Evaluating Large Language Models

Deniz Bölöni-Turgut and Leo Lu
Advisors: Claire Cardie, Tanya Goyal, and Wenting Zhao

LLMs are increasingly being used to evaluate long-form text. Are these LLMs reliable, and what evaluation biases do they exhibit?

- Prompted four models to generate different length essay responses to 100 randomly sampled "open-ended" questions
- Graded these essays alongside their verbose and concise counterparts with 4 models (gpt-4o, gpt-4o-mini, llama3-70b, and llama3-8b).

+58.8% avg token length (GPT-4o)

Please respond to the following questions in essay format.

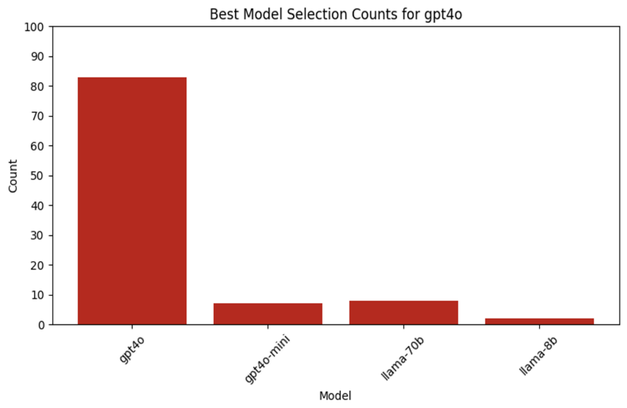
→ Make the previous essay response significantly more ***verbose***.

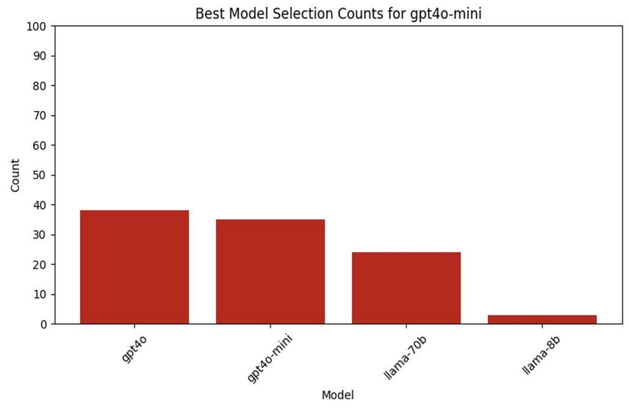
→ Make the previous essay response significantly more ***concise***.

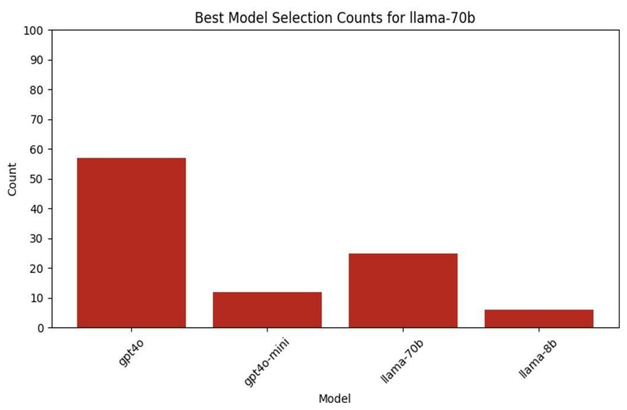
-49.7% avg token length (GPT-4o)

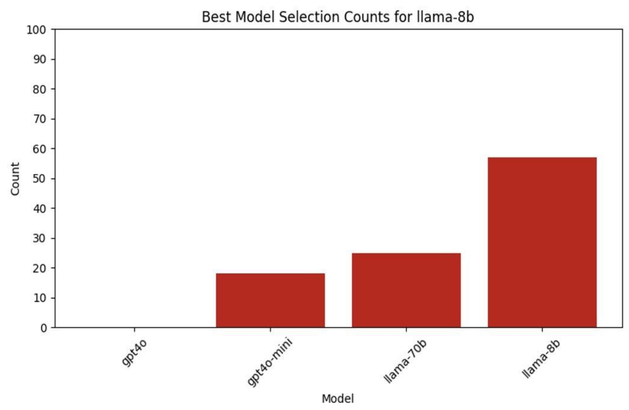
Self-Bias

Evaluation Metric: Direct Comparison of 4 Model Generated Essays





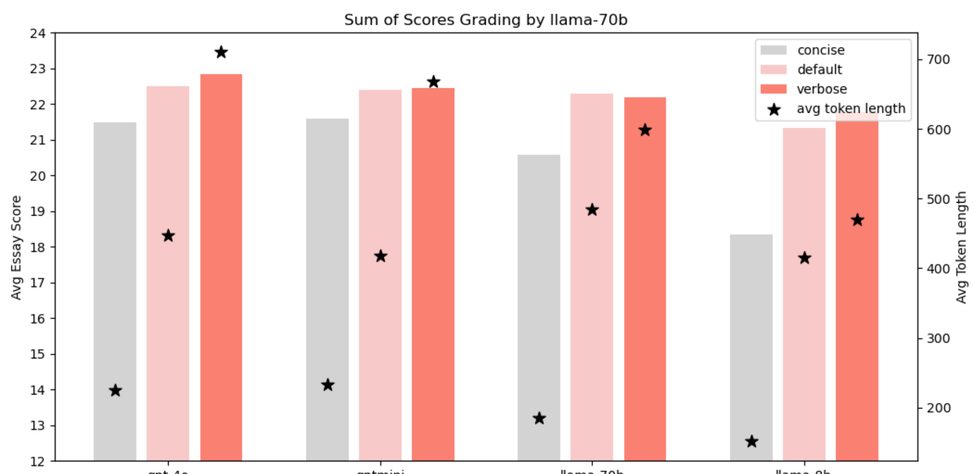




Length Bias

Evaluation Metric: Custom Rubric (0-24 points)

- Thesis and Support
- Organization and Structure
- Critical Thinking and Reasoning
- Awareness of Perspectives
- Use of Sources and Evidence
- Writing Style and Mechanics



Generated Essay Similarities

PROMPT: "how do global factors influence the economy in your country?"

Concise: "Trade allows countries to export goods and services where they have a comparative advantage and import those that are costly to produce domestically."

Default: "International trade allows countries to export goods and services in which they have a comparative advantage and import those that are otherwise expensive or inefficient to produce domestically."

Verbose: "International trade provides countries with opportunities to export goods and services in which they have a comparative advantage, thus allowing them to generate significant revenues and enhance their gross domestic product (GDP)."