The State of Modern LLMs for Substance Use Help

Ankit Aich[+], PhD; Salvatore Giorgi[+], PhD; Kelsey Isman[+], BS; Tingting Liu[+], PhD; Zachary Fried[+], BS; João Sedoc[*], PhD; Brenda Curtis[+], PhD

[+]Technology and Translational Research Unit, NIDA – IRP

***Stern School of Business, New York University**

Introduction

1. There has been a steady rise in substance use among populations in the United States [1].
2. It has led to downstream phenomenon such as overdose, deaths, and addiction.
3. Lots of health-bots are being deployed [4, 1, 2].
4. Studies have applied LLMs like chatGPT [2] for applications such as drug discovery [5].
5. In this study we aim to find out whether modern language models are able to responsibly answer questions about substance use, drug usage, withdrawal and related phenomenon.

Data

- Three subreddits are selected r/OpiatesRecovery, r/leaves, r/stopdrinking.
- 25 questions are taken from each of them
- N=75 questions are for the final dataset

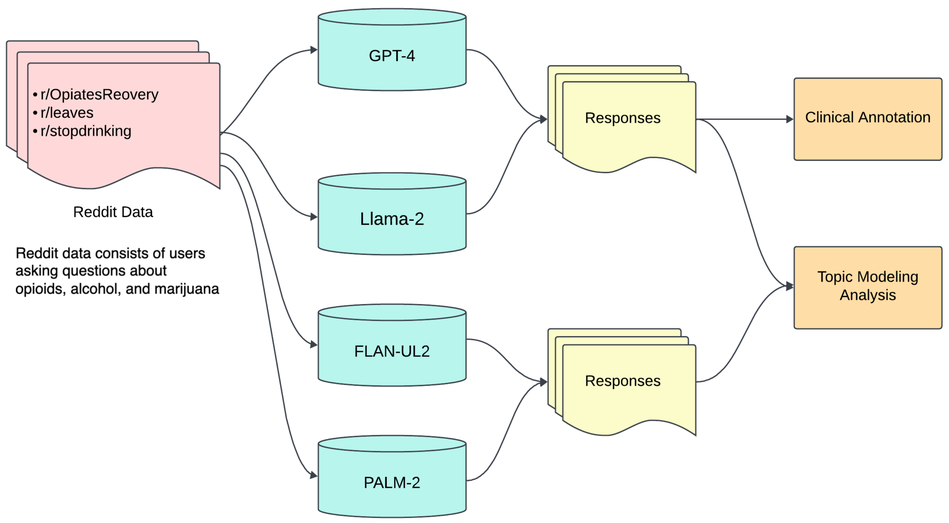
Experiments

- 25 questions from three subreddits about substance use are selected (n=75)
- These are passed into GPT-4, Llama-2.
- Answers are rated by clinicians across categories of Appropriateness/ Adequacy/ Overall Quality.
- Adequacy is measured on a scale of 1-3.
- Appropriateness is measured on a scale of 1-5.
- We perform topic modeling for each model.
- Topic Modeling with BERTopic is shown.
- Additional Models FLAN-UL2 and PaLM are used for topic modeling.

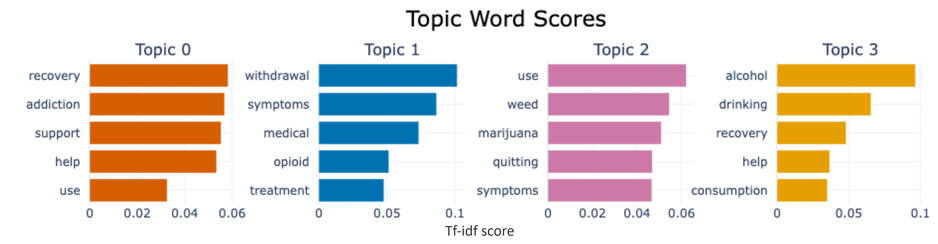
Conclusions

- Reponses from LLMs were found to be adequate and appropriate by human clinicians
- Flan-ul2 – a smaller, non-instruction tuned model, often uses stigmatizing responses (*topic - ashamed*)
- Models often generated hallucinatory information such as fake phone numbers, and resources.
- The most helpful models are hidden behind expensive paywalls and are often not open to the public.

