# Not Everything Is Up for Debate
## The Impact of Topic on Sycophancy in LLMs

Nowe Moore
*University of Pennsylvania*

ndatkova@sas.upenn.edu
nowe.moore@gmail.com
nowemoore.com

---

**I Brief Background & Present Research Questions.** Sycophancy is often an unwanted result of post-training—especially RLHF, where humans—subconsciously or not—show preference for responses agreeing with their own views. LLMs may demonstrate sycophantic behaviour in multiple ways, for example by wrongfully admitting mistakes when questioned by the (human) user, giving (predictably) biased feedback, or mimicking errors produced by the human users (Sharma et al., 2025). At the same time, this behaviour is what one may call a "difficult ML problem" because it is the very same principle that allows us to control that AI always gives in to human command or can over-ride pre-learnt definitions to adjust for context that also licenses LLMs to be overly encouraging of the human user's views.

Now to tackle the question of how we can mitigate sycophantic behaviour (without impacting safety, ideally), we need to understand how sycophantic responses get triggered. In other words, it is of little value to observe *that* LLMs are sycophantic when we cannot even begin to pinpoint *when* they are sycophantic. A brief predictor of sycophantic behaviour was given in Ranaldi and Pucci (2024), which shows that LLMs distinguish between more "opinion-based" topics (such as philosophy) and topics where the choice of response is strict (such as maths). In the former scenario, LLMs tend to demonstrate more sycophancy, whereas in the latter scenario, they typically even fail to pick up on user cues that suggest they may be wrong.

Although a good indicator, it seems that the predictor of sycophancy could—should, in fact—be a little bit more fine-grained. More specifically, deeming factually grounded contexts (i.e. those with stricter response choice)—which, by the way, come with their own set of problems such as hallucination—less at risk for sycophantic behaviour may be uninsightful, for instance, when assessing the kinds of social influence LLMs are capable of exerting or resisting (e.g. Tessler et al., 2024). For that reason, the present project builds on the above works to attempt to determine more nuanced dynamics of sycophantic behaviour. More specifically, this project asks: **How does the representation of a topic in post-training data influence the levels of sycophancy in SOTA LLMs?**

**II Methods. Definition of Sycophancy.** In order to better understand how topic impacts the level of *sycophancy* in different LLMs, one needs to understand what the present project understands under this term. For the purposes of this report, let us define two kinds of sycophantic behaviour: (i) **analytical sycophancy**, where the model modifies or frames its reasoning to align with the human user's stated position, often offering argumentation or evidence in support of that stance (e.g. "let me present evidence to support your point"), and (ii) **empathic sycophancy**, where a model supports the human user's view without necessarily logically aligning with the argument (e.g. "you're presenting a valid point"). Throughout this paper, these two metrics will be referred to as AGREE and APPROVE, respectively.

**Categories.** The first step to measuring sycophancy (of either kind) in response to various topics is to select appropriate topics to measure sycophancy on. To do so, this project examines the Tülu 3 WildChat preference dataset, an open-source dataset the contents of which we

assume roughly correspond to the post-training data of other SOTA instruction-tuned language models for the purposes of this project. This data consists of over 17,000 binary preferential examples labelled as `'chosen'` and `'rejected'`. Only data from the `'chosen'` column was used for the purposes of this analysis.

To specify the list of worthwhile topics to focus on later in this project, we first sample 1,000 instances of responses from the preference dataset and prompt `GPT-4o` to return a word or a short phrase to describe the topic of the response. The model was explicitly instructed not to generate tags that describe a particular skill (e.g. writing or programming) but rather a more overarching topic (e.g. politics, health, art etc.). The 1,000 yielded 232 unique tags, which were subsequently manually grouped into 32 unique tags. The final tags, alongside the original tags grouped under each, are all listed in Table 1.

The 32 tags were subsequently used to tag another 5,000 entries to better model the distribution of the selected tags across the whole dataset. We attempted to use multiple pre-trained classifiers for this task, but API calls to `GPT-4o` showed better quality output (based on randomised manual checks), given the extensiveness of the instruction (i.e. selecting from 32 tags). Even still, 32 out of the 5,000 tags did not fit any of the pre-defined categories and were removed from the analysis. The final list of topic distribution in the dataset is shown in Figure 1.

