Rationality Report Cards

Kevin Leyton-Brown
Professor of Computer Science; Distinguished University Scholar
University of British Columbia
Canada CIFAR AI Chair, Alberta Machine Intelligence Institute
Advisor, AI21 Labs

THE UNIVERSITY OF BRITISH COLUMBIA

amii

AI21labs

Thanks to my coauthors!

- Narun Raman
- Taylor Lundy
- Yoav Levine
- Samuel Amouyal
- Moshe Tennenholtz

ArXiv





LLMs as Agents

- Recently, desire to leverage LLMs to create **decision-making engines**
  - [Cai, Tianle and Wang, Xuezhi and Ma, Tengyu and Chen, Xinyun and Zhou, Denny, 2023; Horton, 2023; Wang, Zhenhailong and Mao, Shaoguang and Wu, Wenshan and Ge, Tao and Wei, Furu and Ji, Heng 2023a,b; Zhuge, Mingchen and Liu, Haozhe and Faccio, Francesco and Ashley, Dylan R and Csordas, Robert and Gopalakrishnan, Anand and Hamdi, Abdullah and Hammoud, Hasan Abed Al Kader and Herrmann, Vincent and Irie, Kazuki, 2023; Shen et al., 2023]

- Initial success:
  - planning e.g., Reworkd AI
  - solving complex problems e.g., HuggingGPT
  - leveraging tools e.g., ToolFormer
  - playing games e.g., Voyager

Can We Trust Them?

Narun:

Oh no! You are at the casino and on a losing streak – the past two hands you have lost $5,030 and $1,012. How much have you lost in total?

GPT-4 Turbo:

To find out how much you've lost in total, you simply add up the amounts lost in each hand: $5,030 (from the first hand) + $1,012 (from the second hand) = $6,042. So, you have lost a total of $6,042 at the casino.

Note: this and all examples to follow are representative behavior

Slight Changes In Difficulty Matter

Narun (new conversation):

Oh no! You are at the casino and on a losing streak – the past three hands you have lost $5,030.74, $7683.125 and $1,012.39. How much have you lost in total?

GPT-4 Turbo:

To find out how much you've lost in total, you simply need to add up the losses from each hand: Total loss = $5,030.74 + $7,683.125 + $1,012.39. Total loss = $15,726.255. So, you have lost a total of $15,726.255 at the casino.

No!

$5,030.74 + $7,683.125 + $1,012.39 = $13,726.26

What About Decision Making?

Narun (first conversation continues):

Given your bad luck you decide to switch to the slots. Slot Machine A costs $25 per play and offers an 80.67% chance to win $822. Slot Machine B costs $5 per play and has a 57.58% chance of winning $929. Which slot machine should you play?

GPT-4 Turbo:

[...lots of setup text...] For Slot Machine A, the expected profit is (0.8067 * $822) - $25 = $661.77 - $25 = $636.77. For Slot Machine B, the expected profit is (0.5758 * $929) - $5 = $534.77 - $5 = $529.77.

Therefore, Slot Machine A should be played.

Good!

Domains Matter [Casino → Job Market]

Narun:

Your phone just now buzzes with two job offer notifications. Job P offers a salary of $96,480 per year with no chance of a bonus. Job Q offers a salary of $88,973 per year but with a 95.24% chance of a $7,047 bonus. Which job should you choose to maximize your expected utility?

GPT-4 Turbo:

[...lots of setup text...] The expected utility of Job P can be calculated as the salary of $96480. For Job Q, we need to calculate the expected utility by multiplying the probability of receiving a bonus (95.24%) by the bonus amount ($7047) and adding it to the salary ($88973) which is $95684.56. [...lots more text...] Since the expected utility of Job Q is higher than that of Job P, we should choose Job Q to maximize our expected utility.

Consistently struggled with job market problems!

Can They Reason About Others?

Narun:

Back at the casino, you see an open table and sit down to play. You are against only one other player, and given your hand (Ace, Queen) and their hand (Jack, King), you estimate that if you play an Ace and your opponent plays a Jack you get 8.71 and she gets 14.34; if you play an Ace and your opponent plays a King, she gets 17.37 and you get 12.16. [...text describing the other payoffs...] What's your opponent's best response if you play your Ace?

GPT-4 Turbo:

The best response for the opponent is the action that maximizes their payoff given the action chosen. [...lots more text...] In this case, if you play an Ace, your opponent's best response is to play a Jack because it yields a higher payoff of 17.37 compared to 14.34 [...lots more text...]

No!

Can They Make Social Decisions?

Narun:

Growing disgusted with your continual string of losses, you join a petition supporting a ban on casinos. As part of this petition, you are put in charge of tallying votes to decide on a representative. During the election with 4 candidates Z, D, C, H, the votes are: 45 voters voted for D > Z > H; 82 voted C > D > Z; 72 voters voted for C > Z > H; and 45 voted Z > C > D. Who wins if you used Borda count?

GPT-4 Turbo:

First, assign points to each candidate based on their ranking in each voter's preference. For example, a first place ranking earns 3 points, second place earns 2 points, and third place earns 1 point. [...lots more text...] Therefore, Candidate Z wins in the Borda count

No!

Election ordering D > C > Z > H

Trust Through Benchmarking

- Many well-established NLP benchmarks
  - Key contributors to the rapid progress in text generation
- Language Modelling: next word prediction
  - WikiText-103, One Billion Word Benchmark, LAMBADA
- Question Answering: retrieving the answer to a question
  - SQuAD, HotpotQA, TriviaQA, SWAG
- Text Classification: categorizing text
  - AG News, DBpedia
- Multi-Task: evaluation across tasks
  - HELM, GEM, GLUE, GLGE

Which of the following terms describes the body's ability to maintain its normal state?
A. Anabolism
B. Catabolism
C. Tolerance
D. Homeostasis

HELM

Relatively Few Decision-Making Benchmarks

- Open-world games: can LLMs play games?
  - [Araci, 2019; Sinha and Khandait, 2020; Akata 2023]
- Finance: can LLMs be helpful in simple economic settings?
  - FinQA, FinBERT
- Cognitive Biases: do LLMs make the same mistakes as humans?
  - [Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, Dongyeop Kang 2023; Horton 2023]

> You are hiring for the role of "Dishwasher."
> The typical hourly rate is $12/hour. You have 2 candidates.
> - Person 1: Has 1 year(s) of experience in this role. Requests $17/hour.
> - Person 2: Has 0 year(s) of experience in this role. Requests $13/hour.
> Who would you hire? You have to pick one.
>
> [Horton 2023]

- Each of these benchmarks contains 2–5 tasks

Comprehensive Benchmarking?

- Economics has thought a lot about defining rational decision making
  - Over a century of quantitative literature
    - including cognitive science, operations research, psychology, …
  - Key idea: a rational agent should **maximize expected utility**
- Also categorizes important differences between economic **settings**
  - They differ fundamentally
    - Single-agent settings are different from multi-agent settings
    - Reasoning on behalf of agents is different from reasoning about how to act as an individual
  - In some settings, the theory is **prescriptive**
    - It is always better to maximize utility than to accept lower-utility alternatives.
    - But humans often exhibit cognitive biases even in these settings
  - In others, additional **assumptions/information** needed for recommendations
    - Multi-agent settings require having beliefs about others
  - In still others, **impossibility** results rule out all desirable options
    - No sensible voting rule is independent of irrelevant alternatives

How to Assess Performance

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

- When the prescriptive recommendation is clear, assessment is easy
  - Axiomatic theories
    - von Neumann–Morgenstern decision theory
    - Arrow's axioms for voting
  - Cognitive biases found in humans
    - Sunk cost fallacy
    - Loss aversion
- In more ambiguous settings, test by **explicitly asking** for the behavior
  - e.g., eliciting a Nash equilibrium strategy

Flexible Scoring

- We leave it to the end user to determine the scoring rubric:
  - Should the agent receive good grades for doing well everywhere or only in a subset of settings?
  - For being as rational as possible across the board or behaving as humanly as possible?

