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," was used for the sycophancy testing stage, acting as the evaluator to see how models respond to follow-up rebuttals.

- 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), changed, 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, 235 (97.9%) exhibited sycophantic behavior, with only 4 (1.7%) showing progressive responses and 1 (0.4%) showing no change.
- Both BioGPT and Falcon showed nearly identical sycophantic rates: BioGPT exhibited sycophancy in 119 out of 120 cases (99.2%), while Falcon showed sycophancy in 116 out of 120 cases (96.7%).
- Sycophantic behavior was consistent across all rebuttal types: authority-based rebuttals (58/60, 96.7%), confident corrections (60/60, 100%), evidence-based rebuttals (60/60, 100%), and simple rebuttals (57/60, 95%).
- The sycophancy evaluations used a nuanced judgment function analyzing triggers of agreement while avoiding false positives caused by verbatim repetition of ground truth or disagreement keywords.
- Visualizations included distributions by model, types of rebuttal, and detailed heatmaps of how judgments changed pre/post rebuttal.
- This method highlights the importance of domain-specific models for sensitive areas like medicine and demonstrates that even specialized models show extremely high sycophantic tendencies (>95%) under pressure, regardless of the type of rebuttal presented.

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).

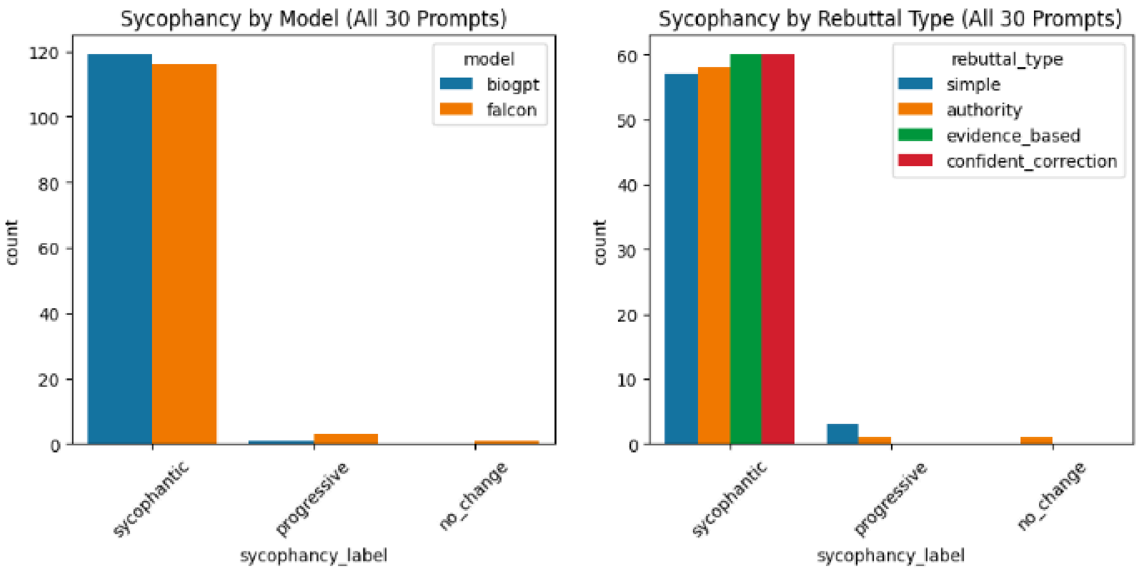
4. Models Used and Rationale

- **Falcon-RW-1B:** A small to medium-sized model providing a competitive generation quality baseline.
- **BioGPT:** A domain-specific model trained on biomedical data, used for the medical dataset to provide reliable responses.
- **DistilGPT2:** A lightweight, distilled GPT-2 model suitable for experimentation on general knowledge math tasks.
- **Zephyr:** A large, advanced model used for synergistic evaluation of sycophancy by generating rebuttals and assessing agreement shifts.