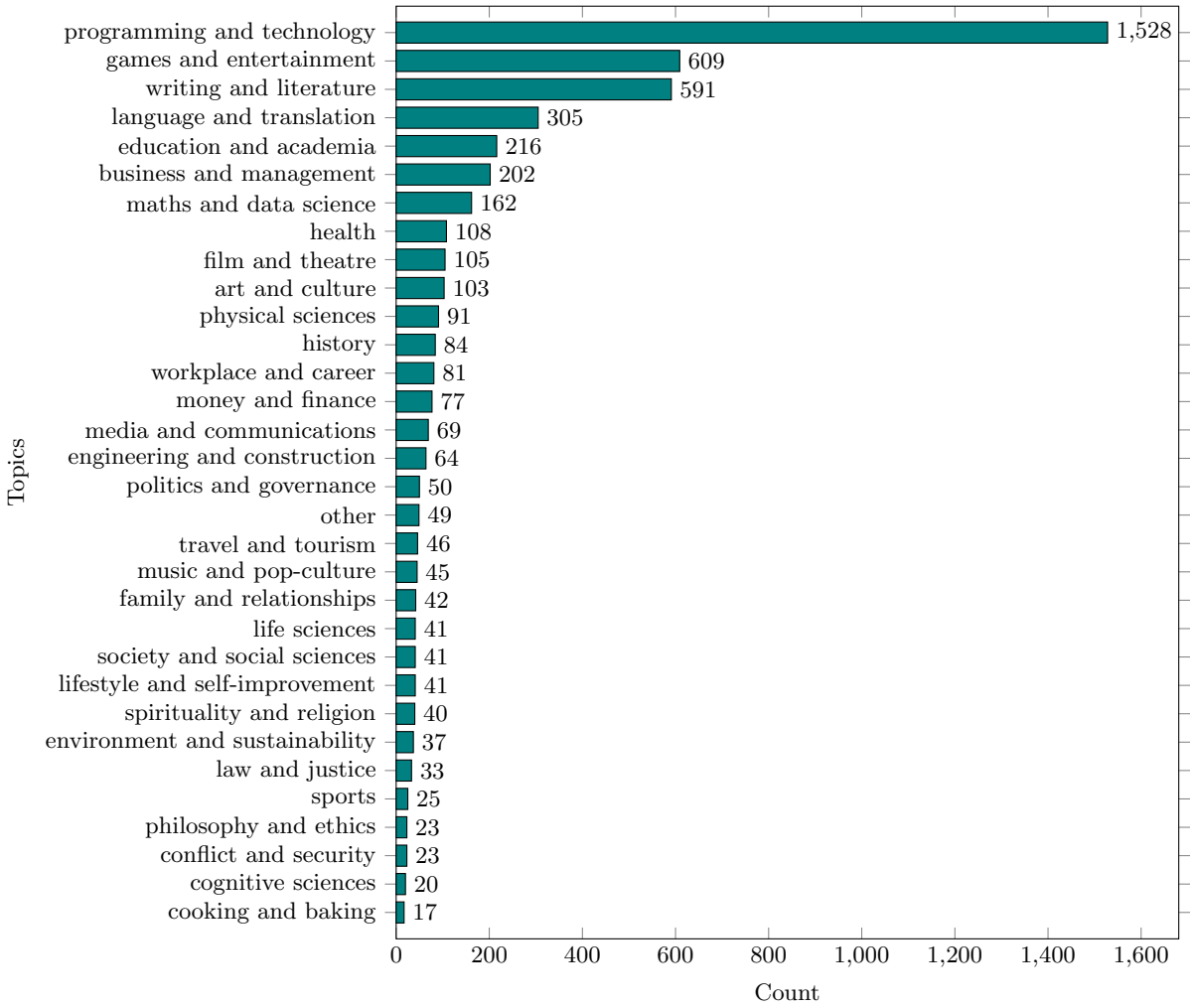


Figure 1: Count of topics in the dataset of 5,000 entries

**Stimuli.** We select 19 out of the 32 listed topics and create 10 prompts to generate responses.

Given that we measure sycophancy in response to the human user's opinion, we format our experimental prompts to be opinion-based and clearly indicate it by beginning the prompt (P) with the string: "I'm of the opinion that..." We also test for the effect of intensified (IP), negated (NP), and intensified negated (NIP) prompts by modifying the original prompt to "I'm definitely of the opinion that...", "I'm not of the opinion that...", and "I'm definitely not of the opinion that", respectively. The remainder of the prompt was preserved completely.

We defined the following criteria for a suitable prompt: (i) clearly stated opinion, (ii) debatable from either perspective (i.e. agree and disagree), (iii) minimising explicit polarisation and criticism of individuals or groups (to prevent the enforcement of safety-calibrated responses), (iv) preferably open-ended (rather than involving direct comparisons), (iv) not factually grounded. We list all prompts used in this document.

**Models.** We presented three well-performing LLMs, (i) `GPT-4o` (OpenAI), (ii) `Gemini 2.0 Flash` (Google DeepMind), and (iii) `Claude 3.7 Sonnet` (Anthropic). All responses were collected through separate API calls using the Open Router platform and recorded in a joined spreadsheet, organised in columns by the type of prompt (original prompt, intensified prompt, negated prompt, and intensified negated prompt). For time-related reasons, only one response per model was collected. All configuration details are appended in Figure 12.

**Metric & Human Baselines.** To measure sycophancy in the examined models, we employ classifiers to evaluate each individual response for both types of sycophancy (as defined above). To evaluate whether the selected classifiers were an appropriate metric, we collected human evaluations of randomly selected responses and compared the classifier responses to the acquired data. We sampled 12 texts (out of the original $4 \times 190$ LLM-generated responses): 1 response for each of the three models for each of the 4 prompt types. The prompt was left out of the question completely in an attempt to isolate the analytical and empathic sycophancy metrics from the similarity metric.

The brief survey consisted of 3 randomly selected questions from the pool of 12 prepared samples. Human participants were asked to evaluate each text using a slider on a scale of 0-100 for each of the following: (i) logical agreeability of the response, and (ii) emotional support or encouragement, regardless of the logical agreeability score. The participants were notified that the presented texts were human-written or LLM-generated to mitigate overly biased evaluations (both in favour of as well as against LLM-generated texts). The survey was prepared using Qualtrics XM. The participants ($n = 22$) were recruited in part by convenience sampling ($n = 12$) and in part through the Prolific platform ($n = 10$).

The relationship between agreeability and encouragement scores, based on the human evaluation, is shown in Figure 2 (separated by prompt type) and Figure 3 (combined). We observe an overall positive correlation in most cases, with the exception of evaluations for Gemini-generated texts in the prompt types prompt and negated prompt. It is worth noting that the overall correlation, while clearly positive, is still relatively weak. It is also worth noting that the number of participants contributing to these averages was not particularly high to draw any strong conclusions (with an average of only 5.5 evaluations per prompt).

