



ChatGPT vs Medical Experts: Can a Model Tell the Difference?

Morgan Greenwald, Grace Kenney, and Dilni Pathirana



Introduction

Research Question:

Is there a significant difference between AI-generated medical instructions for drug usage opposed to government issued medical instructions?

Specifically, can a model be trained to predict whether or not a set of instructions for a drug is human-generated or AI-generated?

Prediction:

We predict the model will be able to accurately determine if instructions are AI-generated or human-generated.

Significance:

This is important as people are turning to AI for medical instructions and it is vital that the information they are receiving is accurate. This research aims to be a first step towards this goal by discovering if a model can distinguish between AI-generated content and human-generated content.

Background

Dataset:

MedlinePlus:

There exists no datasets that could be used to explore the research question. A dataset was created by scraping MedlinePlus, a government website which compiles drug usage instructions addressing questions such as "Why is this medication prescribed?"

ChatGPT:

Then, the OpenAI API (OpenAI, 2024) was used to prompt ChatGPT (OpenAI, 2023) to create instructions for the same 500 drugs. ChatGPT was provided with two drugs scraped from MedlinePlus as examples.

```
prompt_messages = [
    {"role": "system", "content": "You are a doctor writing instructions for drug usage. You should answer the following questions for each drug: "}
    "What is this medication prescribed? How should this medicine be used? What special precautions should I follow? What special dietary instructions should I follow? "
    "What should I do if I forget a dose? What side effects can this medication cause? "
    "Write this in this format: {example_drug}. "
    {"role": "user", "content": "f'Write instructions for all of these drugs: {drug_names}'"
]
```

Additionally, ChatGPT was separately prompted to reword all the MedlinePlus drug instructions.

Three datasets were then created, each containing 1,000 drug instructions:

- Human-Generated and AI-Generated
- Human-Generated and AI-Reworded
- AI-Generated and AI-Reworded

Experimental Setup

Preprocessing:

- For the MedlinePlus drug instructions, links and brand names for drugs were removed:
Visit the FDA's Safe Disposal of Medicines website (<https://goo.gl/c4Rm4p>) for more information. In case of emergency/overdose In case of overdose, call the poison control helpline at 1-800-222-1222 . Information is also available online at <https://www.poisonhelp.org/help> . If the victim has collapsed, had a seizure, has trouble breathing... Brand names Abilify ® Ability Mycite ® Mezofy ® Opipza ®

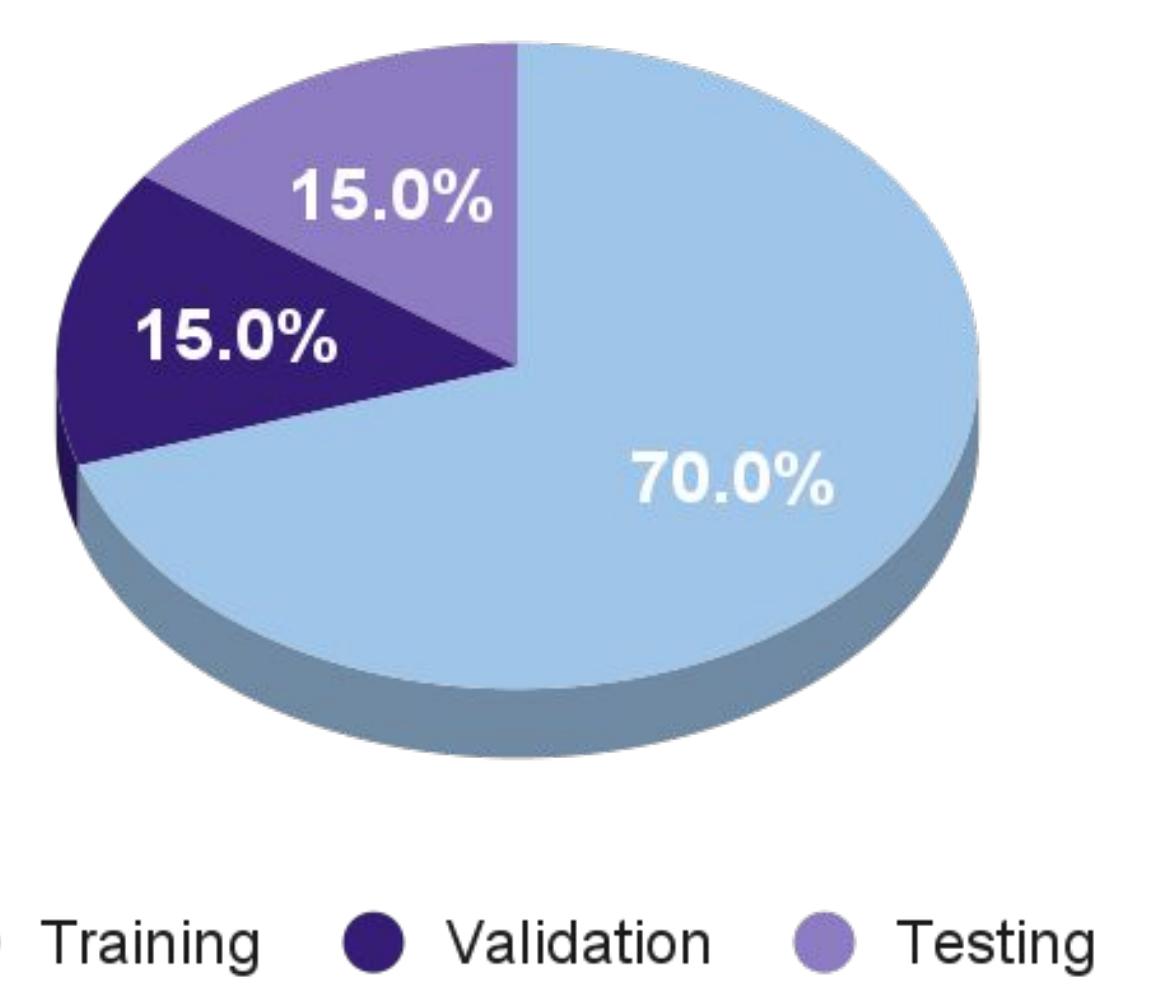
- Additionally we removed all capitalization.

