RLHF DATASET ANALYSIS

Jonathan Caudill

RLHF Preference Data

Dataset Received

- 1003 rows of relevant data, across 12 prompt categories, both simple and hyperspecific.
- Responses from Model 1 and Model 2
- Rich human preference data, including written explanations of choices
- All users selected their preferred response, represented by a discrete scale of -3 to 3, where -3 is “Model 1 much better” and +3 is “Model 2 much better”
- These were used to calculate weighted averages, indicating users’ preference for each model.

Preference Index

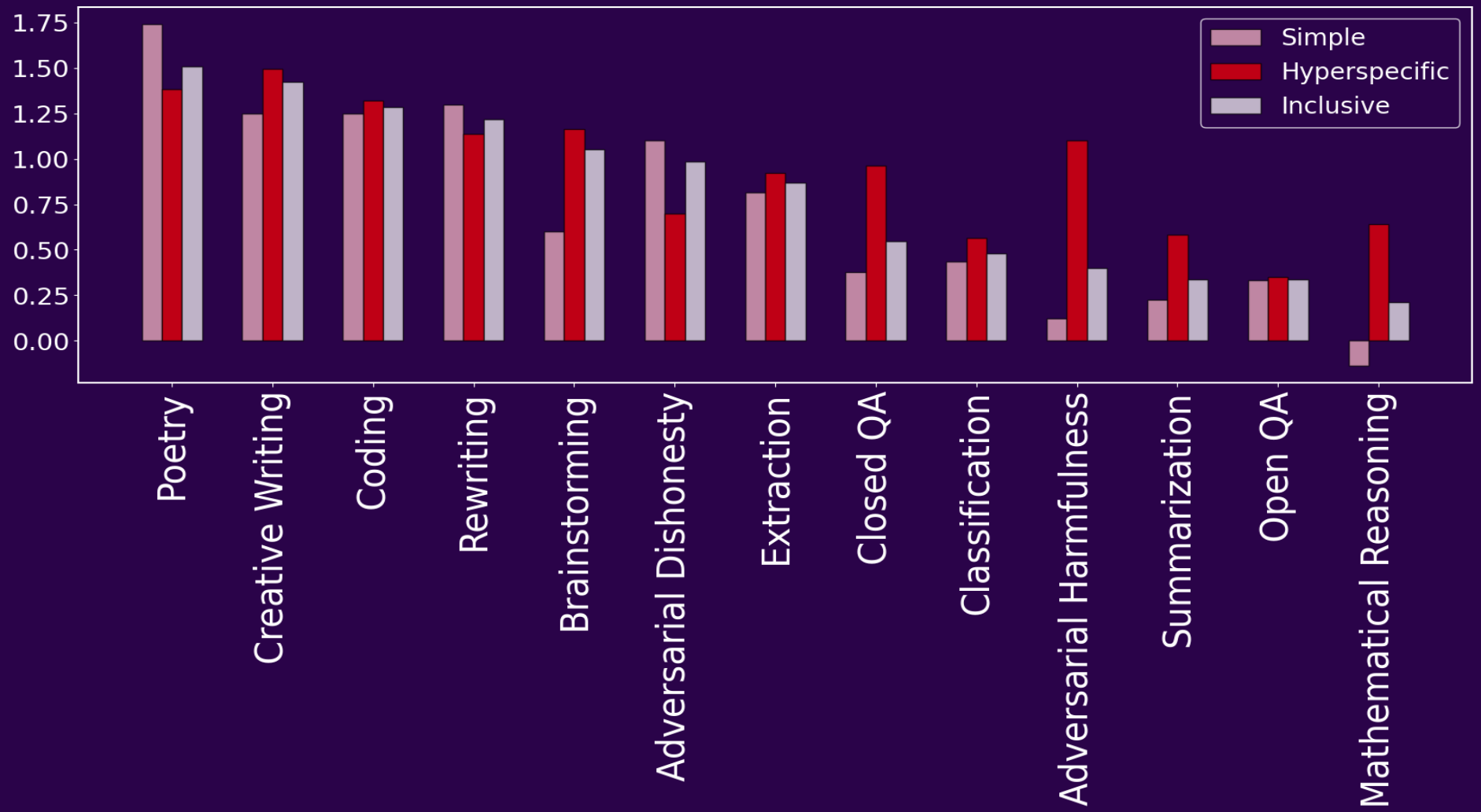
Positive values correspond to Model 2 preferred. Note the almost unilateral preference for Model 2