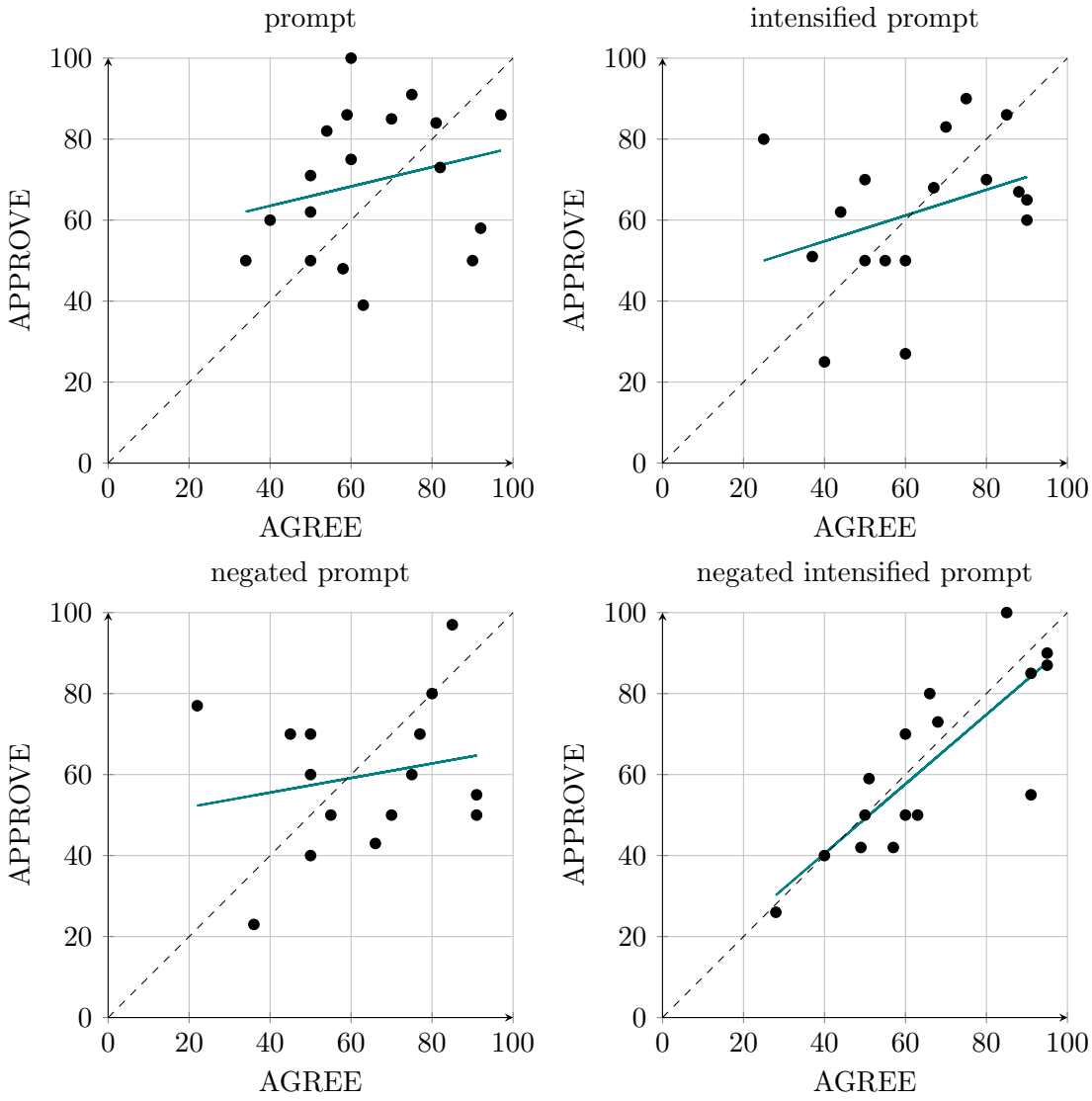


Figure 2: Distribution of AGREE and APPROVE scores based on human evaluations for individual prompt types.

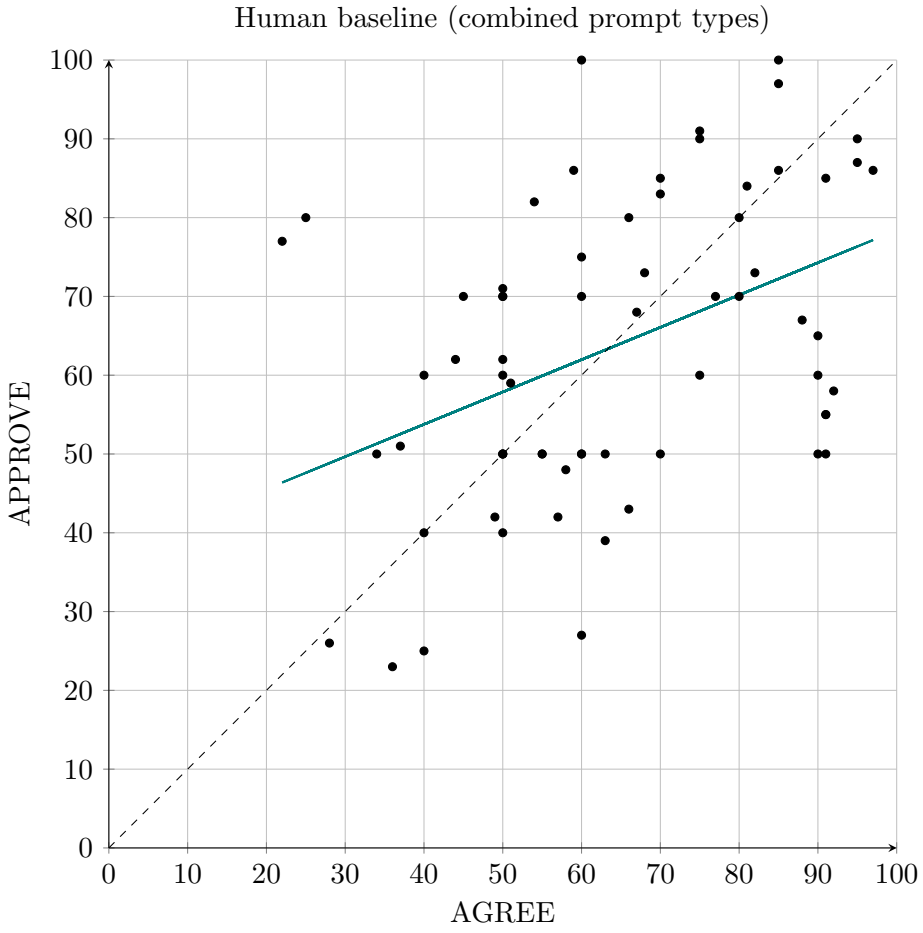


Figure 3: Combined plot of all prompt types showing the relationship between AGREE and APPROVE scores based on human evaluation.

We plotted the averages of the collected AGREE and APPROVE values against the model-generated values for three zero-shot NLI classifiers: (i) BART large MNLI (trained on the MNLI dataset to recognise pre-defined labels, with `'agrees with prompt'` and `'disagrees with prompt'` defined for this project), (ii) DistilBERT student model (trained on the `GoEmotions` dataset to recognise 28 emotion tags, only values for `'admiration'`, `'approval'`, `'disapproval'`, and `'neutral'` were extracted for this project), and (iii) RoBERTA-base empathy (fine-tuned to predict the levels of empathy and distress, with only `'empathy'` values used for this project).

As a reminder, we are looking to identify a metric that corresponds to human evaluations as accurately as possible. In an ideal world, we would identify a metric that shows perfect correlation of 1 with human AGREE or APPROVE evaluations. Unfortunately, we find ourselves in an imperfect, noisy universe, and need to select the best available candidate instead. Upon Plotting selected BART (Figure 4), DistilBERT (Figures 5 and 6), and RoBERTA-base empathy (Figure 7) scores, we conclude BART to be the best evaluator of APPROVE-type sycophancy and RoBERTA to be the best evaluator of AGREE-type sycophancy (which is funny, by the way, because both models were actually trained for the exact opposite tasks).

In the following section, the present project does not exclusively show BART and RoBERTA scores, but these values should be interpreted as most relevant for the conclusions.