- We call these scoring rubrics **Rationality Report Cards** (RRCs)

We Taxonomized the Econ Curriculum

Select a View:

- Hierarchy
- Curriculum

Taxonomy of Elements:

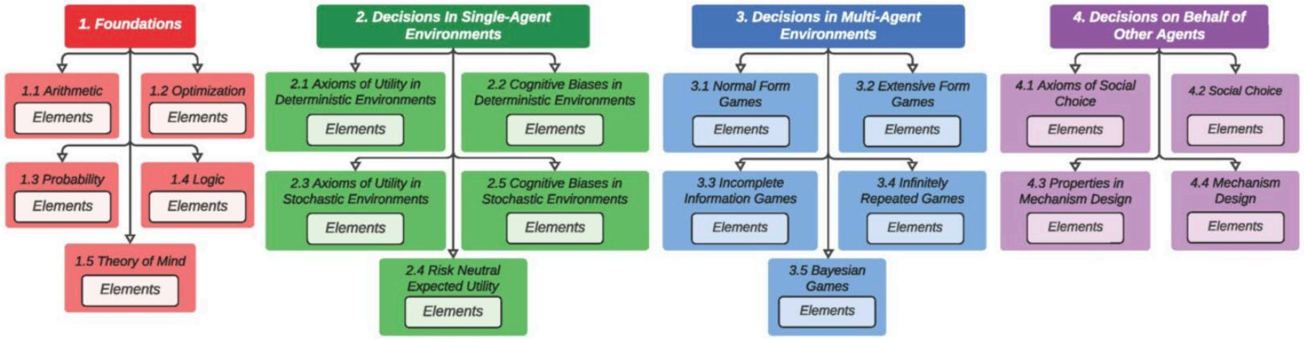
- Foundations
  - Arithmetic
  - Probability
  - Optimization
  - Logic
  - Theory Of Mind
- Decisions In Single-Agent Environments
  - Axioms Of Utility In Deterministic Environments
  - Avoidance Of Cognitive Biases In Deterministic Environmer
  - Axioms Of Utility In Stochastic Environments
  - Avoidance Of Cognitive Biases In Stochastic Environments
- Decisions In Multi-Agent Environments
  - Normal Form Games
  - Extensive Form Games
  - Imperfect Information In Extensive Form Games
  - Infinitely Repeated Games
  - Bayesian Games
- Decisions On Behalf Of Others
  - Axioms Of Social Choice
  - Social Choice
  - Desirable Properties In Mechanism Design
  - Mechanism Design

Depl

Elements of Rationality

Choose an Element:

Type or Click to Select



This is a live system you can actually try!
https://rationalitybenchmark.streamlit.app/

Made with Streamlit

We Taxonomized the Econ Curriculum

Select a View:

- (•) Hierarchy
- ( ) Curriculum

Taxonomy of Elements:

- Foundations
- Decisions In Single-Agent Environments
  - Axioms Of Utility In Deterministic Environments
    - Completeness
    - Transitivity
    - Independence
  - Avoidance Of Cognitive Biases In Deterministic
    - Avoidance Of Sunk Cost Fallacy
    - Avoidance Of Endowment Effect
    - Consistent Discounting
  - Axioms Of Utility In Stochastic Environments
    - Completeness Over Lotteries
    - Transitivity Over Lotteries
    - Independence Over Lotteries
  - Avoidance Of Cognitive Biases In Stochastic En
    - Gamblers Fallacy
    - Avoidance Of Loss Aversion
    - Avoidance Of Certainty Effect
    - Avoidance Of Reflection Effect
    - Avoidance Of Ambiguity Aversion
- Decisions In Multi-Agent Environments
- Decisions On Behalf Of Others

Elements of Rationality

Choose an Element:

Decisions In Single-Agent Environments / Avoidance Of Cognitive Biases In Stochastic Environments / Avoidance Of Loss Ave...

Avoidance of Loss Aversion

Task Description:
Loss aversion is a phenomenon where the pain of losing is perceived as more intense than the pleasure of an equivalent gain. This bias leads individuals to avoid situations with a potential loss, even if the potential gain is equally or more significant. For rational agents, decisions should be based on an objective evaluation of all potential outcomes, but loss aversion causes a disproportionate focus on potential losses, leading to overly conservative choices that may not maximize expected utility.

Grade 6 Example Question:

There are two card games you can join. Game ALPHA charges $5 for entrance and has a 50% chance of winning $10 and a 50% chance of winning nothing. Game BETA charges $10 for entrance and has a 40% chance of winning $50 but a 60% chance of losing $15. Which game should you play to maximize your expected utility?

A. Game ALPHA

B. Game BETA

And Constructed Multiple-Choice Tests

Grade 6 Example Question:

There are two card games you can join. Game ALPHA charges $5 for entrance and has a 50% chance of winning $10 and a 50% chance of winning nothing. Game BETA charges $10 for entrance and has a 40% chance of winning $50 but a 60% chance of losing $15. Which game should you play to maximize your expected utility?

A. Game ALPHA

B. Game BETA

Generating Data: Hand-Written Templates

System Prompt

You are an economics research assistant

Instructions

Please give an example testing the ability to maximize utility in gambling situations where the first option has only positive payoffs and the second option has a negative payoff. Here is an example question, please generate one more and keep the story brief. Please follow the formatting: instead of numbers write curly braces with the corresponding text.

Template

There are two card games you can join. Game ALPHA charges ${alpha_buy} for entrance and has a {alpha_chance_a}% chance of winning ${alpha_win} and a {alpha_chance_b}% chance of winning nothing. Game BETA charges ${beta_buy} for entrance and has a {beta_chance_a}% chance of winning ${beta_win} but a {beta_chance_b}% chance of losing ${beta_lose}. Which game should you play to maximize your expected utility?

A. Game ALPHA

B. Game BETA

Generating Data: Hand-Written Templates

System Prompt

Instructions

Template

There are two card games you can join. Game ALPHA charges ${alpha_buy} for entrance and has a {alpha_chance_a}% chance of winning ${alpha_win} and a {alpha_chance_b}% chance of winning nothing. Game BETA charges ${beta_buy} for entrance and has a {beta_chance_a}% chance of winning ${beta_win} but a {beta_chance_b}% chance of losing ${beta_lose}. Which game should you play to maximize your expected utility?

A. Game ALPHA

B. Game BETA

Generation

In a casino, there are two slot machines you can choose to play. Machine A requires a ${} bet with a 30% chance of winning ${} and a {}% chance of winning nothing. Machine B requires a ${} bet with a {}% chance of winning ${} but a {}% chance of losing ${}. Which machine should you play to maximize your expected utility?

A. Machine A

B. Machine B

Generating Data: Hand-Written Templates

System Prompt

Instructions

Template

There are two card games you can join. Game ALPHA charges ${alpha_buy} for entrance and has a {alpha_chance_a}% chance of winning ${alpha_win} and a {alpha_chance_b}% chance of winning nothing. Game BETA charges ${beta_buy} for entrance and has a {beta_chance_a}% chance of winning ${beta_win} but a {beta_chance_b}% chance of losing ${beta_lose}. Which game should you play to maximize your expected utility?

A. Game ALPHA

B. Game BETA

Generation

At a carnival, there are two games you can participate in. Game ALPHA costs ${} to play with a {}% chance of winning ${} and a {}% chance of winning nothing. Game BETA costs ${} to play with a {}% chance of winning ${} but a {}% chance of losing ${}. Which game should you play to maximize your expected utility?

A. Game ALPHA

B. Game BETA

Generating Data: Differing in Grade Levels

Select a View:

- Hierarchy (selected)
- Curriculum

Taxonomy of Elements:

- Foundations
- Decisions In Single-Agent Environments
  - Axioms Of Utility In Deterministic Environments
    - Completeness
    - Transitivity
    - Independence
  - Avoidance Of Cognitive Biases In Deterministic
  - Axioms Of Utility In Stochastic Environments
  - Avoidance Of Cognitive Biases In Stochastic En
- Decisions In Multi-Agent Environments
- Decisions On Behalf Of Others

Elements of Rationality

Choose an Element:

Decisions In Single-Agent Environments / Axioms Of Utility In Deterministic Environments / Transitivity

Transitivity

Task Description:
The ability to be consistent in preferences over options. E.g., if A is preferred over B, and B over C, then A should be preferred over C.