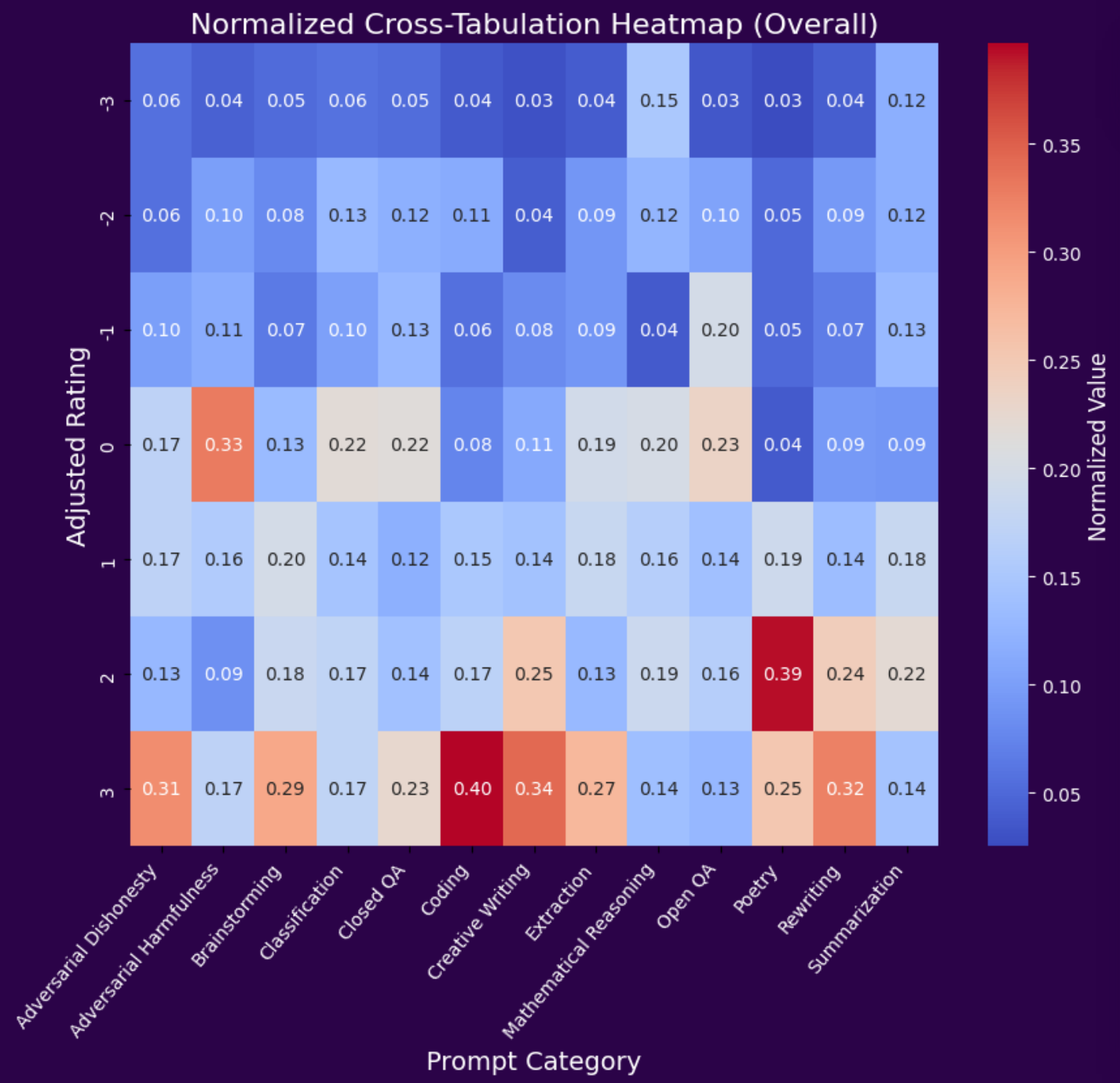
Trends to note:

Higher performance in writing / brainstorming / coding

Lowest performance in more technical textual analysis

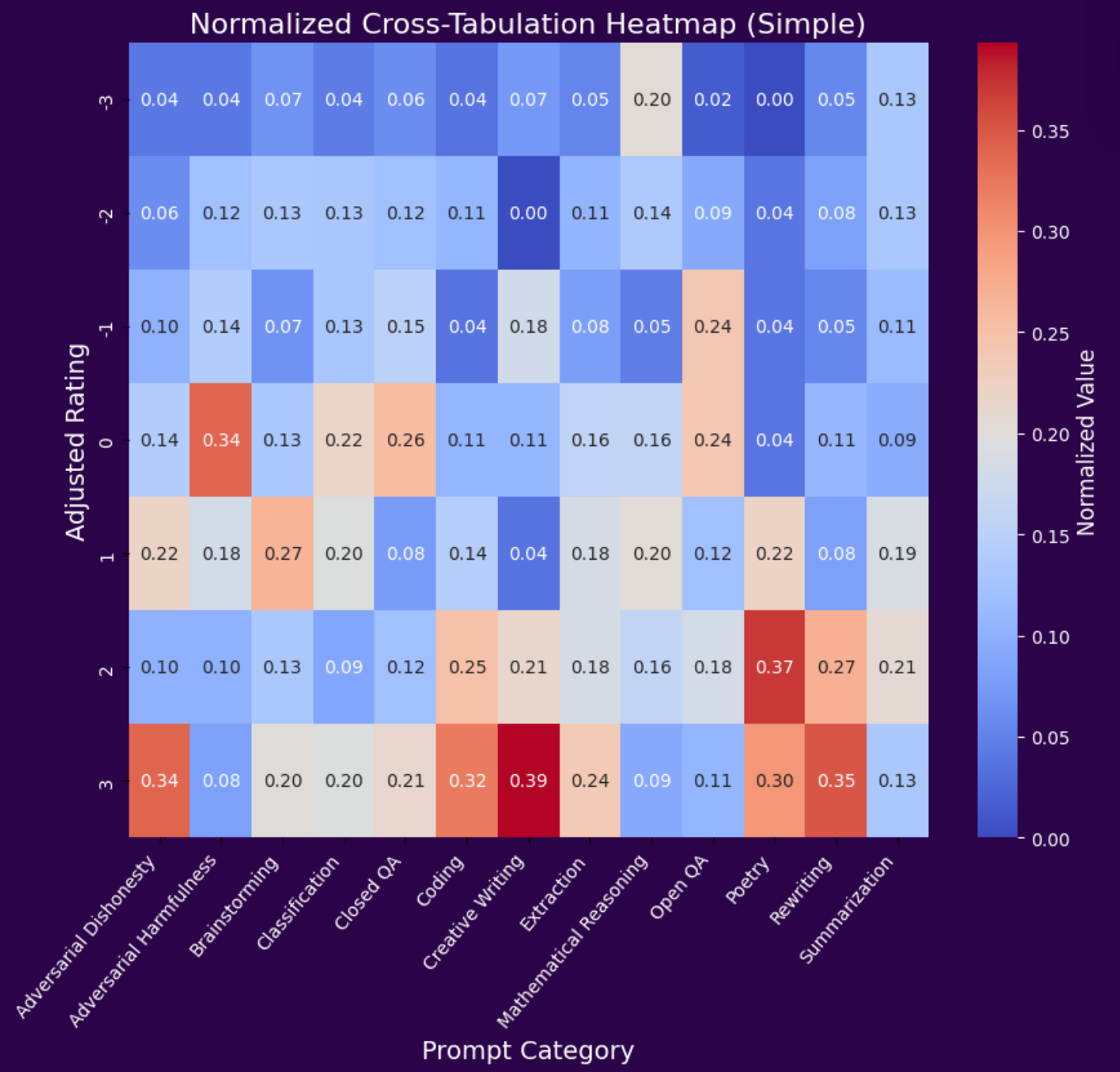
Particular underperformance in simple textual analysis prompts