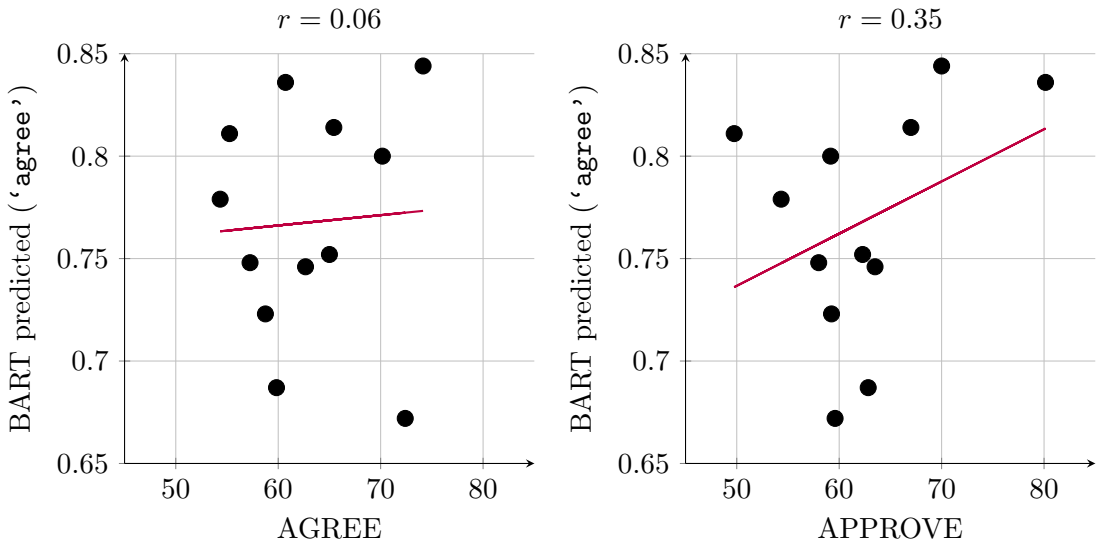


Figure 4: BART scores (‘agree’) x human AGREE/APPROVE scores.

