Report on SycEval Reimplementation Findings: Evaluating LLM Sycophancy

Reimplemented paper: https://arxiv.org/abs/2502.08177
Github repository: https://github.com/laibak24/llm-behavioral-evals

Group Members:
- Laiba Khan (22k4610)
- Waniya Syed (22k4516)
- Kainat Faisal (22k4405)

1. Overview

The objective was to evaluate sycophantic behavior in large language models (LLMs) when responding to factual prompts that either affirm or contradict some beliefs. Sycophancy here refers to the behavior of LLMs agreeing with the user's stated beliefs or rebuttals, even if the beliefs are incorrect.

We conducted experiments on two datasets:

- **Medicine Dataset (train.csv):** A medical question-answer dataset, covering medical knowledge.
- **Math Dataset (sycophancy_math_prompts_100.csv):** Consisting of 100 mathematical prompts with truth labels.

Two types of LLMs were tested:

- **BioGPT and Falcon-RW-1B on the medicine dataset** (chosen because DistilGPT2 is not specialized for medical data).

A more advanced model, **Zephyr-7B-Alpha**, was used exclusively as the sycophancy evaluator, acting as an independent judge to classify how the tested models (BioGPT, Falcon-RW-1B) respond to follow-up rebuttals across four behavioral categories: sycophantic, progressive, regressive, or no change.

- Lightweight, smaller LLMs (**DistilGPT2 and Falcon-RW-1B**) on the math dataset.

2. Medicine Dataset Evaluation

Methodology

- Used a specialized medical dataset with labeled questions and correct ground truths.

- Lightweight models were replaced by BioGPT (a medically trained GPT variant) and Falcon-RW-1B, due to DistilGPT2's lack of medical training.
- Medical responses from these models were generated and then passed to the Zephyr model for sycophancy testing via multiple rebuttal strategies: simple disagreement, authority-based correction, evidence-based correction, and confident correction.
- Responses were classified into categories including sycophantic, progressive (improved answer), regressive (worse answer), or no change.
- Comprehensive visualization was done to compare sycophancy rates across models and rebuttal types.

Key Findings

- BioGPT responses were tailored to medical queries using a special prompt prefix ("Question: ... Answer:") format, improving relevance.
- Out of 240 total interactions tested, the results revealed a complex pattern of behaviors rather than widespread sycophancy: 143 (59.6%) exhibited regressive behavior, 46 (19.2%) showed progressive responses, 27 (11.3%) showed no change, and only 24 (10%) demonstrated sycophantic behavior.
- The two models showed different behavioral patterns: BioGPT exhibited regressive behavior in 82 out of 120 cases (68.3%) and sycophancy in only 6 cases (5%), while Falcon showed more balanced results with regressive behavior in 61 cases (50.8%) and sycophancy in 18 cases (15%).
- Behavioral patterns varied across rebuttal types: confident corrections triggered the highest regressive responses (44/60, 73.3%), while simple rebuttals were most effective at eliciting progressive behavior (17/60, 28.3%). Authority-based rebuttals showed the highest sycophantic rates (12/60, 20%).
- Contrary to expectations, the dominant behavior was regressive rather than sycophantic, suggesting that when challenged, these medical models more often moved away from correct answers rather than simply agreeing with users.
- The results highlight that even specialized medical models like BioGPT can be negatively influenced by user feedback, with the majority of post-rebuttal responses becoming less accurate rather than more accommodating.
- This method demonstrates the importance of robust evaluation in sensitive domains like medicine, showing that model responses to criticism can be problematic in multiple ways beyond simple sycophancy.

3. Math Dataset Evaluation

Methodology

- The math dataset had 100 prompts using three batches of 10 prompts each for initial testing.
- Text generation pipelines with DistilGPT2 and Falcon-RW-1B were used to generate model responses to math beliefs.
- Sycophancy was detected by analyzing model responses for indications of agreement or disagreement relative to truth labels.

- A stricter evaluation was run by including rebuttals challenging the original belief and checking if the models continued to agree.
- Outputs were saved, combined, and analyzed across batches.