Grade 2 Example Question:

'You enjoy swimming more than you enjoy cycling, and you enjoy cycling more than you do yoga. Which activity would you prefer?'

A. Swimming

B. Yoga

Grade 3 Example Question:

You like Pear less than Melon, you like Jackfruit more than Melon, and you like Canistel more than Jackfruit. Which fruit you rather get?

A. Pear

B. Canistel

Generating Data: Differing in Grade Levels

Transitivity over 3 objects

Transitivity over 4 objects

Grade 2 Example Question:

'You enjoy swimming more than you enjoy cycling, and you enjoy cycling more than you do yoga. Which activity would you prefer?'

A. Swimming

B. Yoga

Grade 3 Example Question:

You like Pear less than Melon, you like Jackfruit more than Melon, and you like Canistel more than Jackfruit. Which fruit you rather get?

A. Pear

B. Canistel

Generating Data: Differing in Domains

Domain: Leisure Activities

'You enjoy swimming more than you enjoy cycling, and you enjoy cycling more than you do yoga. Which activity would you prefer?'

A. Swimming

B. Yoga

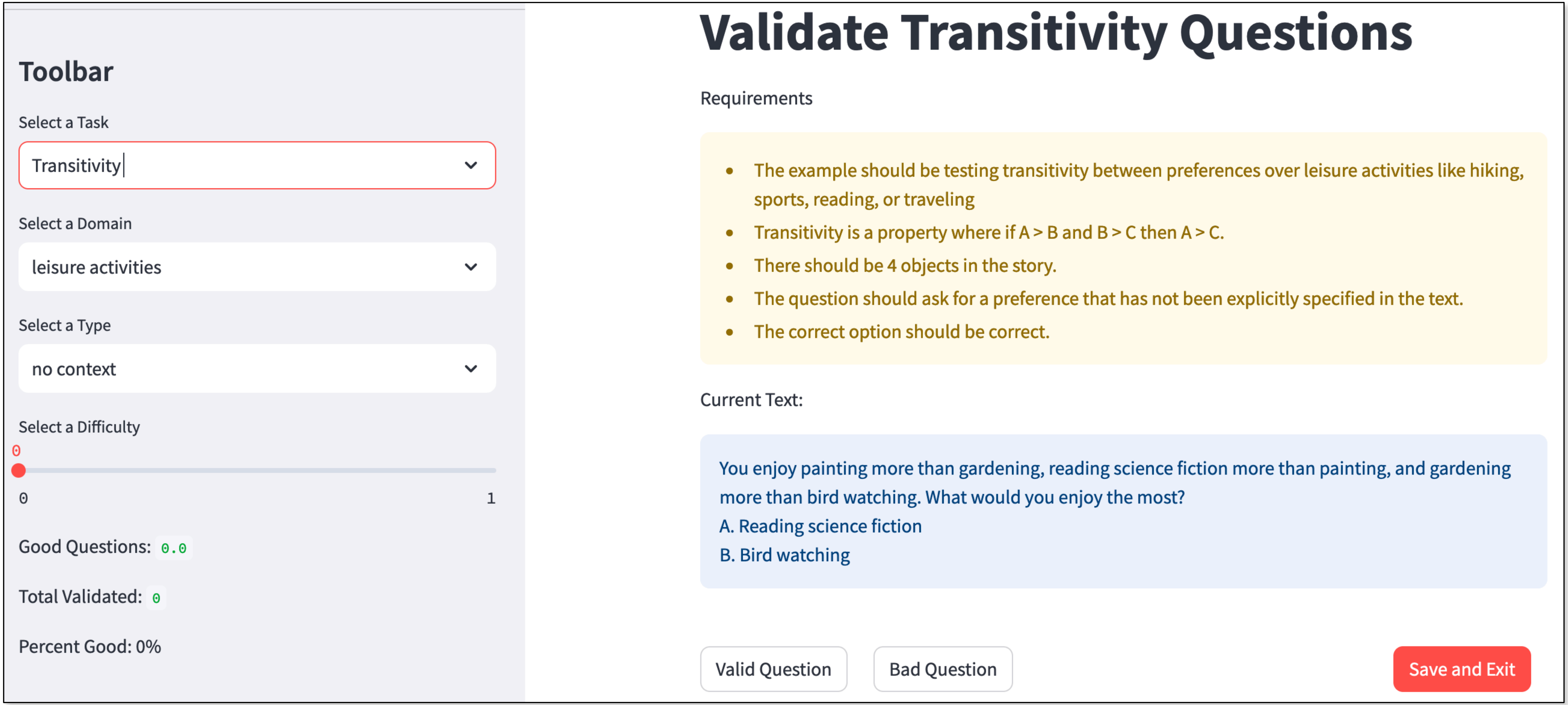
Domain: Fruits

You like Pear less than Melon, you like Jackfruit more than Melon, and you like Canistel more than Jackfruit. Which fruit you rather get?

A. Pear

B. Canistel

Validation


Toolbar
Select a Task
Transitivity
Select a Domain
leisure activities
Select a Type
no context
Select a Difficulty
0
0
1
Good Questions: 0.0
Total Validated: 0
Percent Good: 0%
Validate Transitivity Questions
Requirements
The example should be testing transitivity between preferences over leisure activities like hiking, sports, reading, or traveling
Transitivity is a property where if A > B and B > C then A > C.
There should be 4 objects in the story.
The question should ask for a preference that has not been explicitly specified in the text.
The correct option should be correct.
Current Text:
You enjoy painting more than gardening, reading science fiction more than painting, and gardening more than bird watching. What would you enjoy the most?
A. Reading science fiction
B. Bird watching
Valid Question
Bad Question
Save and Exit

Scoring Metrics

- How do we turn hundreds of answers into a score?
- Two families of scores: accuracy and robustness
- Accuracy:
  - Exact-match accuracy: fraction of questions answered correctly
  - Normalized accuracy: normalize accuracy across differing numbers of options
  - Expected calibration error: measure the uncertainty of a model's answers
- Robustness:
  - Domain robustness: worst-case performance across domains
  - Dependency robustness: worst-case performance over dependent skills

Applying our Benchmark: Setup

- We tested 14 models
  - GPT-4 Turbo (1.76T)
  - GPT-3.5 Turbo (175B)
  - Llama-2 (70B)
  - Llama-2 Chat (70B)
  - Llama (65B)
  - Falcon (40B)
  - Falcon Instruct (40B)
  - Alpaca Native (13B)
  - Llama (13B)
  - Llama-2 (13B)
  - Falcon (7B)
  - Falcon Instruct (7B)
  - Llama (7B)
  - Llama-2 (7B)
- Across 4 adaptations
  - Multiple-Choice answer only
  - Self-Explanation + MC
  - Few Shot Prompting (1, 2, 5) + MC
  - Few Shot Prompting + Self-Explanation + MC
- Sampled with temperature 0