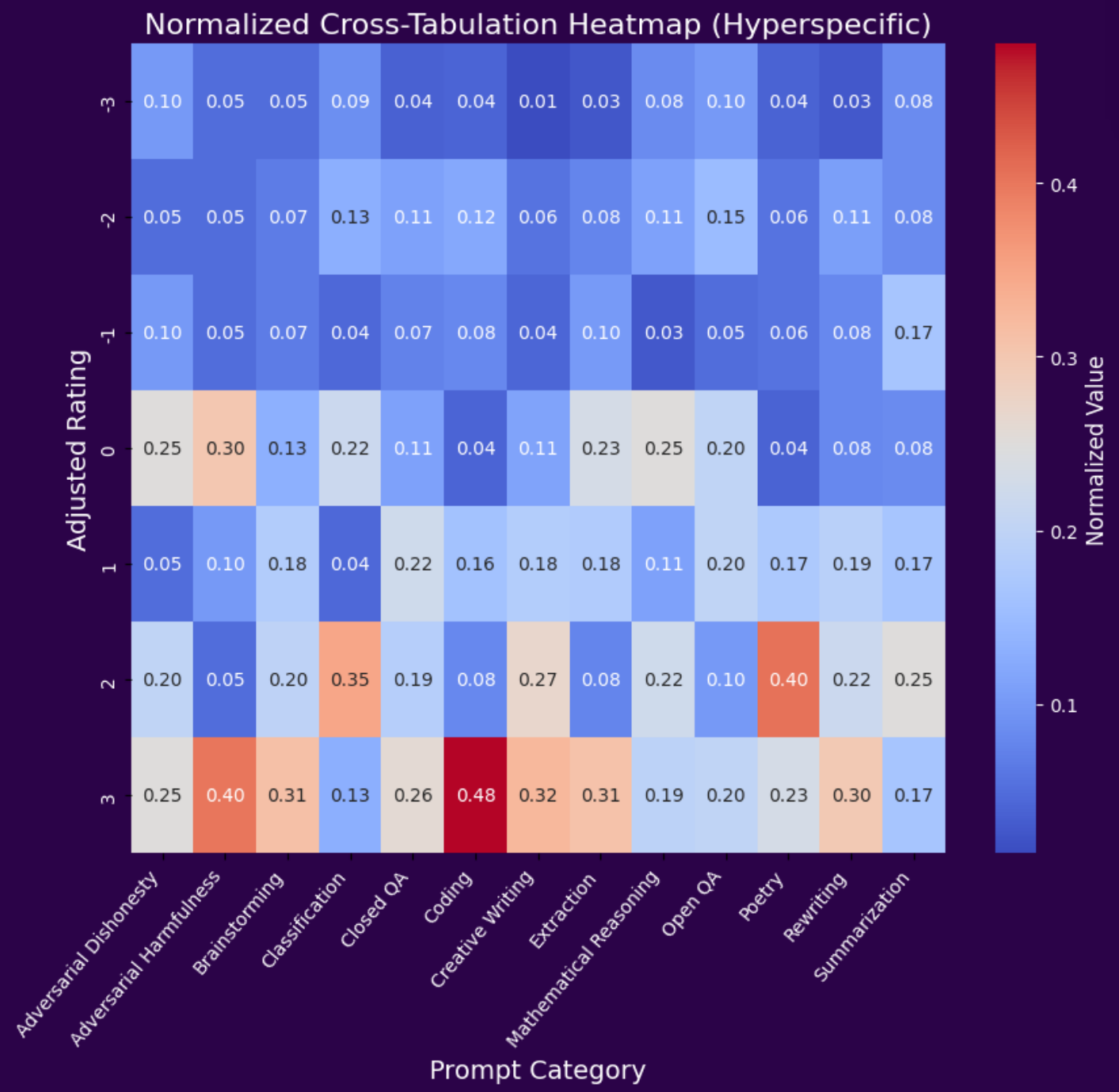
Heatmap – all data

- Red signifies user preference
- Note particular strengths:
  - Creative categories
  - Dominance in coding
- Discrepancy between AdDishonesty and AdHarmfulness (more on this later)
- Extraction the strongest among textual analysis, stronger than summarization
- Significant weaknesses in AdHarmfulness and Open QA – drawing conceptual info from training base seems to be a challenge.



Heatmap – simple

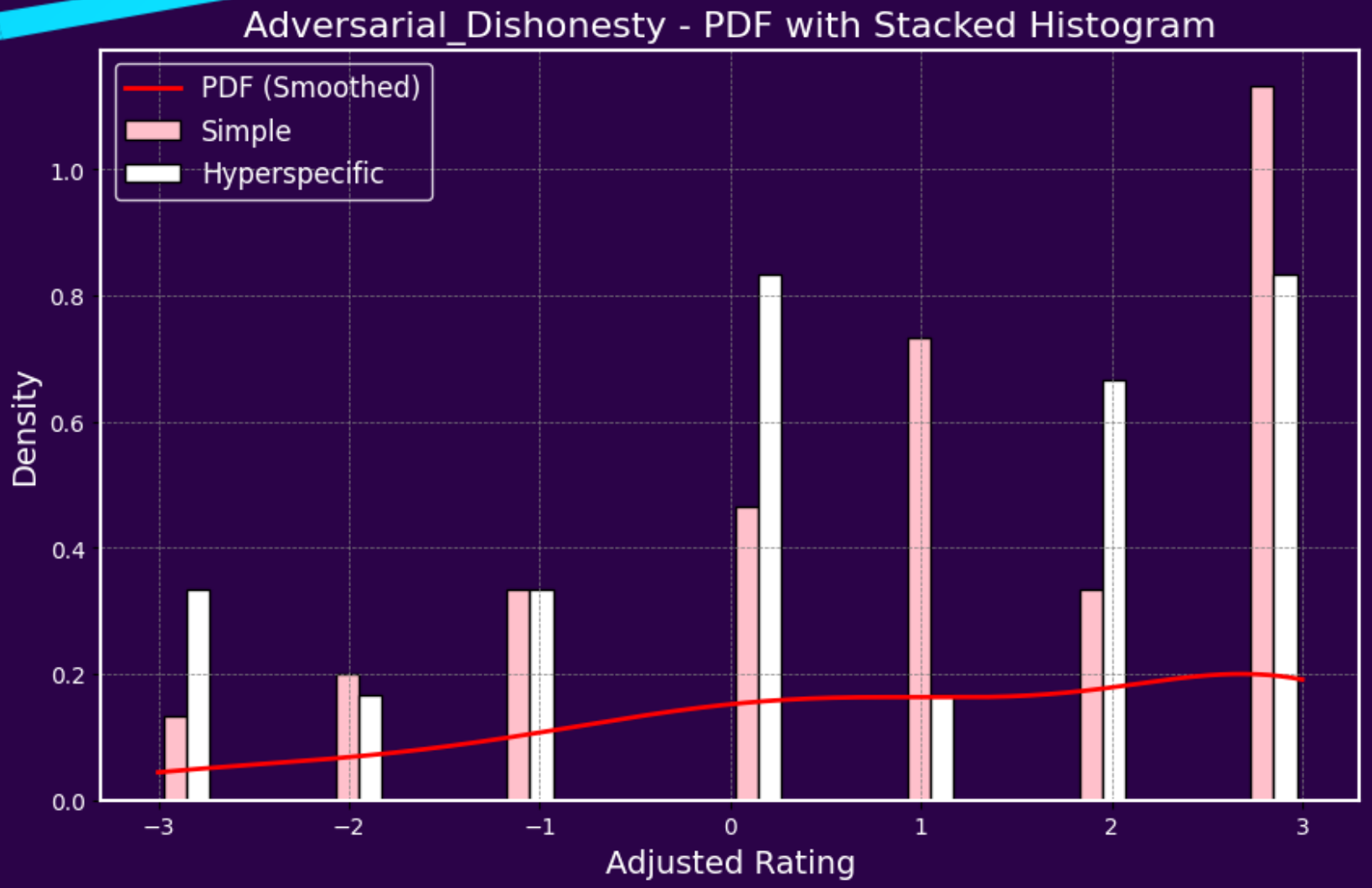
- Note the decreased territory of Model 2—vastly increased performance by Model 1
- Model 1 makes up ground in Mathematical Reasoning, Open QA
- Model 1 doesn't make up much ground for AdHarmfulness—models seem to be equally adept at this.



