

STEER: Assessing Economic Rationality in Large Language Models

Narun Raman¹, Taylor Lundy¹, Samuel Amouyal², Yoav Levine³, Kevin Leyton-Brown¹, Moshe Tennenholtz⁴
1. University of British Columbia, 2. Tel Aviv University, 3. Stanford & AI21 Labs, 4. Technion & AI21 Labs



SCAN ME



LLMs as Economic Agents?

User Prompt

Last year, you decided to invest in a new technology expected to give \$300,000 in profit. To date, \$60,000 have been invested in the project. To complete the project, an additional \$240,000 are required. However, a new competitor's technology has been introduced, reducing the projected profit to \$200,000. Would you:
A. Cut your losses and stop the project.
B. Continue to invest the remaining \$240 in the older technology.

GPT-4 Turbo

The decision should be based on the principle of sunk costs and marginal analysis. The \$60,000 already invested is a sunk cost and should not factor into the decision-making process. What matters is whether the additional \$240,000 investment will yield a profit given the new projected profit of \$200,000. Since the additional investment is less than the revised profit, it would make sense to continue with the project. Correct Answer: B. Continue to invest the remaining \$240,000 in the older technology.

Comprehensive Benchmarking



Economics has a long history in defining rational decision-making



Key idea: a rational agent should maximize expected utility



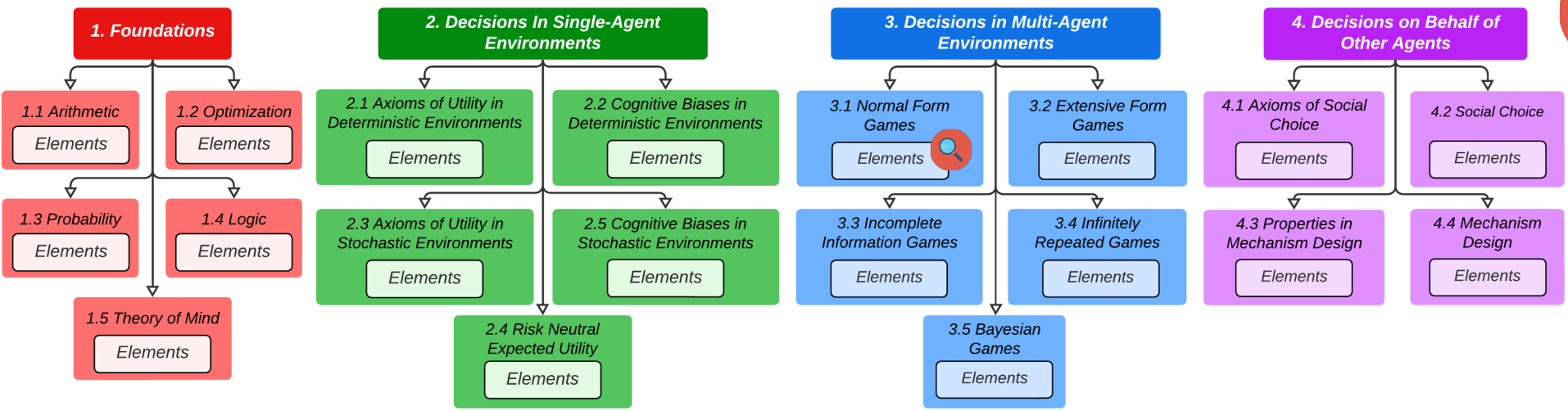
Literature categorizes important differences between economic settings



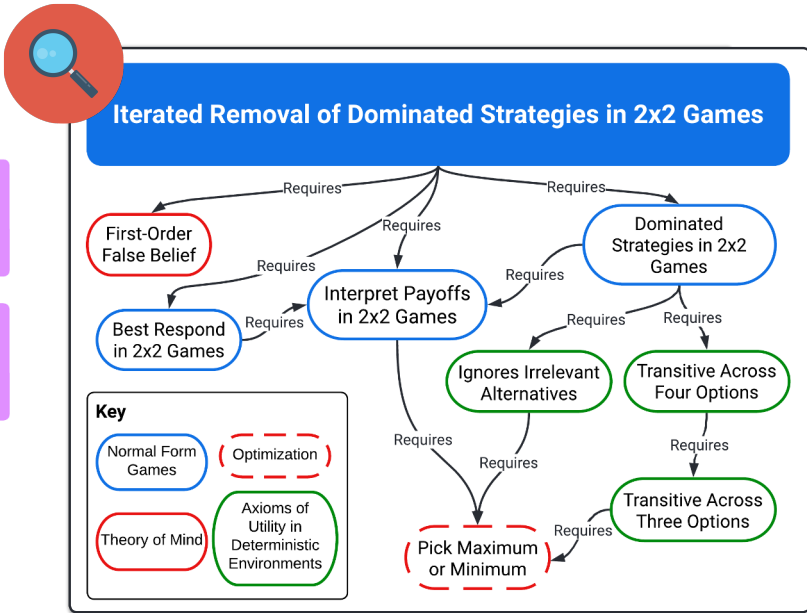
In some settings, the rational option is not clear. Even worse, in others, impossibility results rule out all options