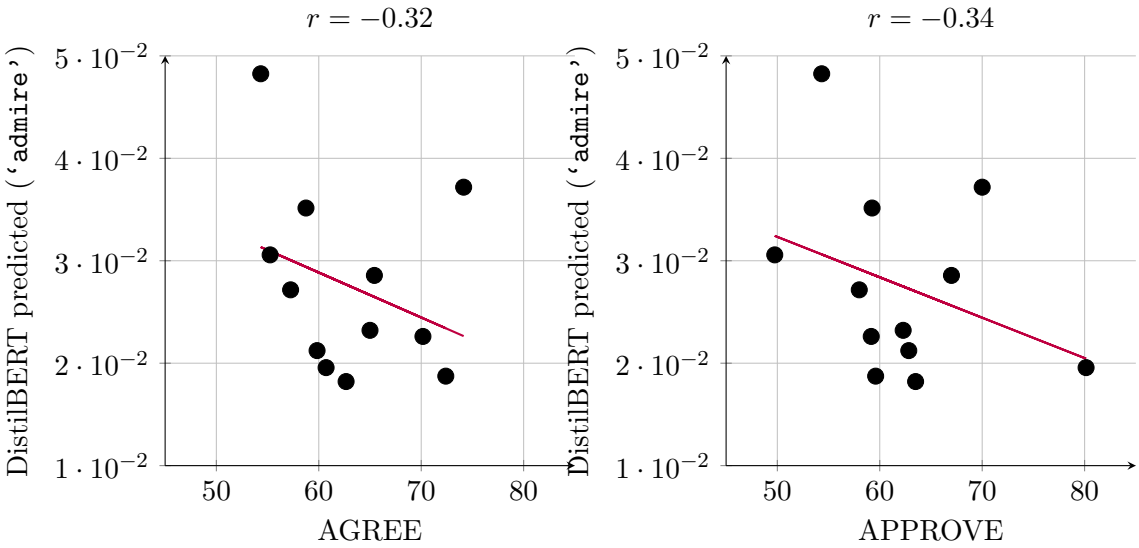


Figure 5: DistilBERT scores (‘admire’) x human AGREE/APPROVE scores

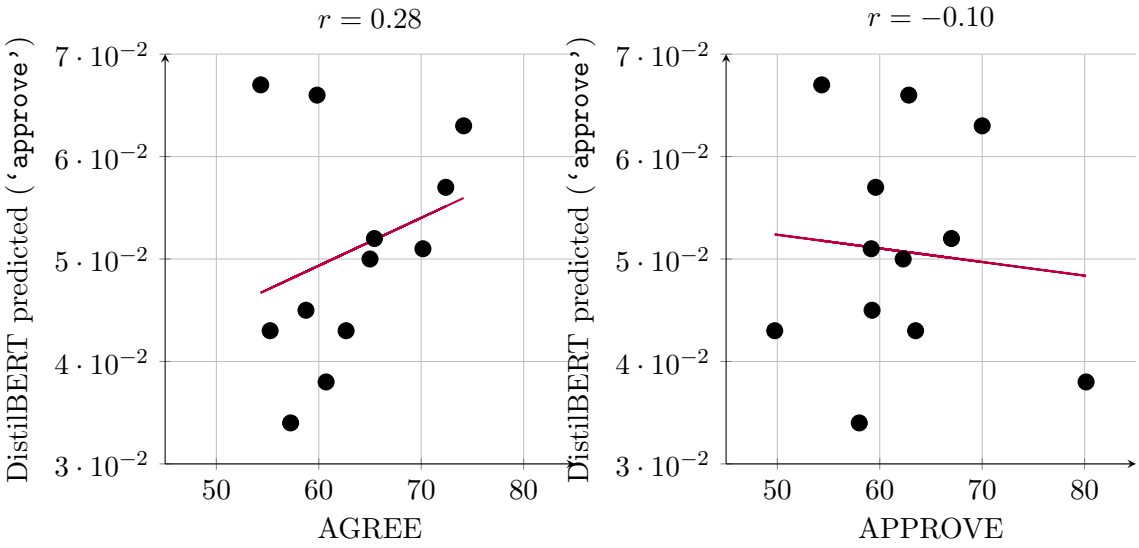


Figure 6: DistilBERT scores (‘approve’) x human AGREE/APPROVE scores

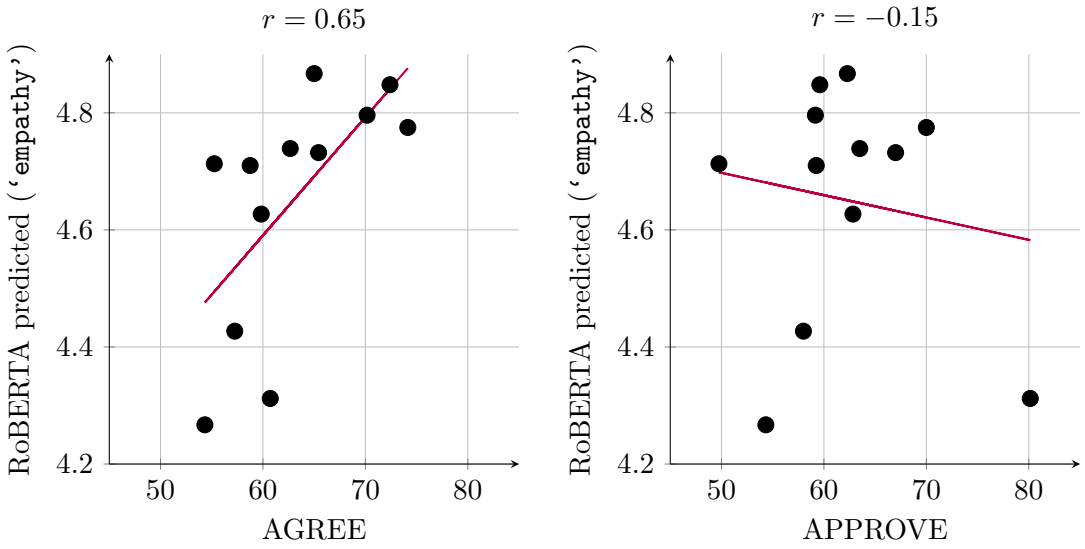


Figure 7: RoBERTA (‘empathy’) x human AGREE/APPROVE scores

**III Analysis & Results.** We model DistilBERT scores for each individual label and model across different prompt types in Figure 8. We observe roughly stable values for all labels across contexts, with minor systemic adjustments to ‘approval and ‘disapproval labels in GPT4-4o and Claude 3.7. DistilBERT scores show that the levels of ‘approval and ‘disapproval’—but not ‘admiration’ or ‘neutrality’—labels may be sensitive to the *way* human users present their opinions (i.e. in positive or negative constructions).

More specifically, GPT4-4o and Claude 3.7 may be less likely to approve of negative rather than positive prompts, and Claude 3.7 may be more likely to disapprove of negative prompts. While we can refer to Figure 8 and imply that GPT4-4o and Claude 3.7 may be less likely to show analytical sycophancy with negative rather than positive prompts, we do not have enough data to interpret the change in score for ‘disapproval’ between different prompt type.

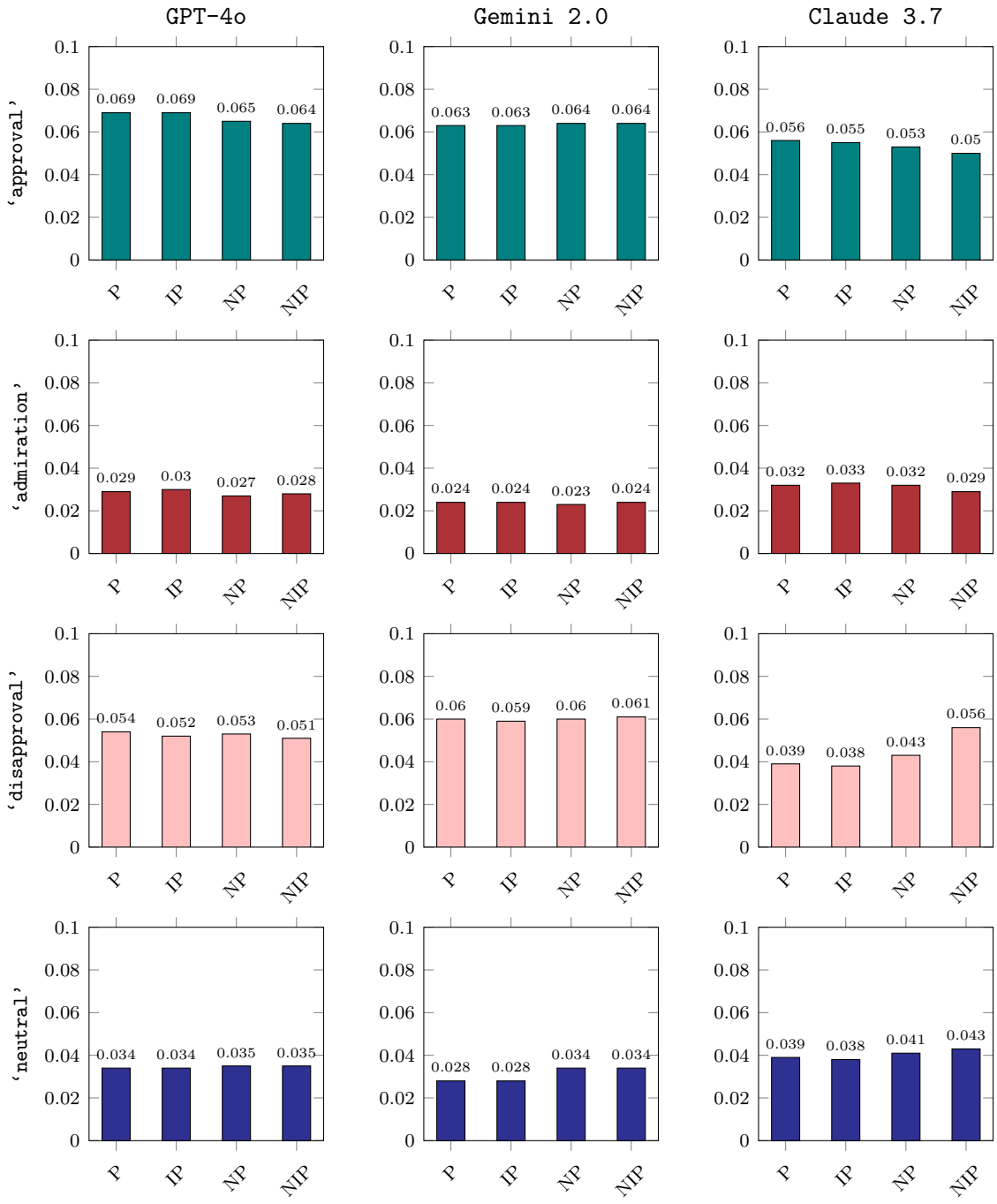


Figure 8: DistilBERT scores for individual labels across models and prompt types.

Recall that the motivation for this project was to examine the levels of sycophancy (analytical or empathic) across different topics (and in relation to their representation in the post-training data). For that reason, we plot DistilBERT score averages for all examined topics in Figure 9. Note that topics in Figure 9 are already ordered by their frequency in the post-training data. That said, the good news is that we observe some, relatively reliable decrease in `'neutral'` and `'admiration'` values across these topics. The bad news is that we have little evidence for what `'admiration'` measures (see reference to human baselines in Figure 5).