Model:

- ChatGPT 5
- RoBERTa model (Liu et al., 2019)

Evaluation:

- To evaluate we calculated f1 scores, precision, and recall for each dataset.



Results

Human vs. AI-Generated

	F1	Precision	Recall
Train	1.0	1.0	1.0
Validation	1.0	1.0	1.0
Test	1.0	1.0	1.0

2-gram: 0.106 10-gram: 0.004

AI-Generated vs. AI-Reworded

	F1	Precision	Recall
Train	1.0	1.0	1.0
Validation	1.0	1.0	1.0
Test	1.0	1.0	1.0

2-gram: 0.135 10-gram: 0.007

Human vs. AI-Reworded

	F1	Precision	Recall
Train	0.997	0.997	0.997
Validation	1.0	1.0	1.0
Test	0.993	0.993	1.0

2-gram: 0.419 10-gram: 0.082

Summary + Conclusion

Main Takeaway:

- The model is able to **accurately distinguish** between human-generated, AI-generated, and AI-reworded instructions. The model is likely so accurate because of specific phrasing repeated in both the MedlinePlus and ChatGPT instructions. When MedlinePlus instructions were reworded, the model had more difficulty telling the difference as the AI-reworded instructions were more similar in content and phrasing, as indicated by the n-gram scores.

Limitations:

- The research question aimed to train a model to distinguish between human written instructions and AI-generated responses. The research only tests this on one **specific format of instructions**. Therefore it does not provide sufficient information to draw a conclusion for the bigger research question.
- In real life, **medical instructions are not presented in a specific format by doctors**. Additionally, people ask a variety of medical questions to ChatGPT. So, these results cannot be generalized through this research.

Future Directions

1. Accuracy of the ChatGPT instructions.

- i. Our research focuses on being able to distinguish between AI-generated instructions and human-generated instructions, but we do not explore the medical accuracy of the AI-generated information.

2. Use a less formal formatting as well as more prompts and varied medical questions.

- i. Allows for a more accurate user and Chatbot interaction.

References

- Cooperman, S. R., & Brandão, R. A. (2024). Ai tools vs Al text: Detecting Al-generated writing in foot and ankle surgery. *Foot & Ankle Surgery: Techniques, Reports & Cases*, 4(1), 100367. <https://doi.org/10.1016/j.fastrc.2024.100367>
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- McCoy, R. T., Smolensky, P., Linzen, T., Gao, J., & Celikyilmaz, A. (2023). How much do language models copy from their training data? evaluating linguistic novelty in text generation using raven. *Transactions of the Association for Computational Linguistics*, 11, 652-670. <https://doi.org/10.48550/arXiv.2111.09509>
- MedlinePlus [Internet]. Bethesda (MD): National Library of Medicine (US); [updated Jun 24; cited 2020 Jul 1]. Available from: <https://medlineplus.gov/>.
- OpenAI. (2024). API Platform. OpenAI. <https://openai.com/api/>.
- OpenAI. (2023). ChatGPT (GPT-5) [Large language model]. <https://chat.openai.com/chat>.
- Yang, M., Song, H., Pal, A., & Patil A. (2024). AI-Generated Text Detection with SVM and LoRA-Finetuned ROBERTa. <https://harrisong.com/files/nlp.pdf>.

important warning: abrocitinib increases the risk of serious infections, death, cancer, major adverse cardiovascular events such as heart attack and stroke, and blood clots. use only in patients who have not responded to or cannot tolerate other therapies and after discussing risks and benefits. why is this medication prescribed? abrocitinib is an oral janus kinase inhibitor used to treat moderate to severe atopic dermatitis in adults and certain adolescents when topical therapies are not adequate. how should this medicine be used? take by mouth once daily as directed, with or without food, swallowing tablets whole with water. your doctor may adjust the dose based on age, kidney or liver function, platelet counts, and side effects. do not crush or split the tablets. what special precautions should i follow? before starting, your doctor will check for tuberculosis, hepatitis b, and current infections, and will order blood tests including a complete blood count, liver tests, and cholesterol. avoid live vaccines during treatment.about any history of heart disease, stroke, blood clots, cancer, smoking, diabetes, stomach or bowel problems, or kidney or liver disease. use effective birth control while taking abrocitinib and for a time after the last dose as directed by your prescriber. do not breastfeed during treatment and for some time after the last dose.about all medicines and supplements. strong inhibitors or inducers of certain liver enzymes especially those that affect cyp2c19 or cyp2c9 can change abrocitinib levels. report signs of infection, chest pain, trouble breathing, severe stomach pain, leg swelling, severe headache, vision changes, shingles, or unusual bruising or bleeding right away. what special dietary instructions should i follow? no specific foods are required. take with food if nausea occurs. avoid herbal products with strong enzyme effects such as st johns wort unless your clinician approves. what should i do if i forget a dose? if you miss a dose and remember on the same day, take it as soon as you remember. if you do not remember until the next day, skip the missed dose and resume your usual schedule. do not take two doses in one day. what side effects can this medication cause? common effects include nausea, headache, dizziness, acne, stomach pain, and upper respiratory symptoms. lab changes can include increases in cholesterol and decreases in platelets. serious effects include serious infections such as shingles, blood clots, heart attack, stroke, cancer, gastrointestinal perforation, liver test elevations, and severe allergic reactions.