We restrict to tests where rational/focal answer is well defined

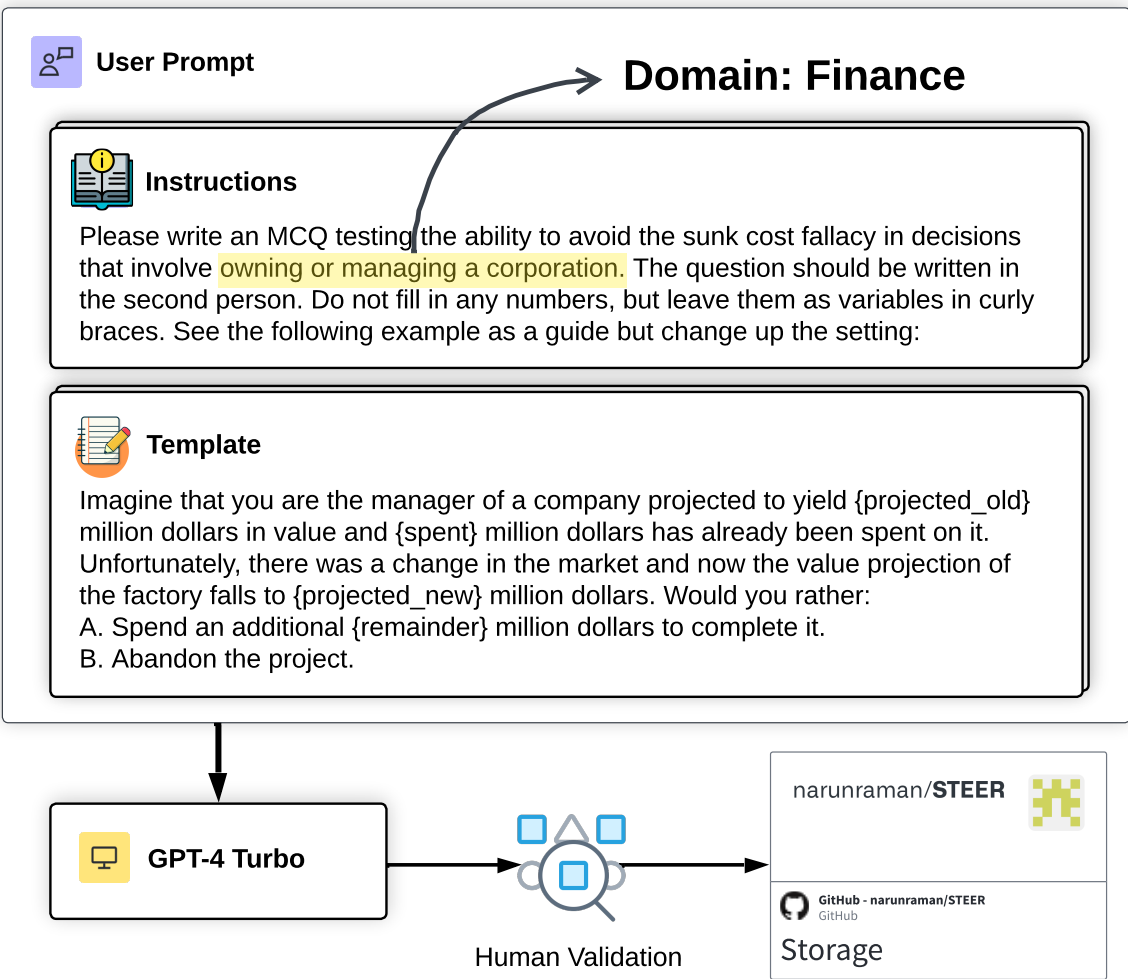
STEER: Systematic and Tunable Evaluation of Economic Rationality



