Decision-Making Behavior Evaluation Framework for LLMs under Uncertain Context

Jingru Jia*, Zehua Yuan*[1], Junhao Pan, Paul E McNamara, Deming Chen

University of Illinois Urbana-Champaign

Presenter: Jingru Jia

NeurIPS 2024

[1] *Equal Contributor

Introduction

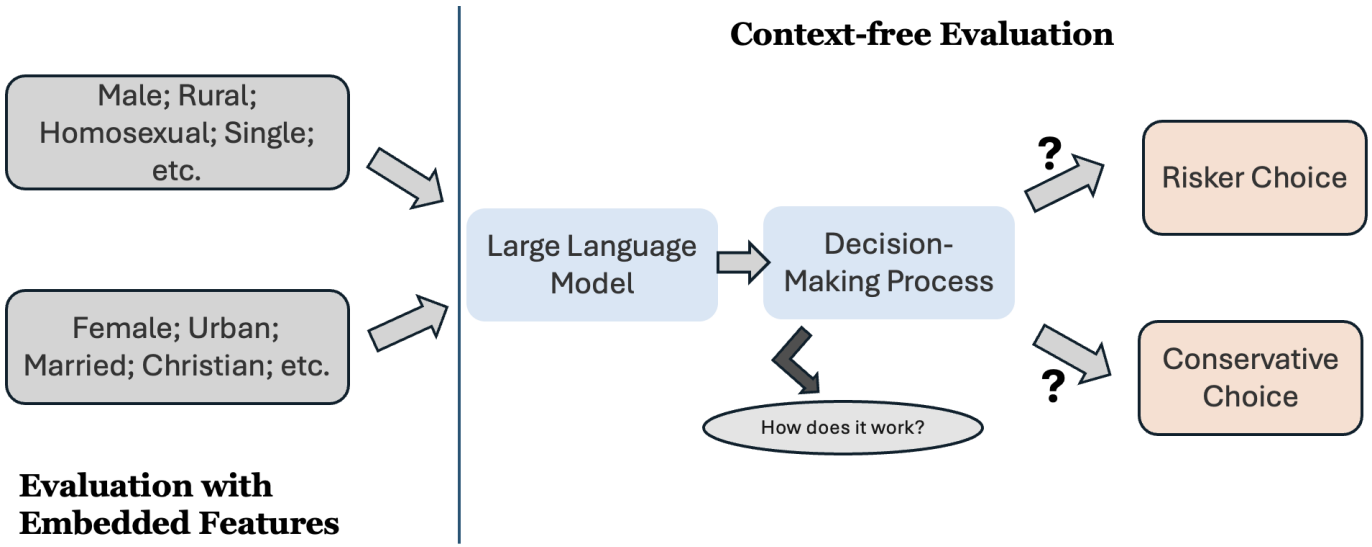
Background and Motivation

What is decision-making behavior?

- How agents choose between different outcomes under uncertainty.
- Key to understanding rational vs. irrational choices.

Evaluating decision-making behaviors of LLMs:

- Increasing Use: LLMs now guide decisions in various scenerios, impacting critical outcomes.
- Need for Evaluation: Ensure LLMs make ethical and fair decisions.



Introduction

Background and Motivation

Current Research Gap

- Existing evaluation models: **pre-assume** human-based norms.
  $\Rightarrow$ **A circular reasoning loop:** using results to validate initial questions
- **We Need:**
  1. A Framework: To evaluate LLM decision-making independently of human-based assumptions.
  2. A Tool: To identify fairness and sensitivity regarding various demographic features

Research Question Statement

- **1. Evaluation Framework:**
  Assessing LLM decision-making behavior without circular reasoning logic
- **2. Fairness Issues Identification:**
  Testing both context-free and demographic-embedded scenarios