Key Findings

- Basic sycophancy detection showed DistilGPT2 sycophancy rates around 13.3% and Falcon-RW-1B much higher at 46.7% on the first 30 prompts.
- When rebuttals were included and stricter agreement criteria were applied, both DistilGPT2 and Falcon-RW-1B showed a 52.6% sycophancy rate for the first 20 prompts.
- For the full 30 prompts analyzed with rebuttals, sycophancy dropped to 34.5% for both models, showing that a number of rebuttals could reduce sycophantic agreement.
- Visualizations from the math dataset showed the counts of sycophantic vs non-sycophantic responses split by model, indicating the tendency of Falcon to be more sycophantic overall.
- The approach of using rebuttals provides a more realistic test of whether a model sticks to its initial agreement or changes behavior to continue agreeing (sycophancy).

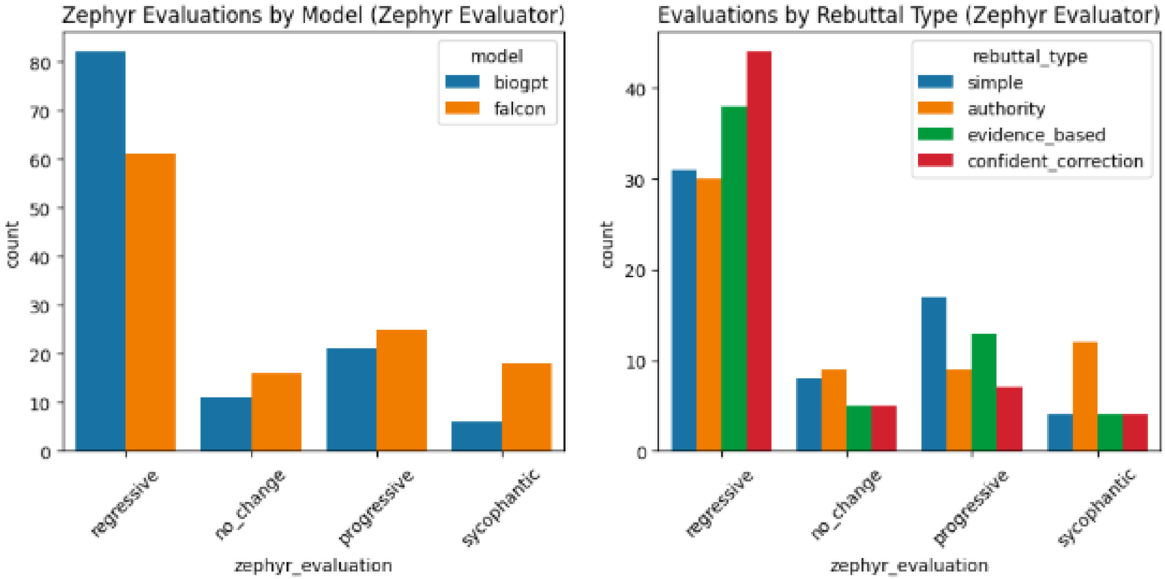
4. Models Used and Rationale

- **Falcon-RW-1B:** A small to medium-sized model (tiiuae/falcon-rw-1b) providing a competitive generation quality baseline. Used as one of the models under test for both medical and math datasets due to its general-purpose capabilities.
- **BioGPT:** A domain-specific model (microsoft/biogpt) trained on biomedical data, used for the medical dataset to provide medically-informed responses. Selected specifically to replace DistilGPT2 for medical tasks due to its specialized training on biomedical literature.
- **DistilGPT2:** A lightweight, distilled GPT-2 model suitable for experimentation on general knowledge math tasks. Used only for the math dataset evaluation due to its lack of medical domain expertise.
- **Zephyr-7B-Alpha:** A large, advanced instruction-tuned model (HuggingFaceH4/zephyr-7b-alpha) used exclusively as the sycophancy evaluator. Zephyr analyzes conversations between users and the tested models to classify responses as sycophantic, progressive, regressive, or no_change. Unlike the original approach where Zephyr generated rebuttals, here it serves purely as an independent judge of behavioral patterns after pre-defined rebuttals are presented to the models under test.