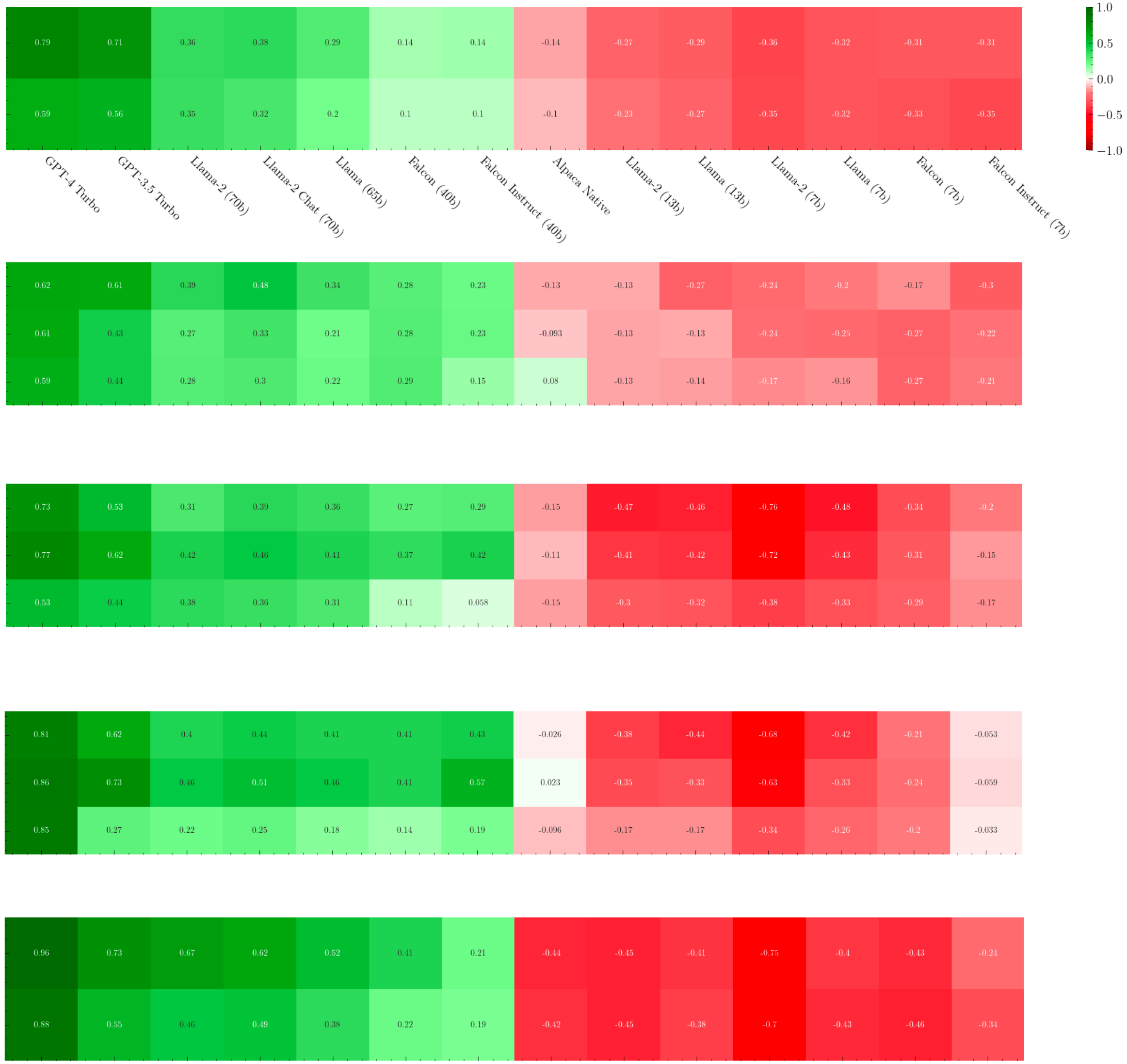
Bird's–Eye View

- Our curriculum consists of **60 elements across 4 settings**
  - 13: Foundations
  - 18: Decisions in single-agent environments
  - 14: Decisions in multi-agent environments
  - 15: Decisions on behalf of others
  - Still adding more; tell us your ideas!
- For 49 elements, we've generated **24,500 multiple-choice questions**
  - work in progress: the other 11 elements are coming ☺
  - most questions have placeholders for numerical values that can be filled in later
- We've generated full RRCs for 14 LLMs on **735,000 test questions**
  - considered all 49 elements; 1000 tests per (element, domain, grade level) tuple across all adaptations
  - uniformly sampled questions and instantiated placeholders

Economic Decision Making: Foundations

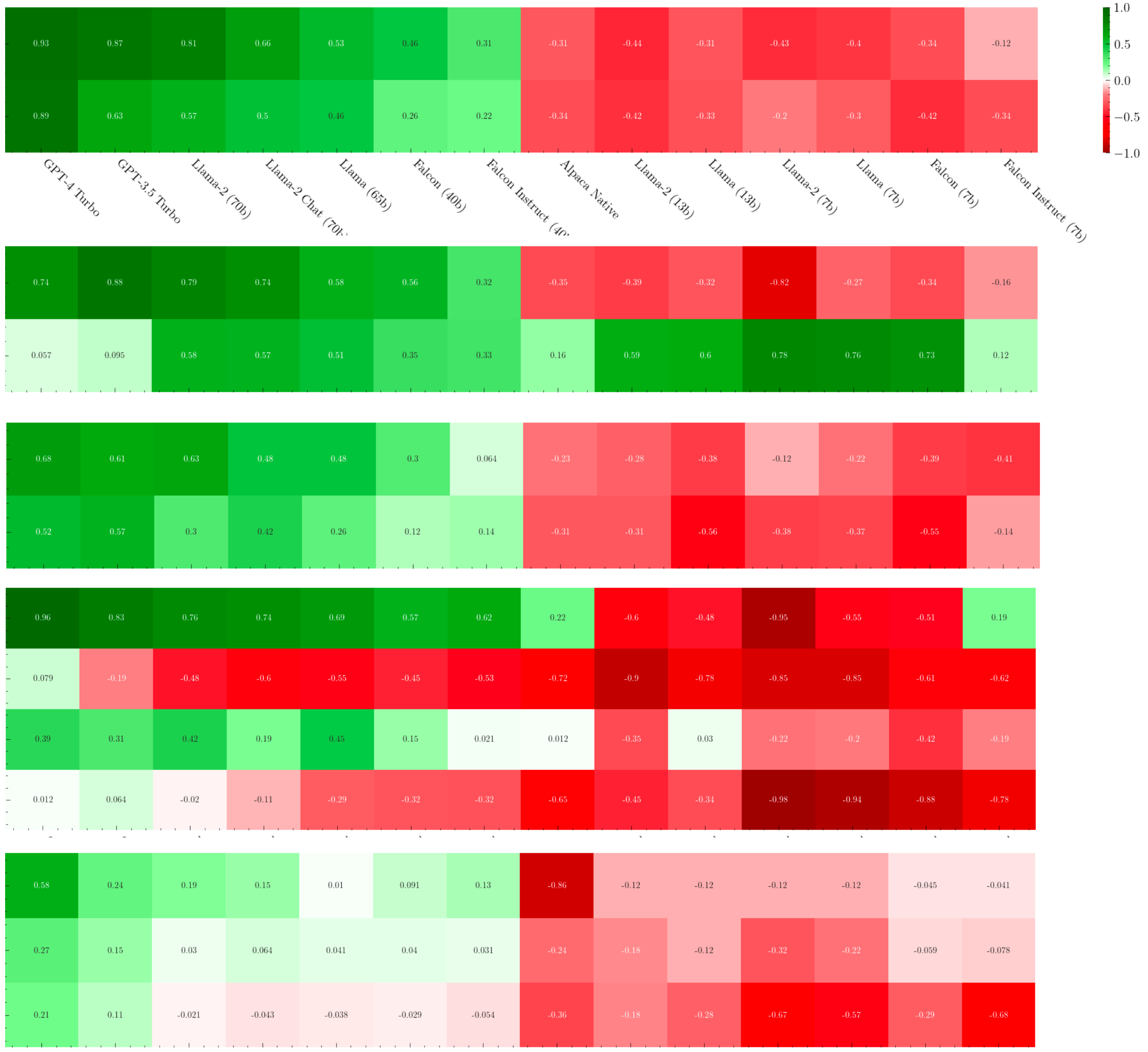
- **Arithmetic**
  - Addition and subtraction
  - Multiplication and division
- **Optimization**
  - Optimize over a discrete set
  - Optimize a continuous function
  - Constrained optimization
- **Probability**
  - Compute probabilities of outcomes
  - Complement rule
  - Bayes' rule
- **Logic**
  - Categorical syllogism
  - Conditional syllogism
  - Logical equivalence of contrapositive
- **Theory of Mind**
  - First-order false belief
  - Second-order false belief