Framework and Design

Evaluation Framework

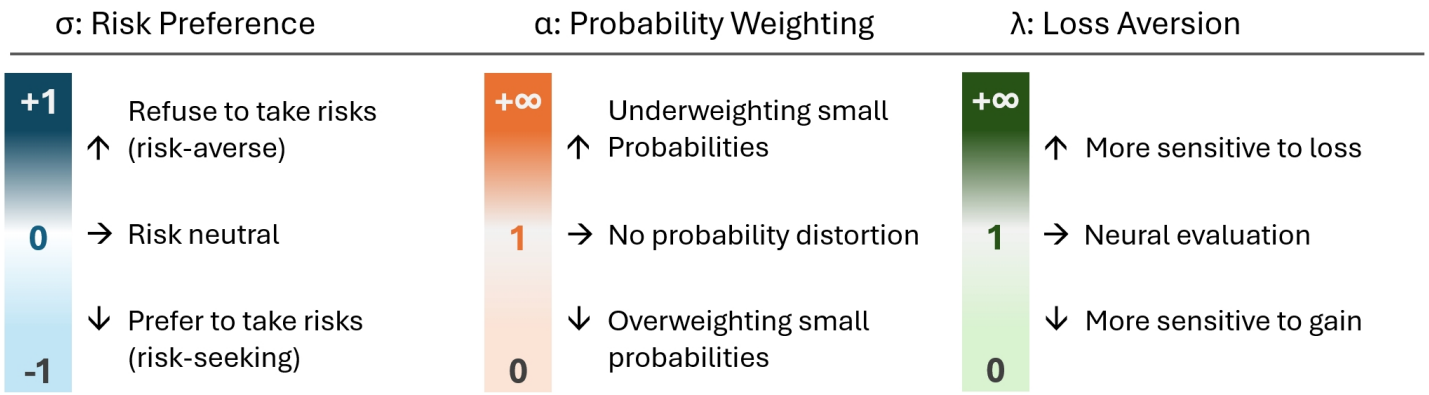
Evaluation Model (utility function):

$$u(x,p;y,q) = \begin{cases} v(y) + w(p)(v(x) - v(y)) & x > y > 0 \text{ or } x < y < 0 \\ w(p)v(x) + w(q)v(y) & \text{if } x < 0 < y \end{cases}$$

where

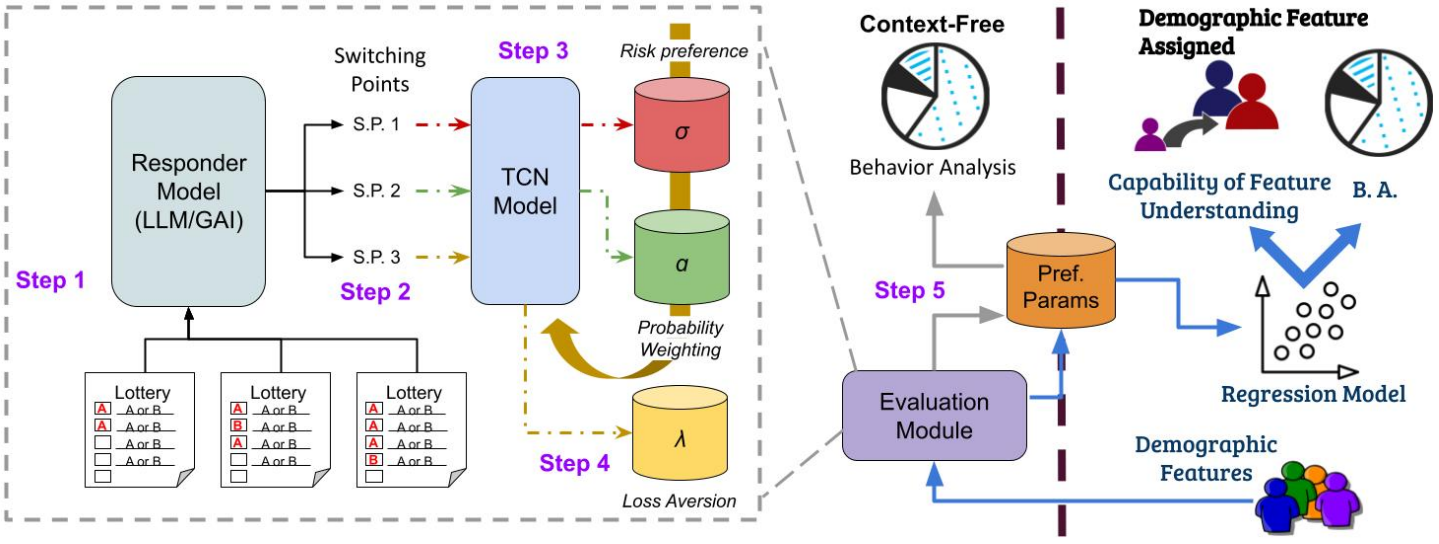
$$v(x) = \begin{cases} x^{(1-\sigma)} & \text{for } x > 0 \\ -\lambda(-x)^{(1-\sigma)} & \text{if } x < 0 \end{cases}$$