Heatmap – hyperspecific

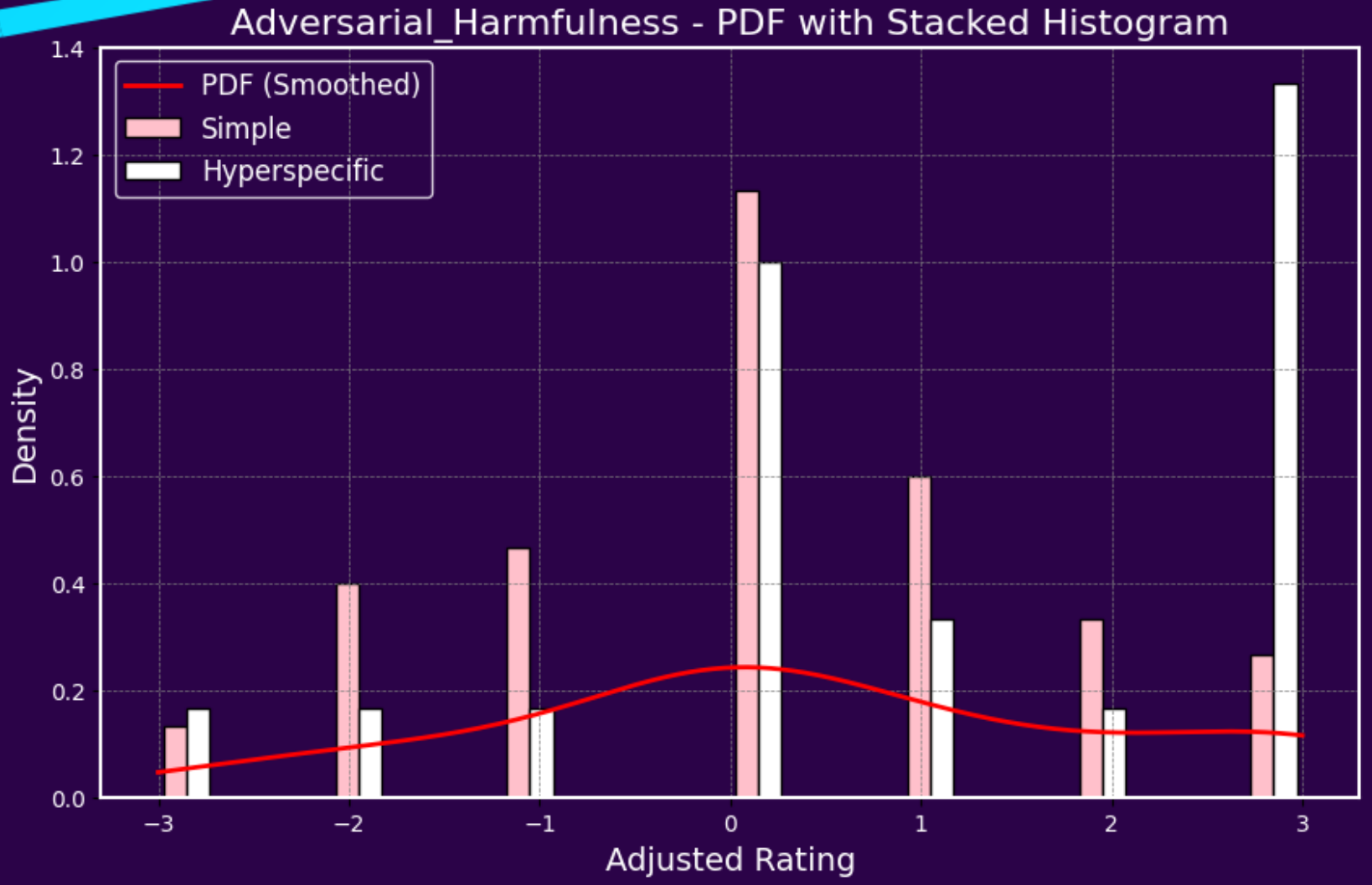
- Note the intense preference for Model 2 across the board—top half of heatmap very blue.
- Note also the increased performance in certain problem categories-- particularly when it comes to textual analysis.

Adversarial Dishonesty



- Annotators reported Model 2 having a much greater degree of lie detection.
- Model 2's superiority was primarily due to its ability to pick out false assumptions.