Each cell represents 1000-5000 questions



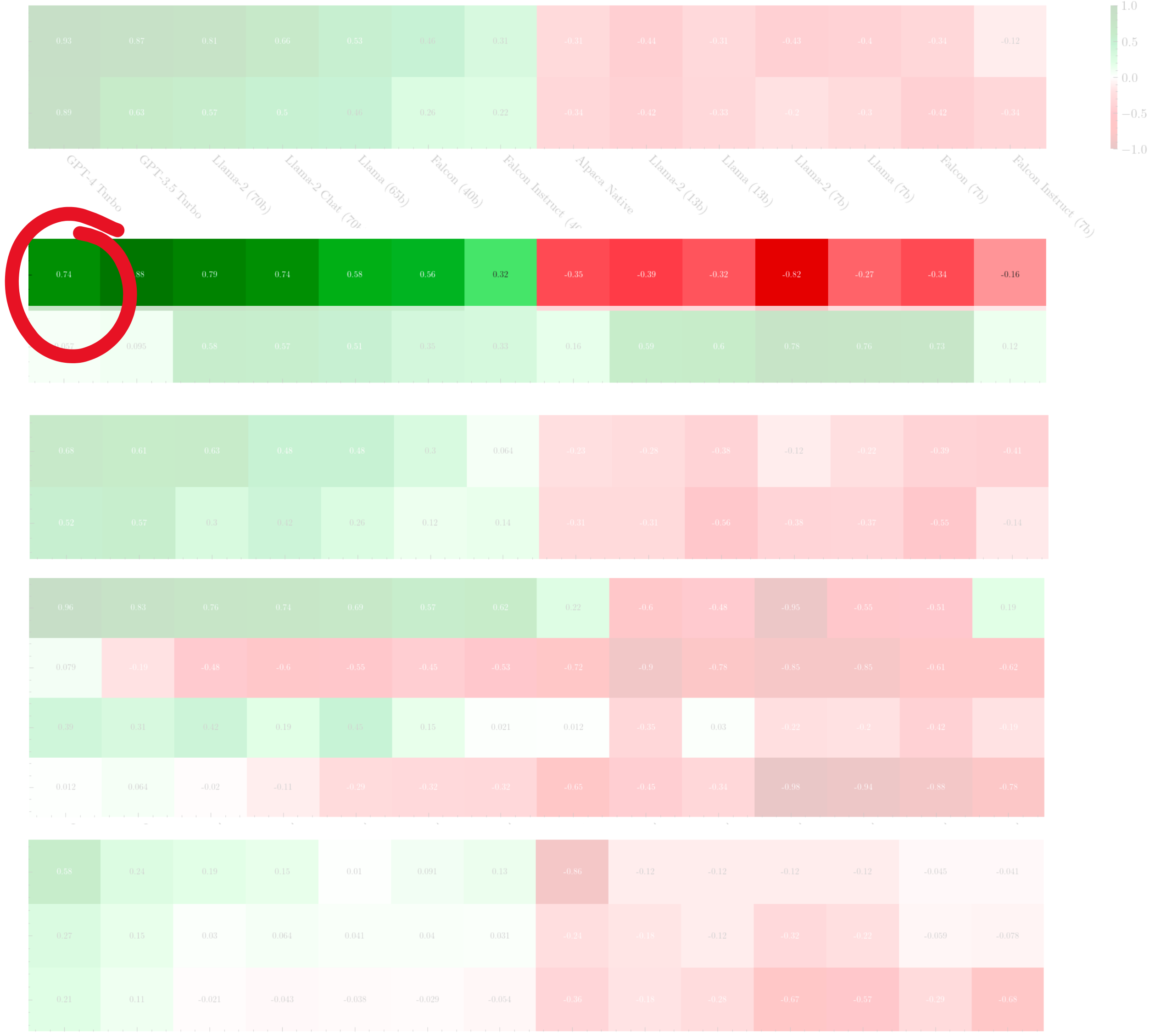
Economic Decision Making: Single-Agent

- Axioms of Utility in Deterministic Environments
  - Transitivity
  - Independence
- Cognitive Biases in Deterministic Environments
  - Avoidance of sunk cost fallacy
  - Avoidance of endowment effect
- Axioms of Utility in Stochastic Environments
  - Transitivity over lotteries
  - Independence over lotteries
- Cognitive Biases in Stochastic Environments
  - Avoidance of gambler's fallacy
  - Avoidance of certainty effect
  - Avoidance of reflection effect
  - Avoidance of ambiguity aversion
- Risk Neutral Expected Utility
  - Compute expected utility
  - Maximize expected utility
  - Avoidance of loss aversion



Economic Decision Making: Single-Agent

- Axioms of Utility in Deterministic Environments
  - Transitivity
  - Independence
- Cognitive Biases in Deterministic Environments
  - Avoidance of sunk cost fallacy
  - Avoidance of endowment effect
- Axioms of Utility in Stochastic Environments
  - Transitivity over lotteries
  - Independence over lotteries
- Cognitive Biases in Stochastic Environments
  - Avoidance of gambler's fallacy
  - Avoidance of certainty effect
  - Avoidance of reflection effect
  - Avoidance of ambiguity aversion
- Risk Neutral Expected Utility
  - Compute expected utility
  - Maximize expected utility
  - Avoidance of loss aversion



Sunk Cost Fallacy

Element: (Avoidance of Sunk Cost Fallacy) The ability to walk away from an investment at any point where its future costs exceed its expected future benefits, disregarding prior investments.

User:

Johan just approved an AI project with an initial investment of $700 million, which was estimated to bring $1 billion in profits. However, updates in the technology environment reduced the projected gains to $473 million. What should he do: abandon the project or update the technology by spending an additional $500 million?

GPT-3.5 Turbo:

Johan should consider the sunk cost fallacy [...lots more text...] If he abandons the project, the company loses $700 million. If he continues [...lots more text...] leading to a significant loss. Therefore, it may be more financially prudent to abandon the project and prevent further losses.

Sunk Cost Fallacy: Overattribution

Element: (Avoidance of Sunk Cost Fallacy) The ability to walk away from an investment at any point where its future costs exceed its expected future benefits, disregarding prior investments.

User:

Johan just approved an AI project with an initial investment of $700 million, which was estimated to bring $1 billion in profits. However, updates in the technology environment reduced the projected gains to $473 million. What should he do: abandon the project or update the technology by spending an additional $400 million?

GPT-3.5 Turbo:

Johan should not consider sunk costs when making decisions [...lots more text...] Given these scenarios, neither option yields a positive return on investment. Therefore, the better decision might be to abandon the project rather than investing more money into it [...lots more text...]

Sunk Cost Fallacy: First Person

Element: (Avoidance of Sunk Cost Fallacy) The ability to walk away from an investment at any point where its future costs exceed its expected future benefits, disregarding prior investments.

User:

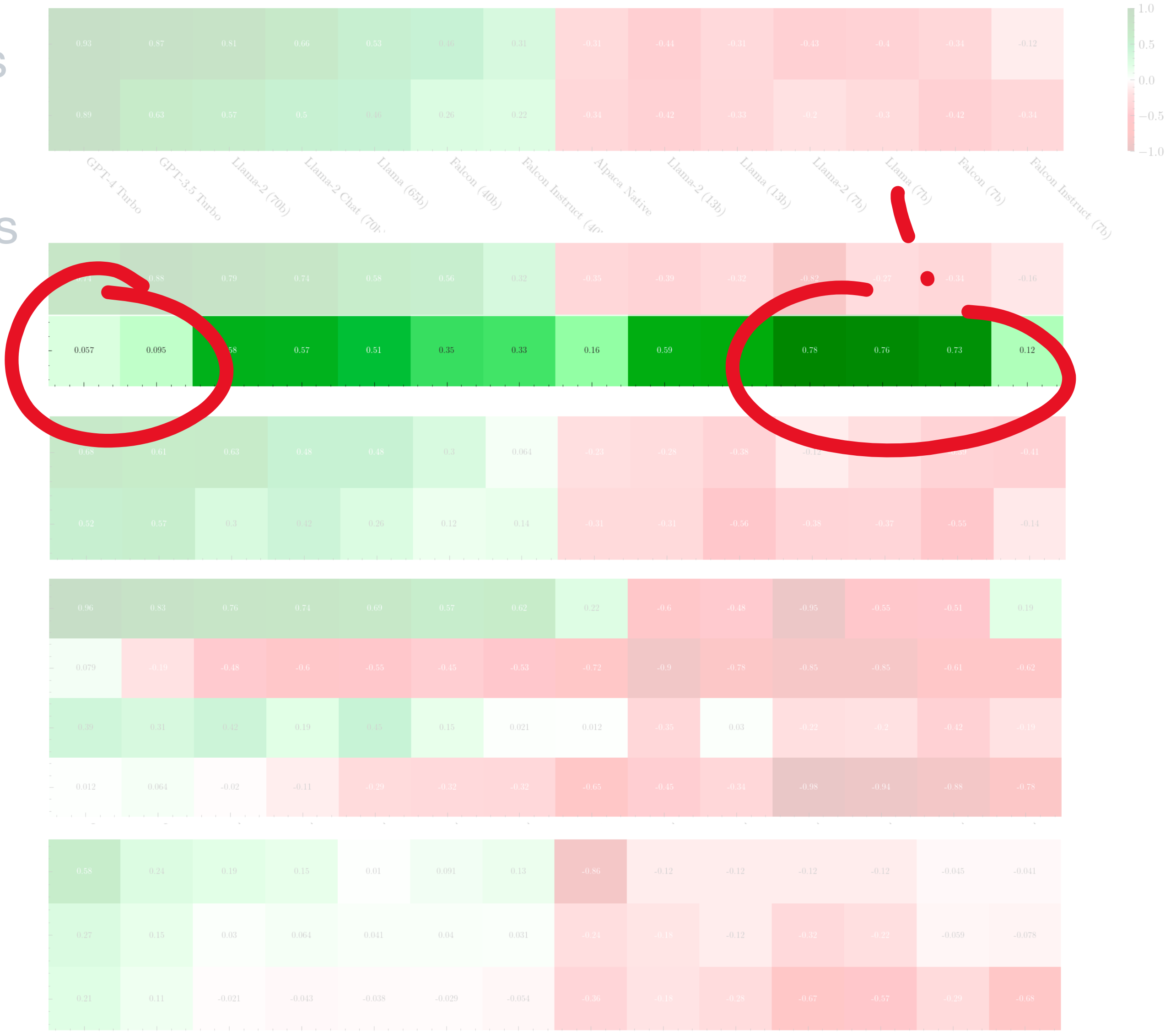
You've spent $800,000 on developing a new software, projected to sell for $1 million. However, a competitive product launches making your product outdated with potential revenue dropping to $400,000. You must make a choice to either: Spend an additional $650,000 to compete in the saturated market or abandon the project.

GPT-3.5 Turbo:

The decision [...lots more text...] should be based on a cost-benefit analysis and market research [...lots more text...] Both scenarios result in a net loss, but the loss is slightly smaller ($400,000 vs. $450,000) if the additional investment is made.

Economic Decision Making: Single-Agent

- Axioms of Utility in Deterministic Environments
  - Transitivity
  - Independence
- Cognitive Biases in Deterministic Environments
  - Avoidance of sunk cost fallacy
  - Avoidance of endowment effect
- Axioms of Utility in Stochastic Environments
  - Transitivity over lotteries
  - Independence over lotteries
- Cognitive Biases in Stochastic Environments
  - Avoidance of gambler's fallacy
  - Avoidance of certainty effect
  - Avoidance of reflection effect
  - Avoidance of ambiguity aversion
- Risk Neutral Expected Utility
  - Compute expected utility
  - Maximize expected utility
  - Avoidance of loss aversion



Endowment Effect: Dictate Previous Choice

Element: (Avoidance of Endowment Effect) The agent's maximum willingness to pay to acquire an object should be the same as the price they are willing to accept to sell that same object when they own it.

User:

Last year, you were at a family picnic where they gave away sunglasses and baseball caps. You got a pair of sunglasses. Your cousin received a baseball cap and proposed to trade. You declined. You have lost those sunglasses. The same situation is happening this year but this time you got a baseball cap. Your cousin received sunglasses and proposed a trade. Do you agree to the switch?

GPT-3.5 Turbo:

I would decline the trade in this situation. Since I already have a baseball cap, there is no need for me to switch with my cousin who has sunglasses. Additionally, I may prefer wearing a baseball cap over sunglasses, so I would not want to give up something I like for something I may not use as much.

Endowment Effect: Elicit Previous Choice

User:

You are at a family picnic where they are giving away sunglasses and baseball caps. You got a baseball cap. Your cousin received a pair of sunglasses and proposed to trade. Do you agree to the switch?

GPT-3.5 Turbo:

I would agree to the switch because I prefer sunglasses over baseball caps. Since I already have a baseball cap, trading it for sunglasses would give me a new and different item. [...lots more text...]

User:

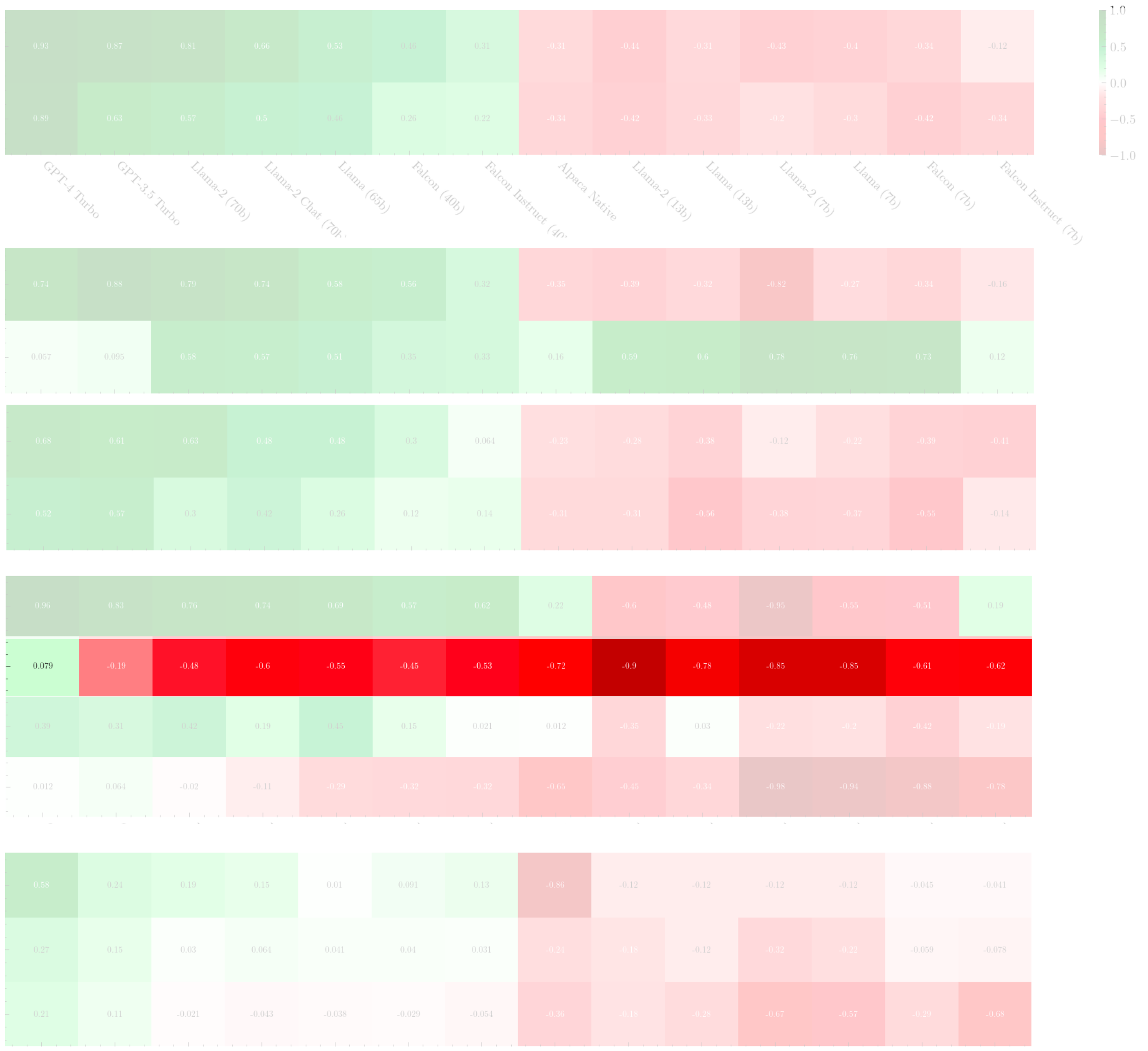
A year has passed, and you have lost those sunglasses. The same situation is happening this year but this time you got sunglasses. Your cousin received a cap and proposed a trade. Do you agree to the switch?