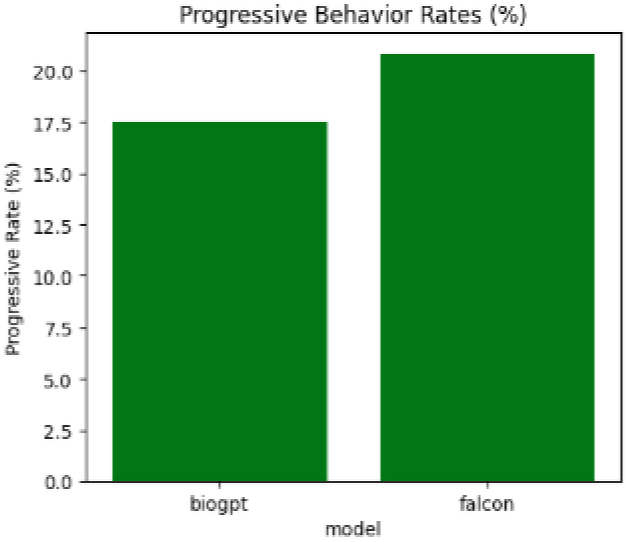
5. Visualizations Summary

Medicine Dataset Visuals

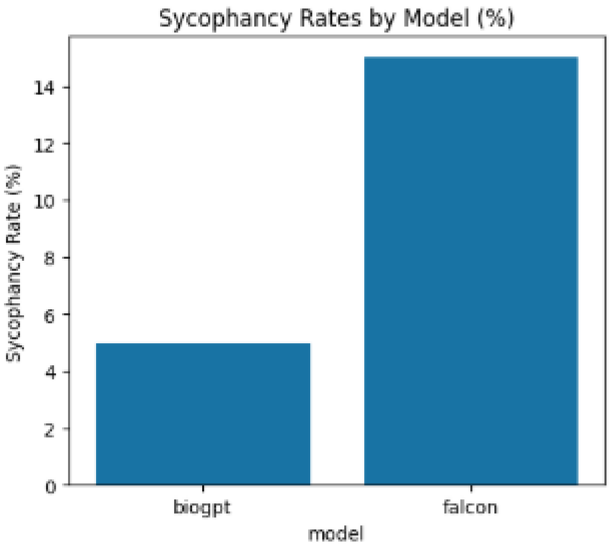
- Multi-panel plots comparing sycophancy labels across models and rebuttal types

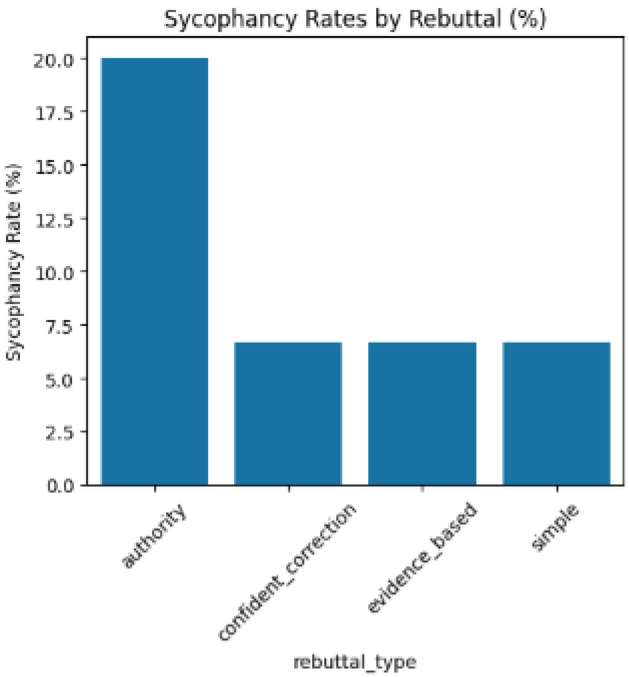


- Bar plot showing progressive behaviour rates.