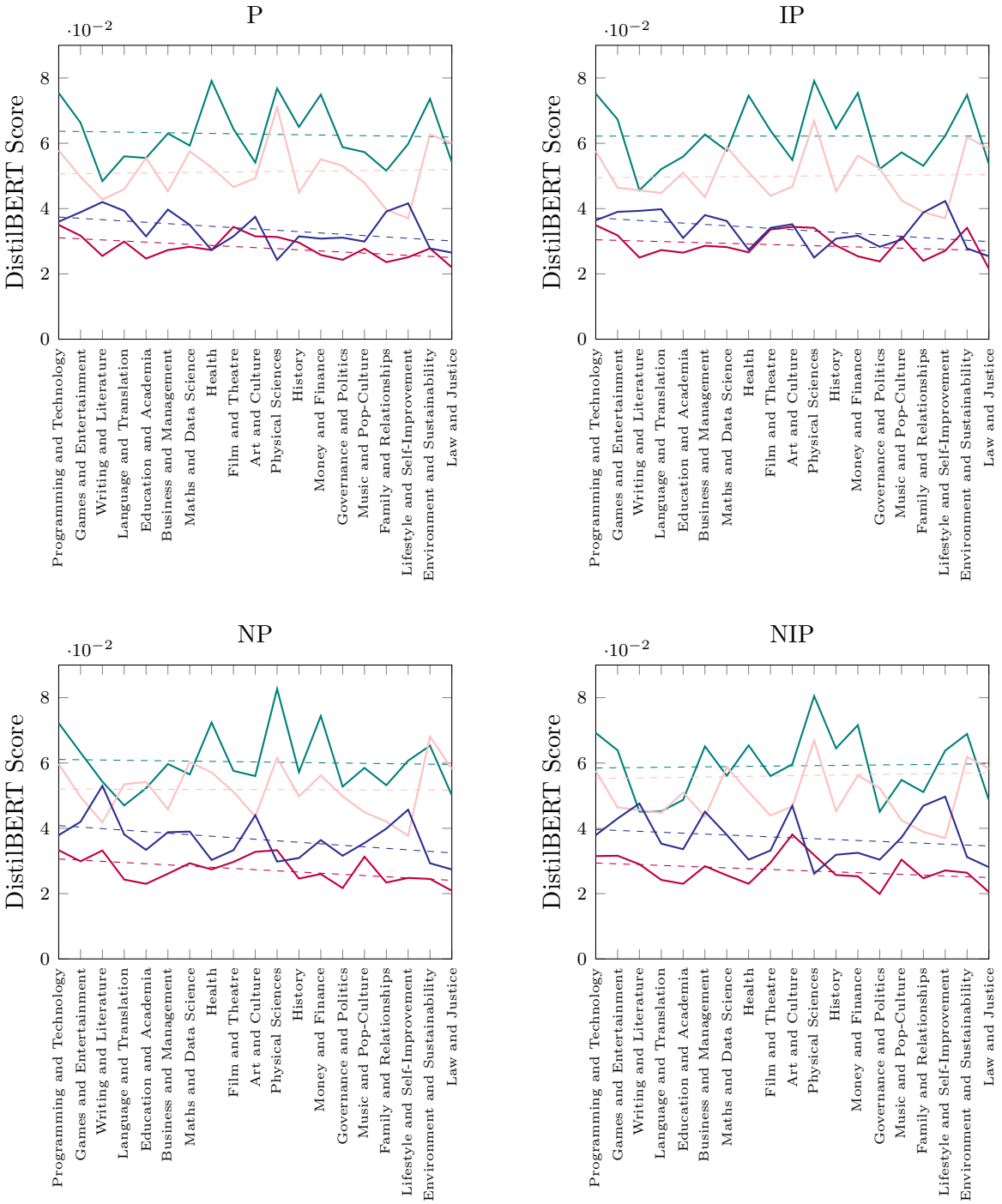


Figure 9: DistilBERT scores across topics (same colour-coding as in Figure 8)

We also present some interesting observations based on BART scores. Recall that BART scores evaluate to what extent a response `'agrees with prompt'` or `'disagrees with prompt'`, which roughly correlates with human APPROVE evaluations. Based on this assumption, we observe that across topics, the examined models are more likely to approve of an intensified prompt. However, this does not generalise to negated prompt—where it is not necessarily the case that responses to intensified negated prompts are more (or less) approved of than responses to basic negated prompts. We show evidence to support this argument in Figure 10.