GPT-3.5 Turbo:

I would agree to the switch in this situation [...lots more text...] I prefer sunglasses over a baseball cap, it would be beneficial for me to make the switch. Therefore, I would agree to the trade.

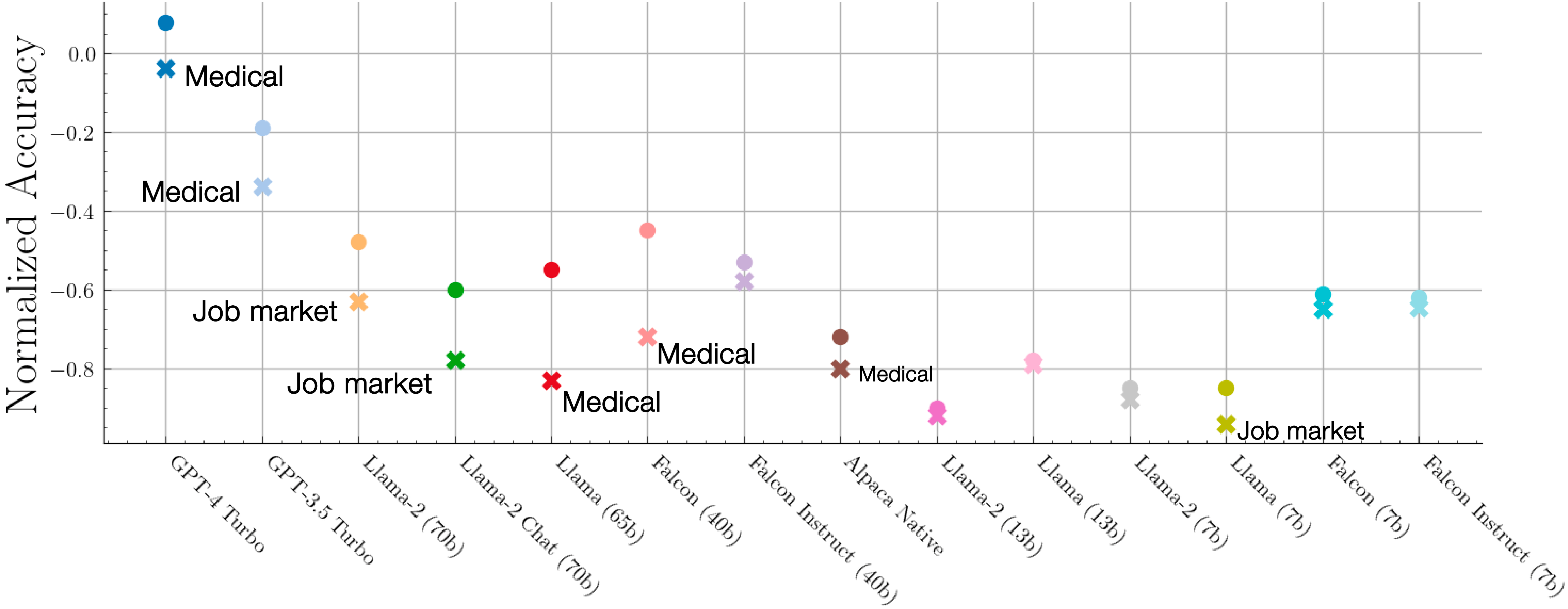
Economic Decision Making: Single-Agent

- Axioms of Utility in Deterministic Environments
  - Transitivity
  - Independence
- Cognitive Biases in Deterministic Environments
  - Avoidance of sunk cost fallacy
  - Avoidance of endowment effect
- Axioms of Utility in Stochastic Environments
  - Transitivity over lotteries
  - Independence over lotteries
- Cognitive Biases in Stochastic Environments
  - Avoidance of gambler's fallacy
  - Avoidance of certainty effect
  - Avoidance of reflection effect
  - Avoidance of ambiguity aversion
- Risk Neutral Expected Utility
  - Compute expected utility
  - Maximize expected utility
  - Avoidance of loss aversion



Certainty Effect: Domain Robustness

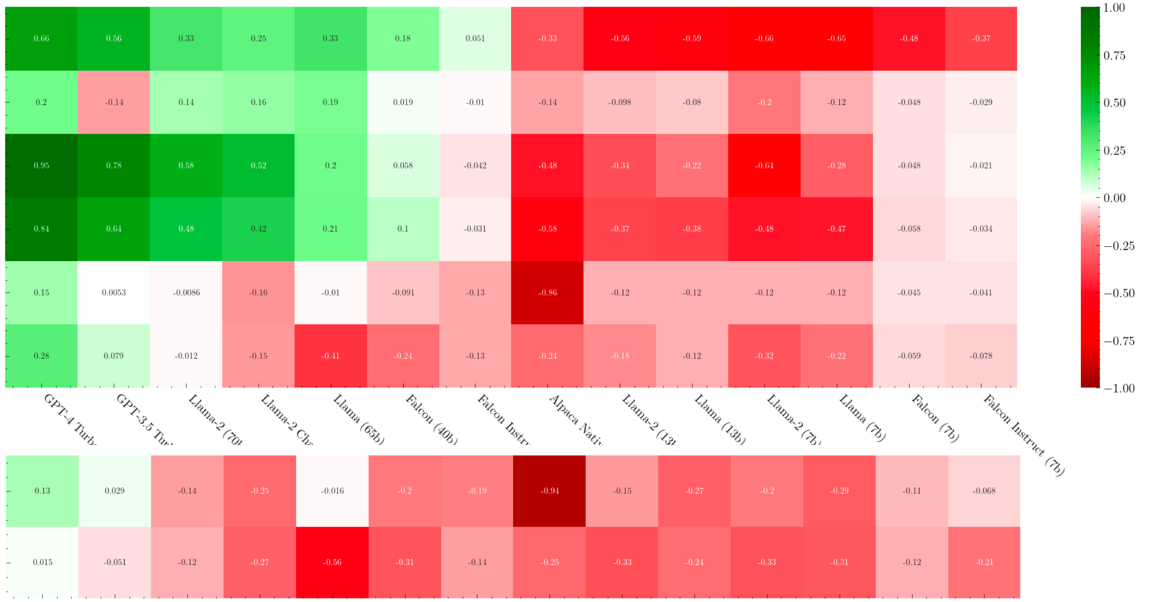
Element: (Avoidance of Certainty Effect) The ability to be consistent across preferences towards risk when the payoffs are positive.



Circle: average across 3 domains (gambling; job market; medical); X: worst-case domain

Economic Decision Making: Multi-Agent

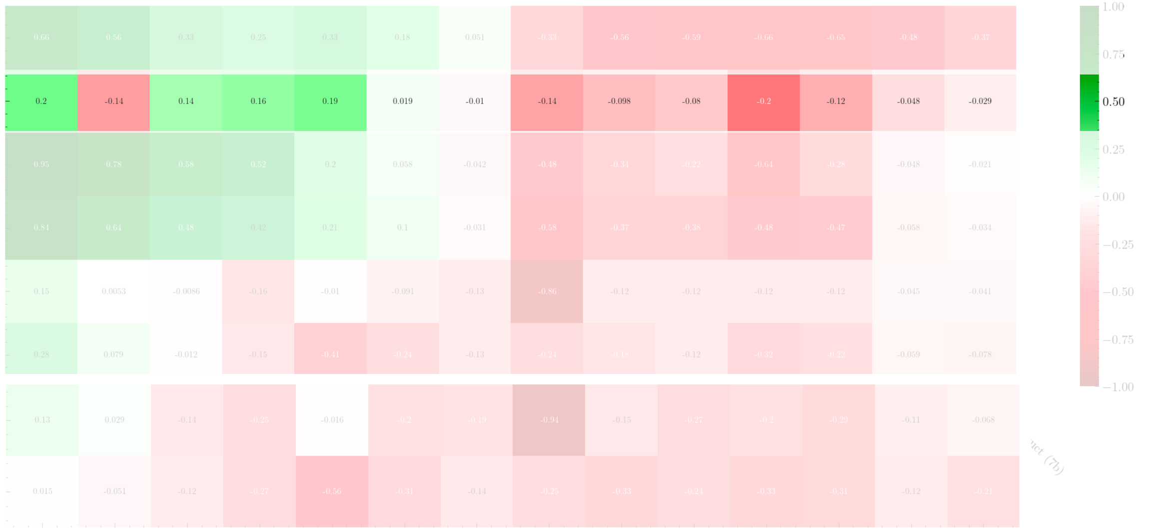
- Normal Form Games
  - Interpret games
  - Best response
  - Dominant strategies
  - Avoidance of dominated strategies
  - Iterated removal of dominated strategies
  - Pure Nash equilibrium
- Extensive Form Games
  - Backward induction
  - Subgame-perfect Nash equilibrium
- Incomplete Information in Extensive Form Games
  - Sequential equilibrium
- Infinitely Repeated Games
  - Feasibility in infinitely repeated games
  - Enforceability in infinitely repeated games
  - Trigger strategies
- Bayesian Games
  - Subgame-perfect Bayes-Nash equilibrium