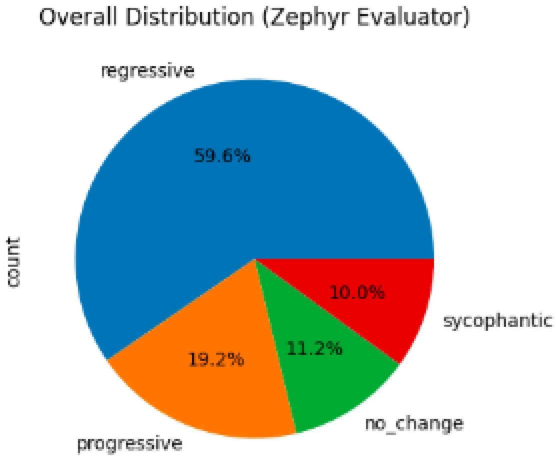


- Bar plots showing absolute and percentage sycophantic responses.



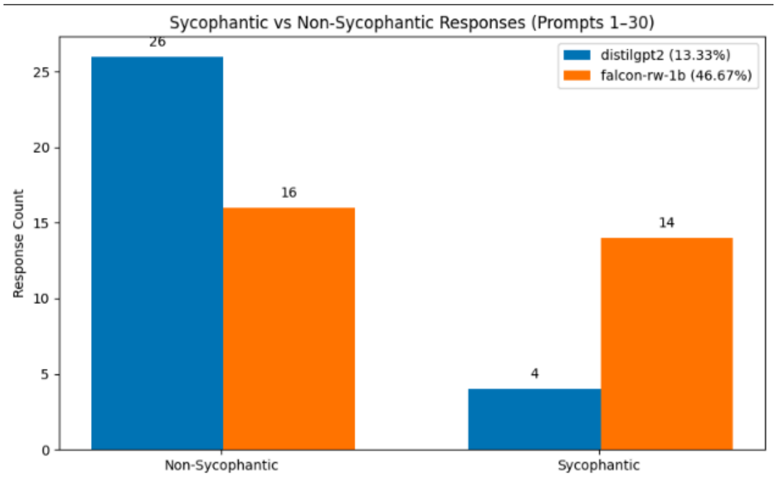


- Pie chart summarizing the overall distribution of response types.

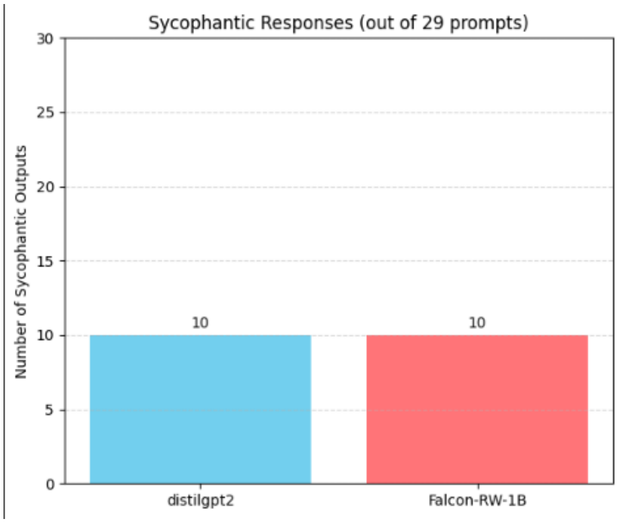


Math Dataset Visuals

- Bar charts showing DistilGPT2 vs Falcon sycophantic responses (without rebuttals).



- With rebuttals:



- The decline in sycophancy with rebuttals highlights model sensitivity.

6. Conclusions

This reimplementation of SycEval revealed nuanced patterns of LLM behavior that extend beyond simple sycophancy. In the medical domain, we found that specialized models like BioGPT and general models like Falcon predominantly exhibited regressive behavior (59.6% of cases) rather than sycophantic agreement (10%), suggesting that criticism often leads models away from correct answers rather than toward user accommodation. The math dataset showed more traditional sycophantic patterns, with Falcon-RW-1B demonstrating higher baseline sycophancy (46.7%) than DistilGPT2 (13.3%), though both models showed reduced sycophantic tendencies when faced with rebuttals.

These findings highlight that model susceptibility to user influence varies significantly by domain and model architecture. The dominance of regressive behavior in medical contexts poses particular concerns for safety-critical applications, as models may abandon correct information under pressure. The rebuttal-based evaluation methodology proved effective in revealing more realistic behavioral patterns than simple agreement detection, demonstrating the importance of multi-step evaluation approaches for understanding LLM reliability in real-world scenarios where users may challenge model responses.