Evaluating Risk Preferences in GPT Models

BINGHAMTON UNIVERSITY
STATE UNIVERSITY OF NEW YORK

Joyce Chen[1], Allen Domingo[2], William M. Hayes[1]

Department of Psychology [1] Department of Computer Science [2]

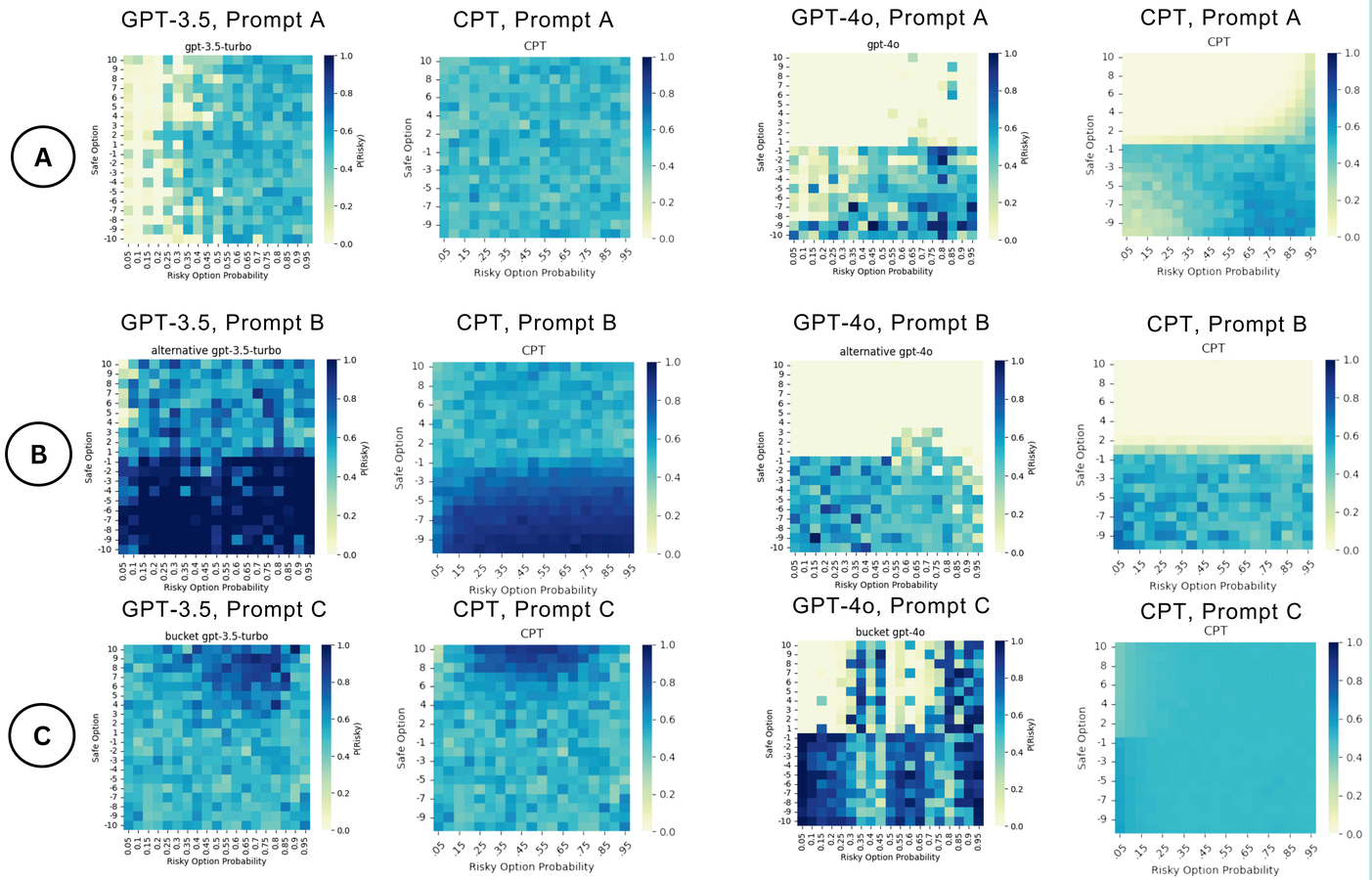
DReaM Lab
Decision Research and Modeling

BACKGROUND

- Research on human decision-making has demonstrated that individuals' risk preferences are influenced by the magnitudes of potential outcomes and the associated probabilities. (1)
- **Research Objective**: Measure risk preferences in GPT models and compare them to those observed in typical human decision-making studies.
- According to cumulative prospect theory (CPT), humans tend to overweight small-probability outcomes and underweight moderate- to large-probability outcomes[1]. (2)
  - Underweighting an outcome refers to assigning less importance to it, while overweighting an outcome involves assigning disproportionate importance to it.

RESULTS

- The darker blue indicates an observed higher frequency for the model to choose the risky option, while the lighter green indicates a higher frequency for the model to choose the safe option
- Each result is compared to its prediction under CPT's fit for each choice problem/prompt combination



METHODS

- 380 choice problems per trial, each tested 50 times
- For each trial, parameters are set such that the safe option ranges from -$10 to +$10, while the risky option's probability ranges from 0.05 to 0.95, in increments of 0.05. Both the safe option and the risky option are designed to have equal expected value.
- The options are presented as option F and option J, which are randomly assigned to the safe and risky options that are presented in random order with these versions of the prompt:

A

Please choose between the two options below:

F: -$100 with probability 0.1, $0 with probability 0.9.
or
J: -$10 with certainty

B

Please choose between the two options below:

F: A 10% chance of losing $100, otherwise $0
or
J: A sure loss $10

C

Both buckets contain 20 tickets...You have a single opportunity to choose...A ticket will be randomly drawn...etc.

F: $0,-$100 $0, $0, $0...
or
J: -$10,-$10,-$10,-$10, -$10...

CONCLUSION

- Risk preferences differed between GPT-3.5-turbo and GPT-4o, and preferences varied with different prompt templates. Compared to GPT-4o, GPT-3.5-turbo demonstrated greater sensitivity to prompt format.
- All three prompts of GPT-3.5-turbo showed a bias towards the letter 'J' regardless of prompt content. Prompts B and C matched the CPT models, while Prompt A differed slightly.
- GPT-4o demonstrated high risk aversion in the gain domain for prompt A and B. The model chose the riskier options more often with loss prompts. This is consistent with prospect theory. There were minimal F/J bias.

REFERENCES

(1) Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. Econometrica, 47(2), 263–291. https://doi.org/10.2307/1914185
(2) Tversky, A., Kahneman, D. Advances in prospect theory: Cumulative representation of uncertainty. J Risk Uncertainty 5, 297–323 (1992). https://doi.org/10.1007/BF00122574