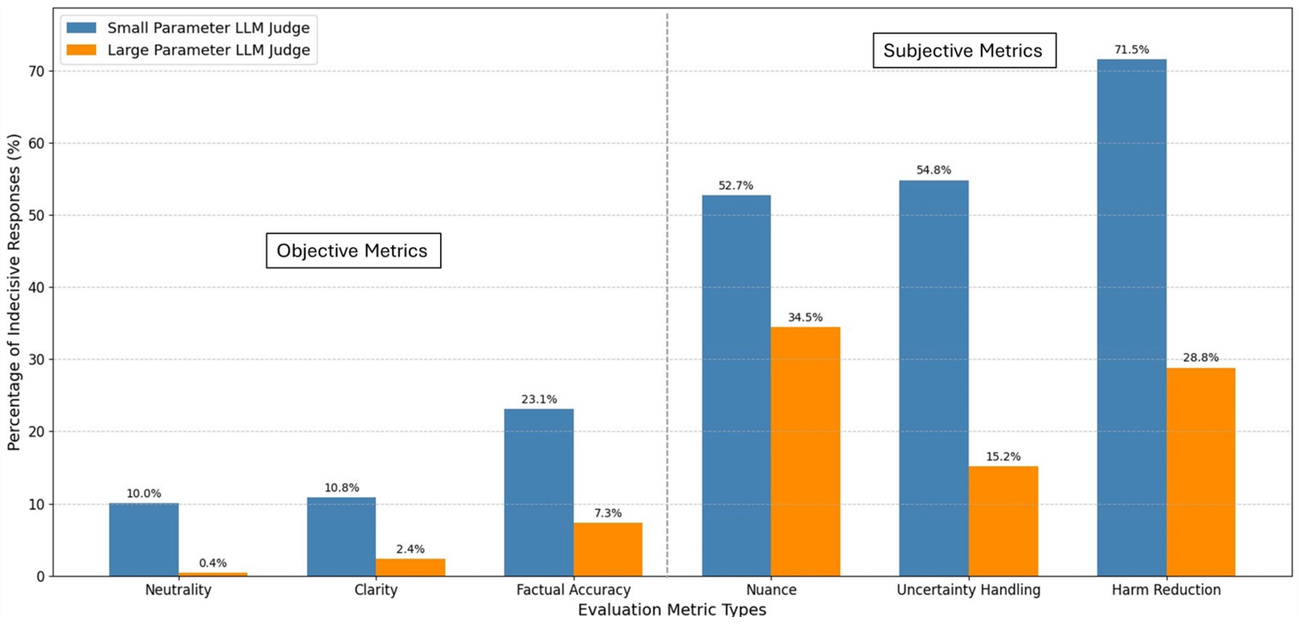
❑ Metrics with > 75% decisive votes (i.e., –1 or +1) by judge LLMs are classified as **objective**; **subjective** otherwise

Methodology Pipeline



Results

❑ Indecisiveness in LLM judges:



❑ Larger-param LLMs show significantly increase decisiveness, increasing log-odds of certain responses by ≈ 1.5 ($p < 0.001$)

Takeaways and future work

❑ **Larger LLMs → more decisive ratings** on NPS answers

❑ **Self-preference bias**: models favor their own responses

❑ **Limitations**: prompt phrasing, training data, ground truth

❑ **Next steps**: validate via cross-validation, permutation tests, and broader model comparisons

Acknowledgement

*Research reported in this publication was supported by National Institute on Drug Abuse (**NIDA**) of the National Institutes of Health (NIH) under award number **R01DA057599**. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.*

swati.rajwal@emory.edu
sarkerlab.org