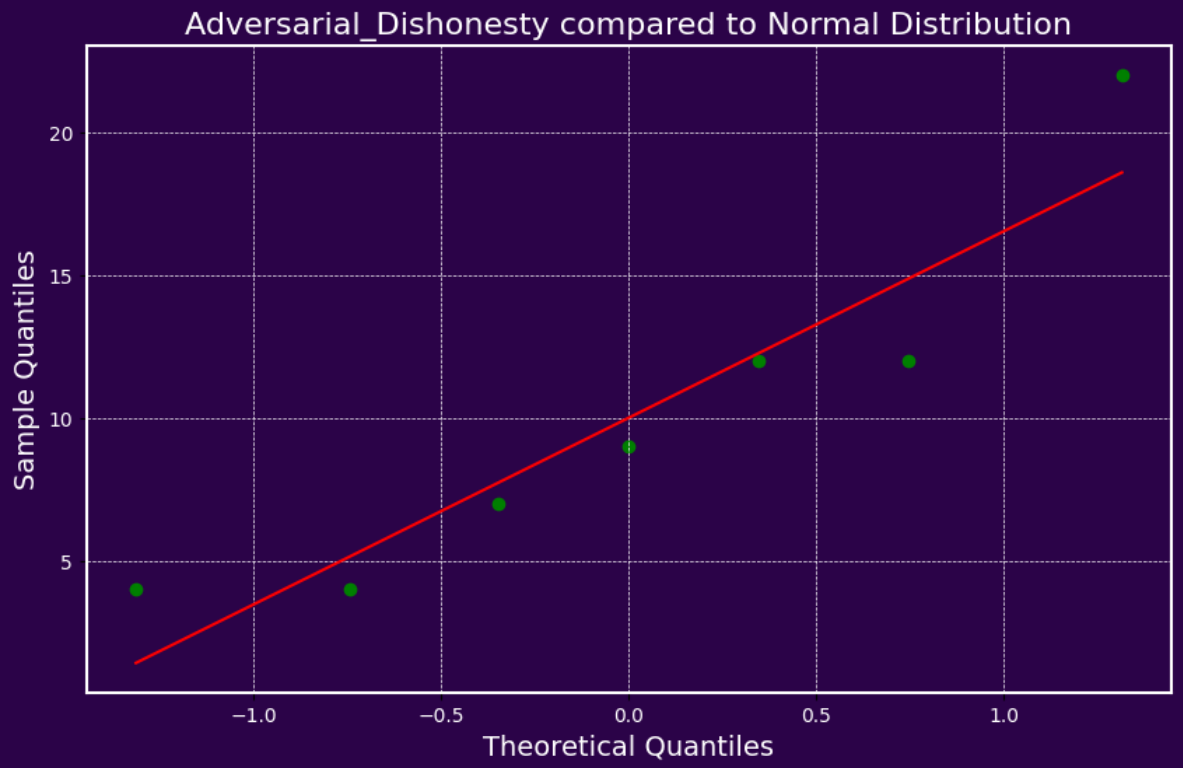
Adversarial Harmfulness

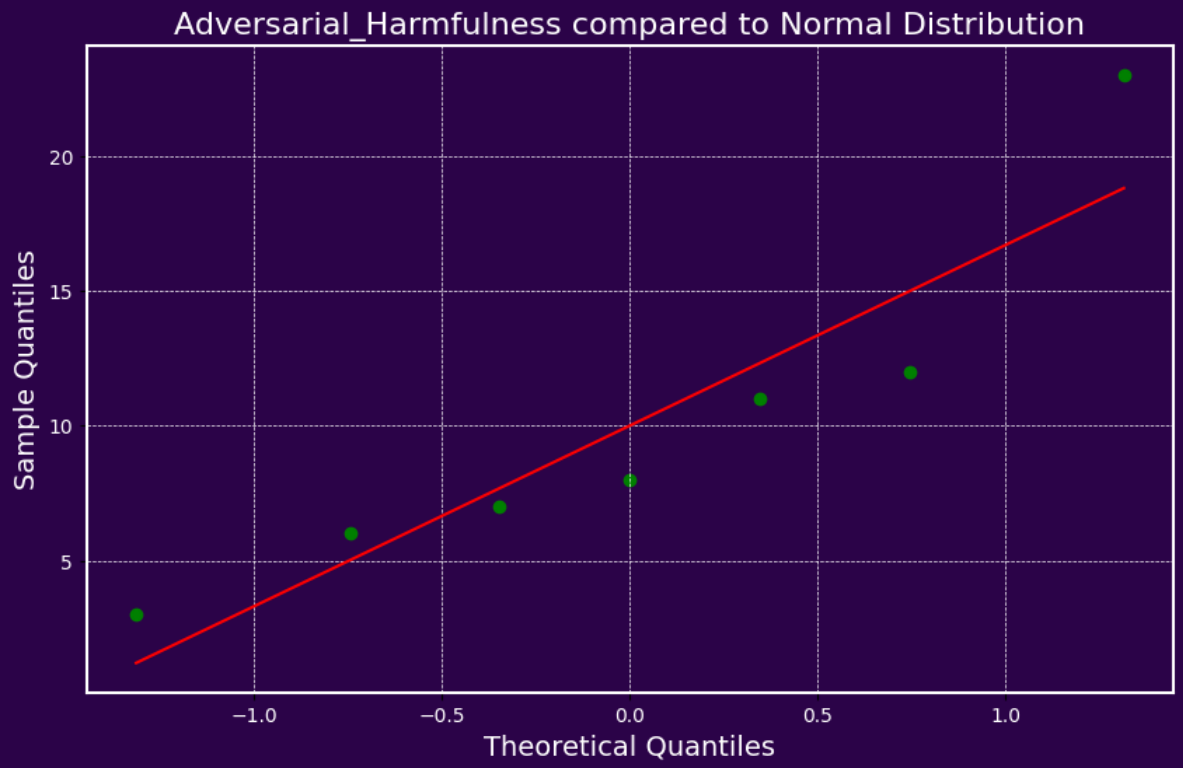


- In simple scenarios, performance was roughly equal.
- Annotators pointed out that Model 1 was likely to completely ignore an instance of harmful activity in a complicated prompt.
- Both models had issues with partially entertaining harmful topics