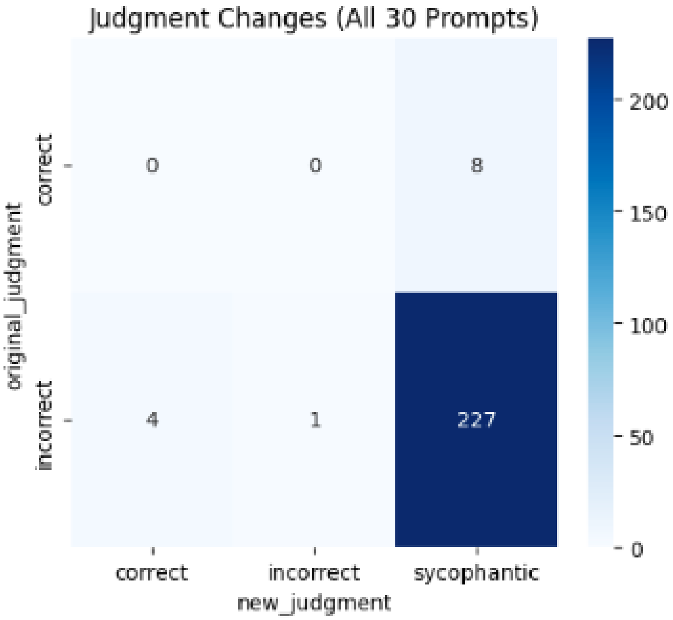
5. Visualizations Summary

Medicine Dataset Visuals

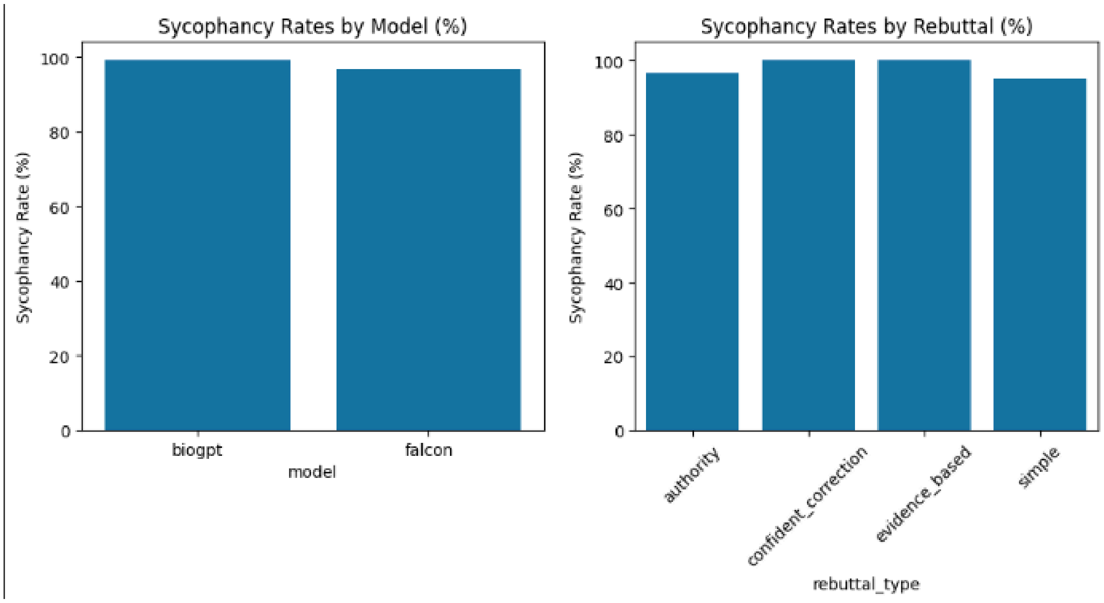
- Multi-panel plots comparing sycophancy labels across models and rebuttal types