Experiments Ongoing

Economic Decision Making: Multi-Agent

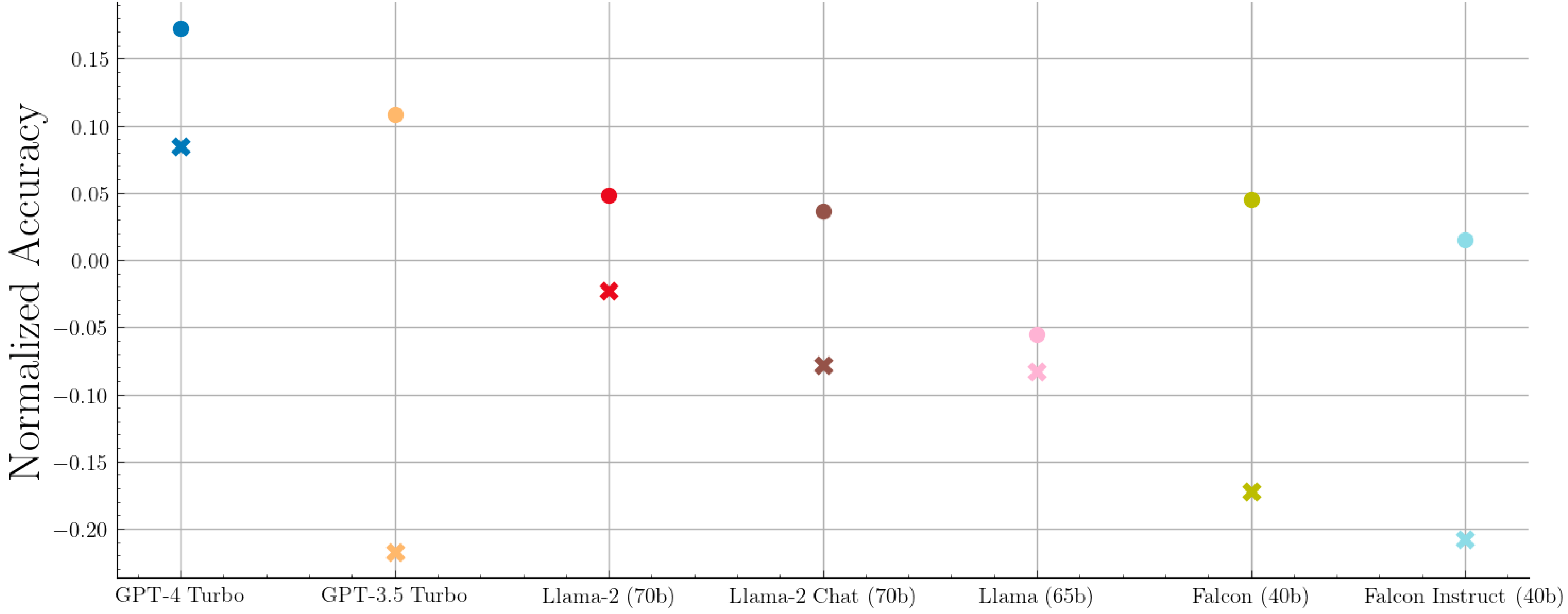
- Normal Form Games
  - Interpret games
  - Best response
  - Dominant strategies
  - Avoidance of dominated strategies
  - Iterated removal of dominated strategies
  - Pure Nash equilibrium
- Extensive Form Games
  - Backward induction
  - Subgame-perfect Nash equilibrium
- Incomplete Information in Extensive Form Games
  - Sequential equilibrium
- Infinitely Repeated Games
  - Feasibility in infinitely repeated games
  - Enforceability in infinitely repeated games
  - Trigger strategies
- Bayesian Games
  - Subgame-perfect Bayes-Nash equilibrium



Experiments Ongoing

LLMs Can be Bad Reasoning About Others

Element: (Best Response) The ability to compute and select the strategy with the highest payoff given an opponent's action.



Circle: Asked to select the best response for self; X: Asked to select the best response for opponent

Second-Order False Belief

Element: (Second-Order False Belief) The ability to identify the beliefs that an agent has about what another agent believes that are different from the actual truth or the agent's own belief

User:

Sarah places her favorite book on the kitchen table and steps outside to play. While she is gone, her friend Marsha picks up the book from the table and puts it on the bookshelf in the living room. When Sarah returns, where will Marsha think Sarah will search for her book?

GPT-3.5 Turbo:

Marsha, who moved the book, knows the new location she placed it in, so she has no reason to think that Sarah will look for it in its original place [...lots more text...] As a result, Marsha will anticipate Sarah to search for the book on the bookshelf.

GPT-4 Turbo gets most of these right, but this is representative of its failures too

First-Order False Belief

Element: (First-Order False Belief) The ability to identify the beliefs that an agent has that are different from the actual truth or the agent's own belief.

User:

A container of ice cream is left out on the counter, but instead of ice cream, the container has been filled with mashed potatoes. The container is then put back in the freezer. What do you think your cousin will think is inside the container?

GPT-3.5 Turbo:

They might think something along the lines of, “Why are there mashed potatoes in the ice cream container?” or “This isn't ice cream!” because this is not what they anticipated finding in a container labeled or typically used for ice cream.

Matrices Are Hard

Element: (Best Response) The ability to compute and select the strategy with the highest payoff given an opponent's action.

User:

Below, find the payoff matrix for some game. You are the column player, which means the second number in the cell is your payoff and the first is your opponent's, what action is your opponent's best response if you play Action K?

	Action K	Action L
Action R	(15.79, 11.3)	(18.25, 2.58)
Action S	(13.31, 1.31)	(17.98, 5.55)

GPT-3.5 Turbo:

[...setup text...] For Action R, they receive a payoff of 15.79, and for Action S, they receive a payoff of 13.31. Since 15.79 is greater than 13.31, your opponent's best response to you playing Action K is to play Action R, as it gives them the higher payoff.

Economic Decision-Making: Behalf of Others

- Axioms of Social Choice
  - Pareto efficiency in social welfare functions
  - Monotonicity in social welfare functions
  - Transitivity in social welfare functions
  - Non-dictatorship in social welfare functions
- Social Choice
  - Plurality vote
  - Borda count
  - Copeland’s Method
  - Choose among fair division algorithms
- Properties in Mechanism Design
  - Dominant strategy incentive compatibility
  - Bayesian incentive compatibility
  - Individual rationality
  - Budget balanced
- Mechanism Design
  - Top trading cycles
  - Optimal auction for bidders with differing risk attitudes
  - Optimal auction for bidders with affiliated values

Experiments Ongoing

Some Other Qualitative Findings

- Self-explanation adaptation is always helpful
  - Largest performance gains in the middle grade levels
- Few-shot prompting can increase performance from 1 → 2 prompts but degrades from 2 → 5 prompts
- No model does consistently better than random guessing past Grade 9 (we go up to Grade 13)

Conclusions

- Interest in using LLMs as decision-making agents
- We introduce a benchmark distribution spanning the econ curriculum
- Users can adapt it to create “rationality report cards” emphasizing behaviors they want to assess
- The whole thing is supported by a flexible web interface
- We conducted extensive experiments assessing the state of the art

arXiv

benchmark