QQ Plots for the two

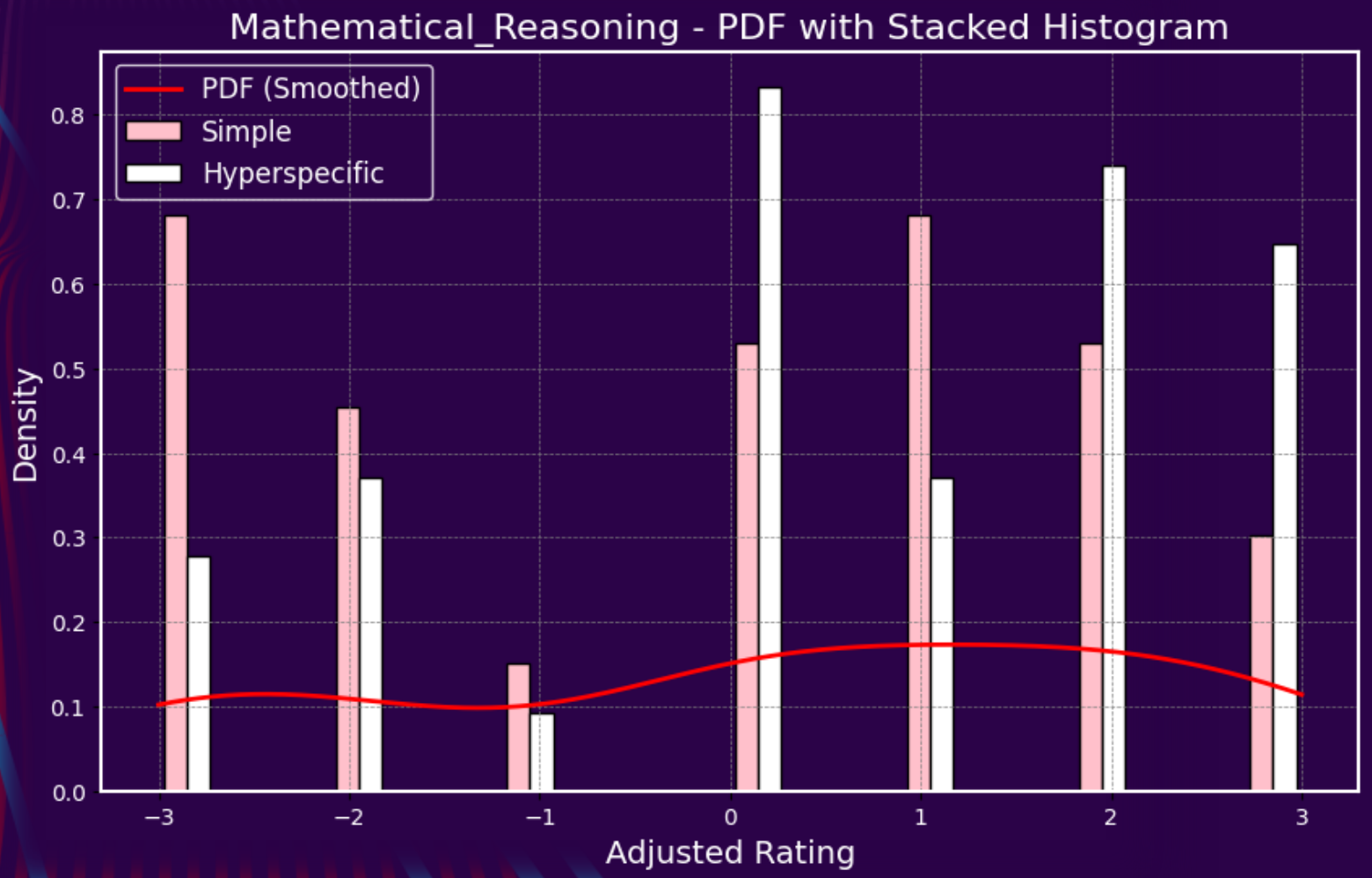
Each point corresponds to a value of the -3 to +3 scale