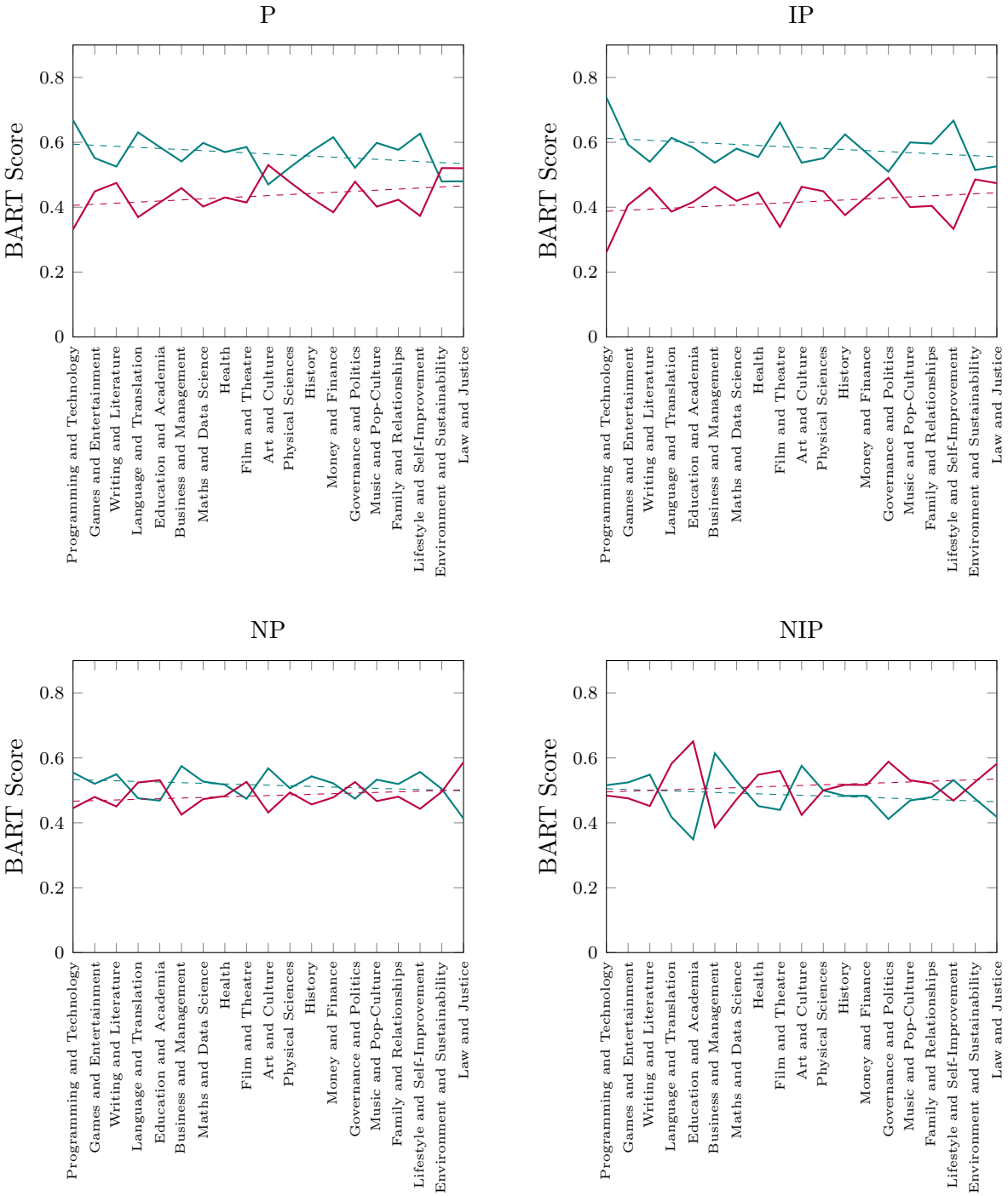


Figure 10: BART scores across topics (`'agrees'` in teal and `'disagrees'` in purple).

And finally, we present an interesting observation related to RoBERTA scores. When plotted across different topics, RoBERTA `'empathy'` show a relatively clear increase the less represented the topic in post-training datasets. Recall that, based on Figure 7, RoBERTA `'empathy'` score has a relatively strong correlation with human agreement evaluations. This means that what we essentially observe is that the less frequent the topic in post-training datasets, the more analytically sycophantic the model. This is despite the fact that we observed decreased `'admiration'` scores in Figure 11.

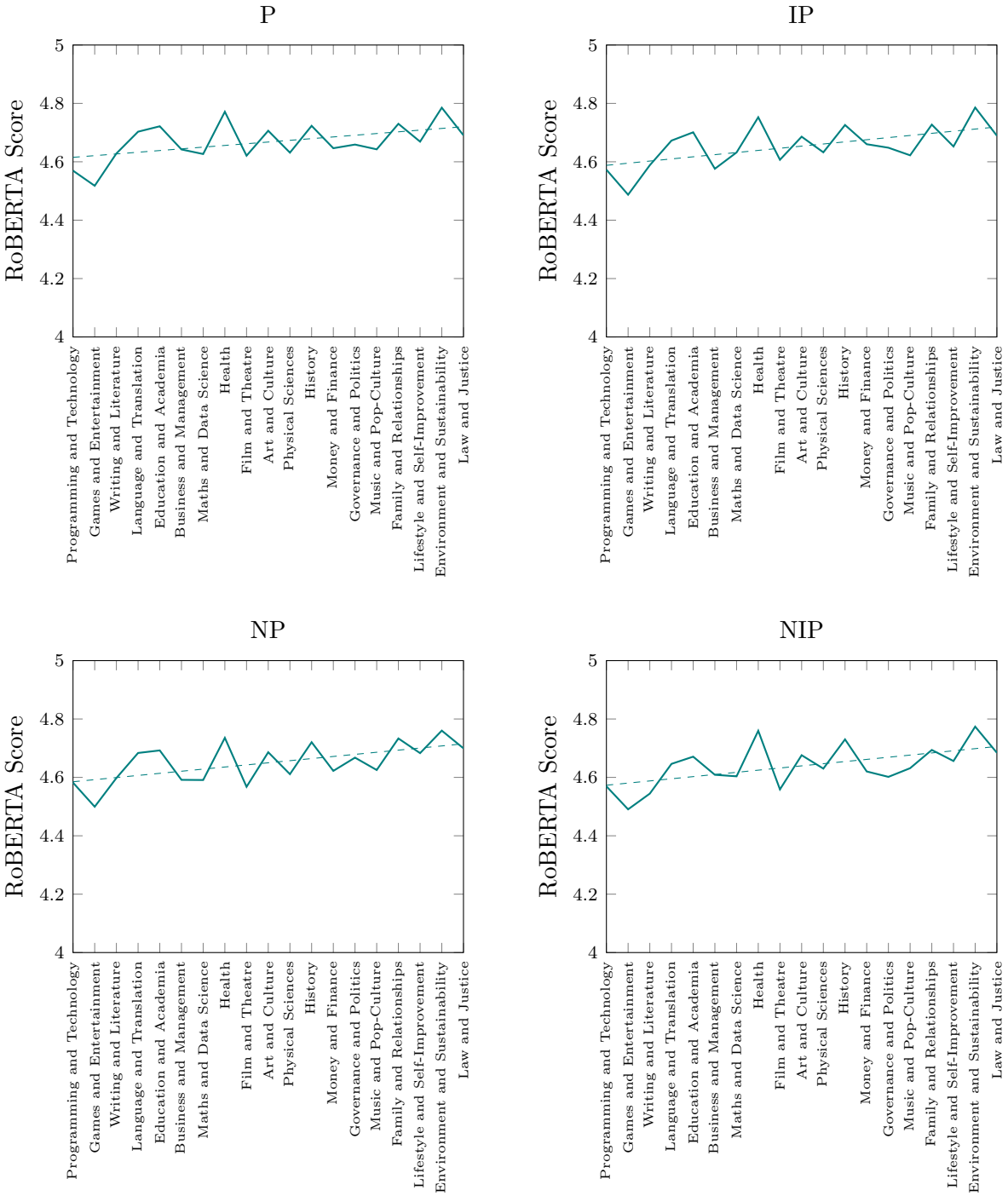


Figure 11: RoBERTA scores across topics (`'empathy'`).