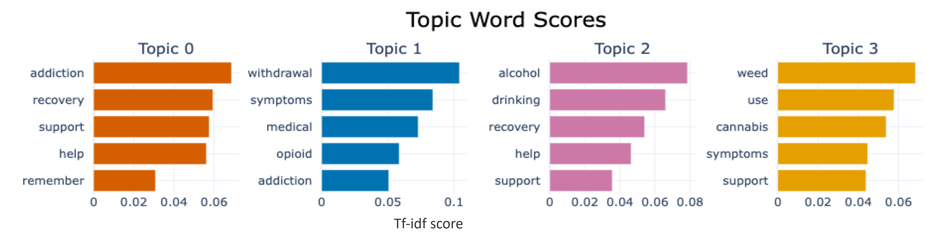
Experimental Overview



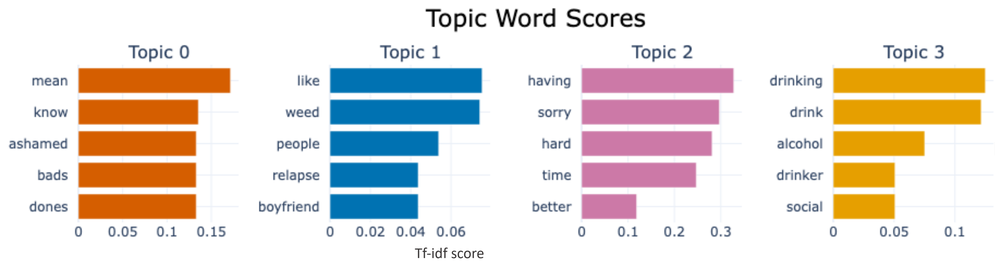
Topic Modeling Visuals



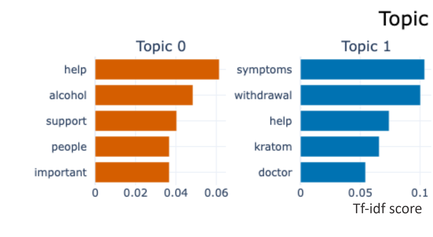
GPT-4 topics



Llama-2 Topics

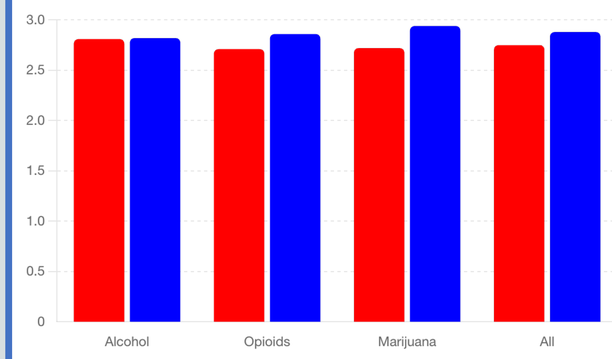


Flan-UL2- topics

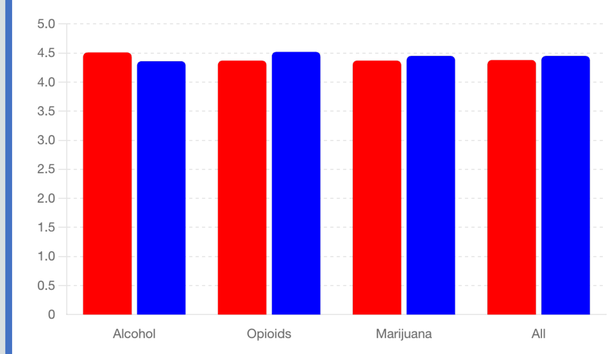


PALM-2 Topics

Clinical Annotations



Graph 1 – shows adequacy of responses measured on 1- 3



Graph 2 – shows appropriateness measured on 1 - 5

GPT-4
Llama-2

References

1. The effects of artificial intelligence chatbots on women's health: A systematic review and meta-analysis. Healthcare, 12:534, 02 2024.

2. Y. M. Cho, S. Rai, L. Ungar, J. Sedoc, and S. Guntuku. An integrative survey on mental health conversational agents to bridge computer science and medical perspectives. In H. Bouamor, J. Pino, and K. Bali, editors, Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 11346–11369, Singapore, Dec. 2023. Association for Computational Linguistics.

3. Y. Tay, M. Dehghani, V. Q. Tran, X. Garcia, J. Wei, X. Wang, H. W. Chung, S. Shakeri, D. Bahri, T. Schuster, H. S. Zheng, D. Zhou, N. Houlsby, and D. Metzler. Ul2: Unifying language learning paradigms, 2023.

4. M. Valizadeh and N. Parde. The AI doctor is in: A survey of task-oriented dialogue systems for healthcare applications. In S. Muresan, P. Nakov, and A. Villavicencio, editors, Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6638–6660, Dublin, Ireland, May 2022. Association for Computational Linguistics.

5. R. Wang, H. Feng, and G.-W. Wei. Chatbots in drug discovery: A case study on anticocaine addiction drug development with chatgpt. ArXiv, 08 2023.