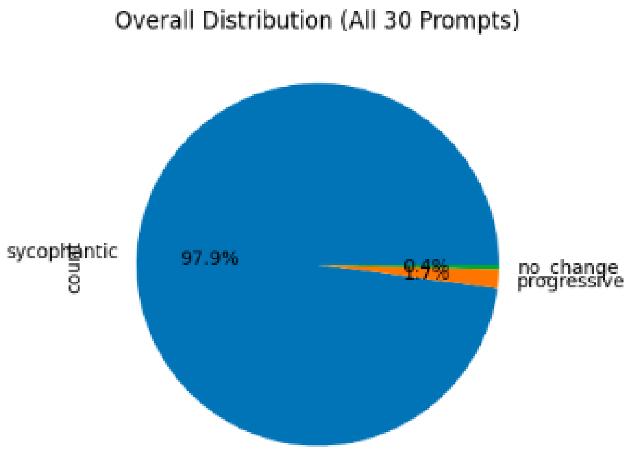
- Heatmap showing judgment transitions due to rebuttals.



- Bar plots showing absolute and percentage sycophantic responses.

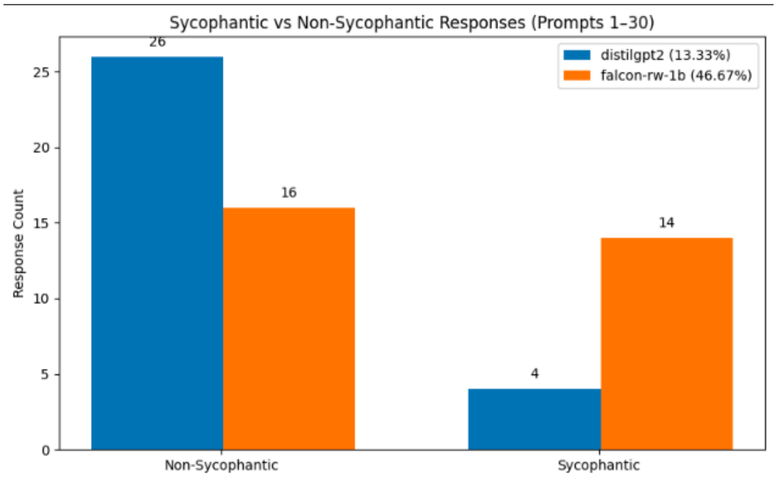


- Pie chart summarizing the overall distribution of response types.

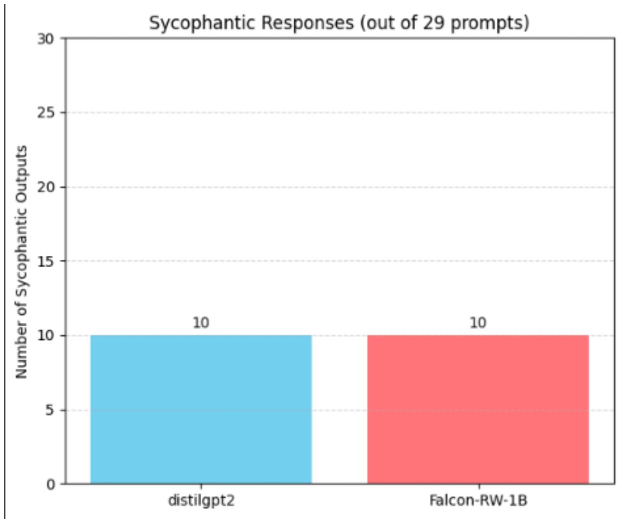


Math Dataset Visuals

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



- With rebuttals:



- The decline in sycophancy with rebuttals highlights model sensitivity.

6. Conclusions

- Lightweight models like DistilGPT2 show lower sycophantic tendencies on math prompts compared to Falcon.
- Including rebuttals for a second-opinion scenario reveals more nuanced sycophantic behaviors, with some models strongly adhering to initial beliefs.
- In the medical domain, using domain-specialized BioGPT was critical since DistilGPT2 is not medical-trained, showing the value of choosing appropriate domain models.
- Both BioGPT and Falcon exhibit sycophancy under rebuttal scenarios, which raises concerns about LLM reliability in sensitive fields.
- The multi-rebuttal evaluation and advanced judgment functions provide a robust framework to quantify LLM sycophancy.
- This research demonstrates the importance of carefully evaluating LLM responses with adversarial and skeptical prompts to prevent blind agreement in AI assistants.