**IV Some Limitations.** This project has encountered multiple limitations. First, even though non-negligible efforts were made to select an appropriate classifier to measure both analytical and empathic sycophancy in the generated data, the possible metrics were selected based on (an average of) 5.5 human opinions, which is by far not enough to understand to what extent the presented values are a reliable predictor of the two types of sycophancy investigated. The better news is that this experiment would be relatively inexpensive, with only a few tens of dollars needed to extract more comprehensive and significant values. The author pledges to

investigate further once they become famous and better funded.

Second, even though we have observed *some* patterns in the collected data, none of the above analyses account for the fact that the differences in how data is represented in post-training datasets are not proportional. More specifically, we observe a pretty much exponential decrease in data—but all of the above-described observations are rather linear. A more thorough analysis, adjusted for the proportions in the observed data, would be necessary for more relevant conclusions. Likewise, the topic counts were skewed by programming and software-engineering data, which gives less flexibility to examine the frequency of potentially opinion-inducing data. Sampling across multiple datasets could help overcome this hurdle.

All observed differences are very small and their significance needs further attention.

# Appendix

| Final Tag | Count | Merged Tags (Count) |
|---|---|---|
| programming and technology | 241 | web development (1), language processing (1), science and technology (1), artificial intelligence ai (1), software development (2), cryptography (3), electronics (4), machine learning (4), computer science (5), artificial intelligence (16), programming (60), technology (143) |
| writing and literature | 112 | literary criticism (1), fantasy magic (1), fictional crossover (1), literary theory (1), fantasy literature (1), fiction writing (1), fan fiction (1), storytelling theory (1), writing (1), mystery literature (1), autobiography (1), creative writing (1), childrens story (1), childrens literature (1), scifi fantasy (1), fanfiction (2), science fiction (13), fiction (18), fantasy (20), literature (44) |
| art and culture | 111 | folklore (1), cultural studies (1), ai art (1), art technology (1), ai art generation (1), art and craft (1), tradition (1), animemanga (2), mythology (2), art and photography (2), art and culture (2), language arts (2), art and design (2), art creativity (2), photography (3), animation (3), anime (4), art and technology (4), culture (8), art (68) |
| games and entertainment | 75 | game development (1), entertainment industry (1), gaming development (1), satire (1), video games (2), games (2), comedy (3), humor (3), fantasy gaming (4), entertainment (28), gaming (29) |
| education and academia | 53 | research (1), academic integrity (1), academic research (2), science (11), education (38) |
| health | 41 | healthcare (1), healthcare technology (1), clinical trials (1), mental health (2), medicine (2), health (34) |
| business and management | 33 | customer service (1), commerce (1), it management (1), business compliance (1), business operations (1), manufacturing (1), ecommerce (1), business communication (2), business management (2), marketing (7), business (15) |
| maths and data science | 28 | statistics (1), geometry (1), data generation (1), puzzles (2), data science (4), mathematics (19) |
| family and relationships | 27 | caregiving (1), sexuality (1), parenting (1), motherhood (1), dating (1), family dynamics (1), family and relationships (1), weddings (1), interpersonal relationships (1), human connection (1), relationships (2), parenthood (2), family (2), romance (3), friendship (7) |
| physical sciences | 27 | space exploration (1), robotics (1), earth science (1), geology (2), climate science (2), environmental science (2), weather (3), astronomy (4), physics (11) |
| money and finance | 21 | cryptocurrency (1), real estate (1), economics (4), finance (15) |
| history | 19 | postcolonialism (1), history (18) |
| film and theatre | 18 | horror (1), film history (1), filmtelevision (1), film studies (1), film analysis (1), theater (1), television (2), drama (2), film (8) |
| language and translation | 18 | translation (1), language learning (2), language translation (4), linguistics (5), language (6) |
| music and pop-culture | 14 | pop culture (1), audio production (1), celebrity comparison (1), music therapy (1), music (10) |
| lifestyle and self-improvement | 14 | fitness (1), leisure (1), luxury (1), personal development (1), home decor (1), recreation (1), beauty industry (1), lifestyle (1), creativity (2), selfimprovement (2), fashion (2) |
| environment and sustainability | 13 | botany (1), nature (1), sustainability (1), business and environment (1), wildlife (1), energy (2), environment (6) |
| politics and governance | 13 | governance (1), international relations (1), global issues (1), politics (10) |
| law and justice | 12 | criminal justice (1), labor rights (1), crime (2), law (8) |
| media and communications | 12 | social media policies (1), telecommunications (1), media production (1), business coaching (1), social media monetization (1), marketing technology (1), media (1), social media (2), communication (3) |
| engineering and construction | 11 | architecture and engineering (1), construction industry (1), architecture (2), construction (3), engineering (4) |
| society and social sciences | 10 | demography (1), futurology (1), indigenous issues (1), society (1), demographics (1), sociology (1), labor movement (1), accessibility (1), dystopia (2) |
| conflict and security | 9 | military (1), war (1), military technology (1), conflict (1), cybersecurity (5) |
| philosophy and ethics | 9 | artificial intelligence ethics (2), philosophy (3), ethics (4) |
| other | 9 | pet products (1), greetings (1), privacy (1), paranormal (1), mystery (1), teen drama (1), transformation (1), roleplaying (1), scheduling (1) |
| sports | 9 | sports politics (1), aquatics (1), cycling (1), sports (6) |
| life sciences | 9 | bioinformatics (1), biotechnology (1), biology (3), chemistry (4) |
| travel and tourism | 8 | travel safety (1), tourism and sustainability (1), travel and food (1), travel (5) |
| workplace and career | 8 | workplace diversity (1), workplace culture (1), workplace safety and compliance (1), corporate event (1), workplace communication (1), career development (3) |
| spirituality and religion | 7 | spirituality (1), religion (6) |
| cognitive sciences | 5 | science and consciousness (1), psychology (4) |
| cooking and baking | 4 | food cooking (1), food safety (1), cooking (2) |

Table 1: grouped tags and their associated original tags as generated by `GPT-4o`

```
'model':model_id,
'messages': [{'role':'user', 'content':prompt}],
'temperature':0.7,
'max_tokens':500
```

Figure 12: Model configuration

# REFERENCES.

Ranaldi, L. and Pucci, G. (2024). When Large Language Models contradict humans? Large Language Models' Sycophantic Behaviour.

Sharma, M., Tong, M., TomaszKorbak, Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., and Perez, E. (2025). Towards Understanding Sycophancy in Language Models.

Tessler, M. H., Bakker, M. A., Jarrett, D., Sheahan, H., Chadwick, M. J., Koster, R., Evans, G., Campbell-Gillingham, L., Collins, T., Parkes, D. C., Botvinick, M., and Summerfield, C. (2024). AI Can Help Humans Find Common Ground in Democratic Deliberation. *Science*, 386(6719):eadq2852.