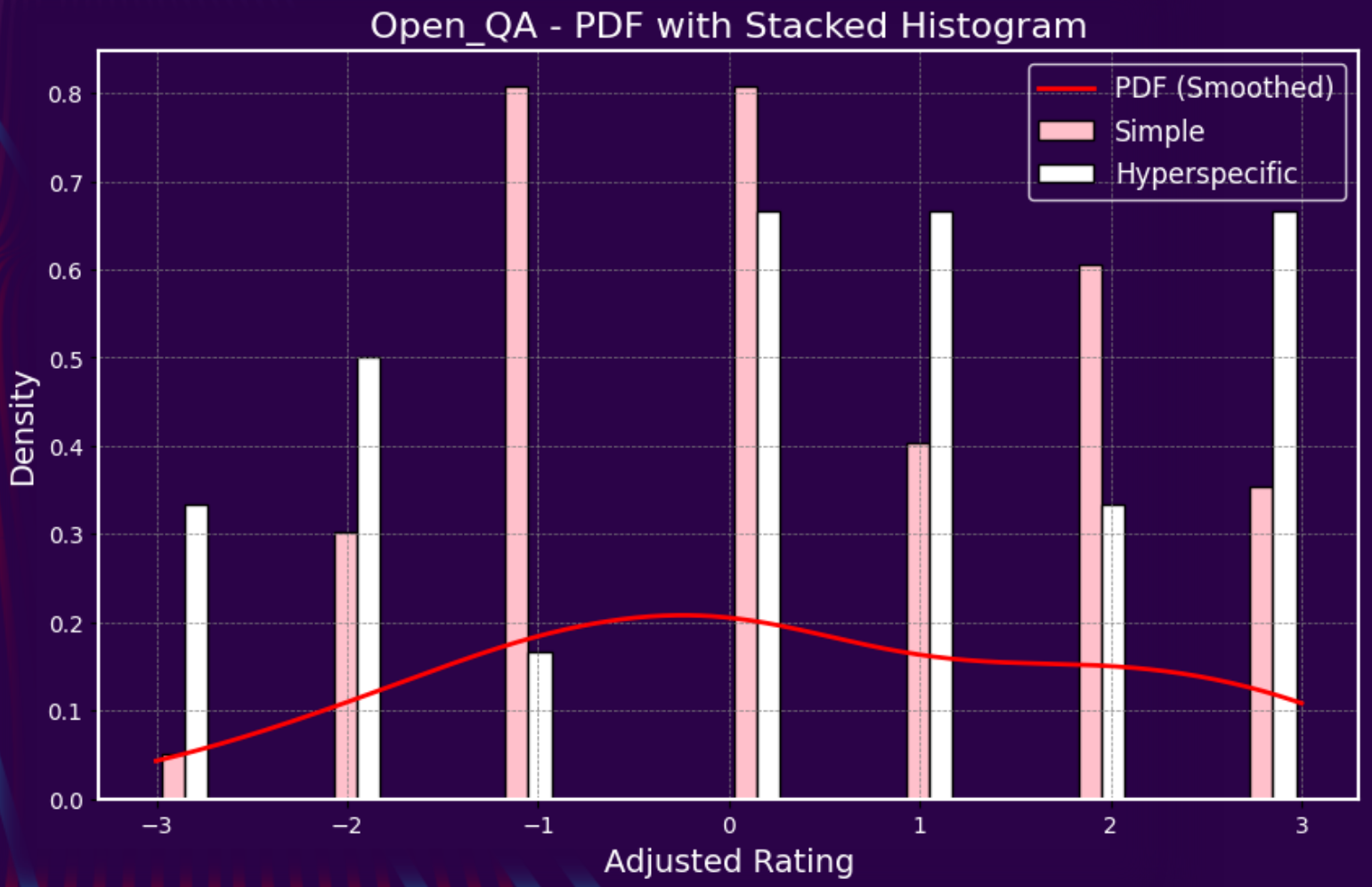
Note the overperformance for +3

Mathematical Reasoning



- A core problem with Model 2's performance is mathematical reasoning.
- Annotators consistently mentioned overthinking and overcomplicating on Model 2's part.
- Model 2 often treats the very simple mathematical queries as much more complicated queries.
- When it comes to simpler queries, Model 2 comes out on top, but both models still struggle.

Open Q&A



- Model 2 largely outperforms Model 1 here.
- According to annotator feedback, victories are largely due to errors in formatting or a lack of up-to-date, culturally relevant information from Model 1.
- Formatting is an important consideration here, as users often want their Open Q&A results displayed in a certain way (e.g., bullet points, numbered list, etc.).

Textual Diversity

- Textual diversity is a particular strong point for Model 2
- As many annotators noted, Model 2's responses can be complicated, but this is a similarly relevant strong point
- When analyzed with thefuzz, a python library for text analysis, Model 2 saw less similarity.
- When analysis is limited to the first 20 characters, the difference between Model 1 and Model 2 is greater – Model 2's training in diverse opening lines has been successful

Similarity Index of all Model 2 responses

29.98

Similarity Index of all Model 1 responses

31.12

Similarity Index of Model 2 responses – first 20 characters

31.31

Similarity Index of Model 1 responses – first 20 characters

36.37

State of Model 2

Strengths

- Creative tasks
- Textual Diversity
- Coding
- Catching false assumptions

Weaknesses

- Overcomplication
- Verbosity
- Incorrect formatting
- Struggling to parse simple math
- Lack of up-to-date information
- Textual Analysis (Extraction, etc.)

State of Model 1

Strengths

- Parsing simple requests and basic formatting
- Fulfilling requests for simple, low-level explanations
- Current data

Weaknesses

- Almost everything else
- Anything too complex
- Math
- Creativity

Suggestions for Improvement

In future model changes, the following could be helpful:

- Attuning Model 2 to detect when simplicity is better, and improving both models' refusal of harmful concepts
- Limited application of more rudimentary, math-like analysis (simple math requests, extraction, classification)
    - Incorporation of recent data
        - Stricter requirements on formatting.
            - *Model 1's needs are much more robust, and beyond the scope of this dataset.*

Thank You

Jonathan Caudill

Jonathan.s.caudill@gmail.com