All of this is viewable on our website: <https://steer-benchmark.streamlit.app/>

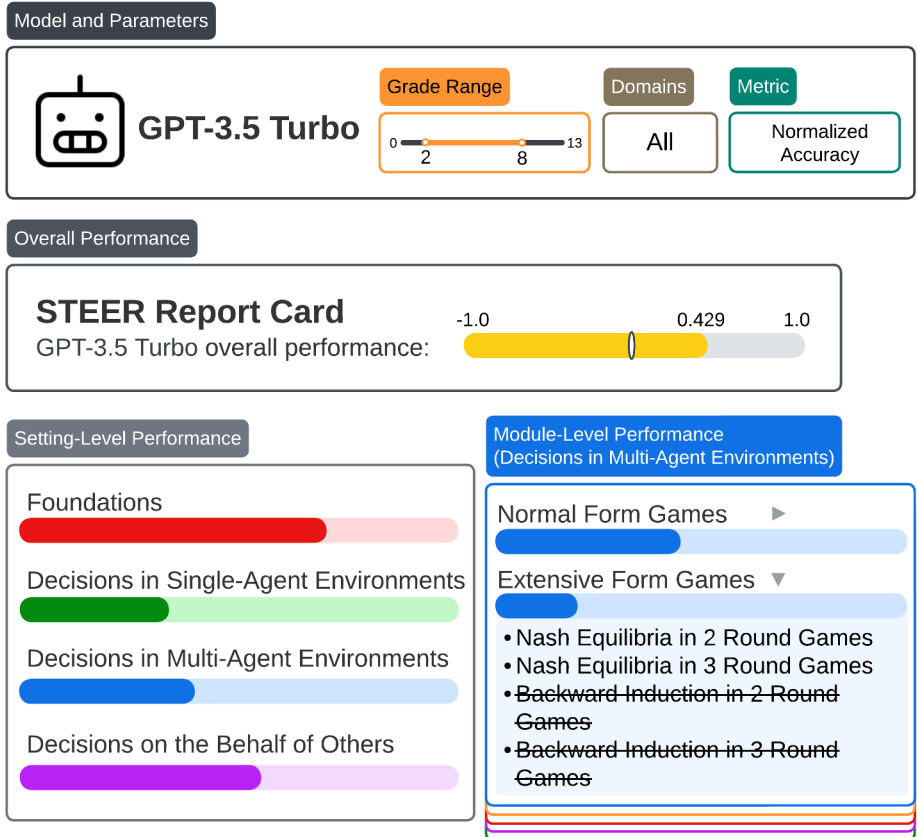


STEER Generation and Validation

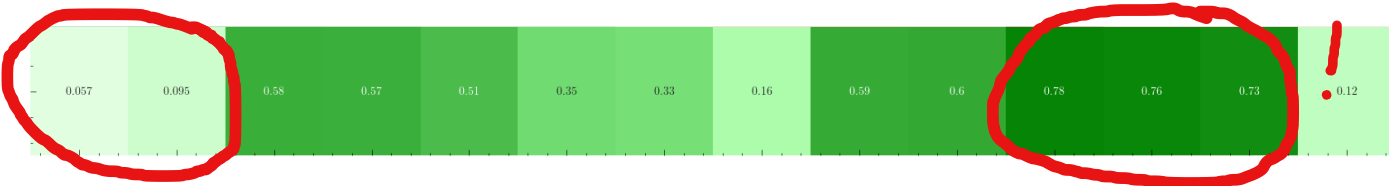


STEER currently contains 926,000 test questions (1000 tests per (element, grade level, domain) tuple) with more being generated every day!

STEER Report Cards



Example: (Avoidance of the Endowment Effect)
GPT models have inconsistent preferences when their preferences are inferred from the context rather than their own decision-making.



Across 4 adaptations:

- Multiple Choice answer only
- Self-Explanation + MC
- Few Shot Prompting + MC
- Few Shot Prompting + Self-Explanation + MC

On 4 metrics:

- Exact-Match Accuracy
- Normalized Accuracy
- Dependency Robustness
- Domain Robustness

We evaluated 20 models:

- GPT-4 Turbo (1.76T?)
- GPT-3.5 Turbo (175B)
- Llama-2 (70B)
- Llama-2 Chat (70B)
- Llama (65B)
- Mixtral (56B)
- Falcon (40B)
- Falcon Instruct (40B)
- Alpaca Native (13B)
- Llama (13B)
- Llama-2 (13B)
- Llama-2 Chat (13B)
- Falcon (7B)
- Falcon Instruct (7B)
- Llama (7B)
- Llama-2 (7B)
- Llama-2 Chat (7B)
- Mistral (7B)
- Databricks Base (7B)
- Databricks Instruct (7B)