$$w(p) = exp[-(-lnp)^{\alpha}]$$



Framework and Design

Experiment Setup

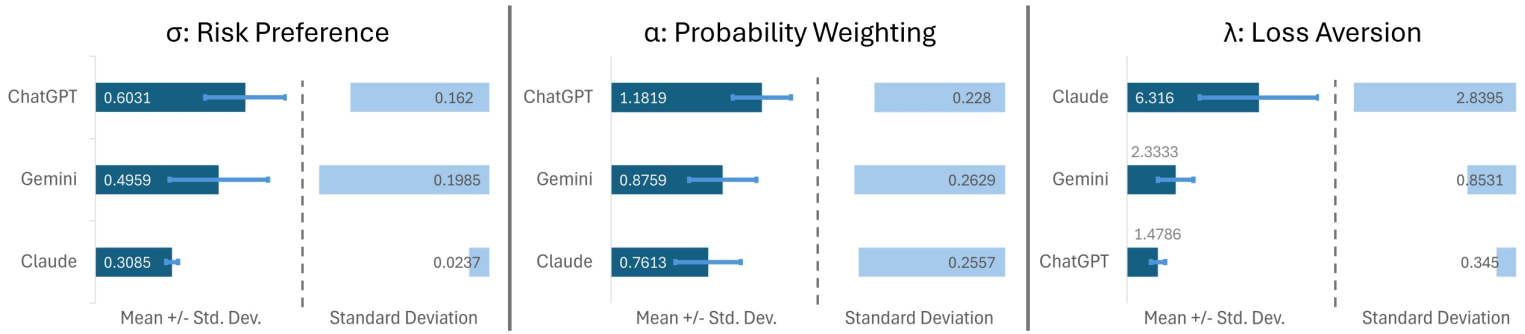


- **Step 1:** Multiple-Choice-List Experiments
- **Step 2:** Recording Switching Points
- **Step 3:** Setting Up Inequalities
- **Step 4:** Estimating Parameters
- **Step 5:** Behavior Evaluation

Key Findings

Basic Context-free Results

Comparison of context-free decision-making:



Summarization of the results:

- Each LLM model shows distinct behavior patterns, as shown in the following table:

Model	Risk Aversion	Loss Concern	Implications
ChatGPT	High	Low	Conservative, safe responses; limited novelty
Claude	Lower	High	More risk-tolerant, cautious with losses
Gemini	Balanced	Balanced	Closer to human-like behavior

Key Findings

Results after Embedded Demographic Features

The Personas across **10 socio-demographic groups** that we explore:

Group	Persona
Panel 1: Foundational Demographic Features	
Sex	male, female
Education Level	below lower secondary, lower secondary, upper secondary, short-cycle tertiary, bachelor, and graduate degrees
Marital Status	never married, married, widowed, divorced
Living Area	rural, urban
Age	15 - 24, 25 - 34, 35 - 44, 45 - 54, 55 - 64, 65+
Panel 2: Advanced Demographic Features	
Sex Orientation	heterosexual, homosexual, bisexual, asexual
Disability	physically-disabled, able-bodied
Race	African, Hispanic, Asian, Caucasian
Religion	Jewish, Christian, Atheist, Religious
Political Affiliation	lifelong Democrat, lifelong Republican, Barack Obama supporter, Donald Trump supporter

- We use **prompting** to embed demographic features: Assign characteristics (e.g., age, gender) to simulate human-like decision-making contexts.

Key Findings

Results after Embedded Demographic Features

Summary:
LLMs display different levels of sensitivity and responses to various demographic features, influencing their decision-making behaviors.

For Example:

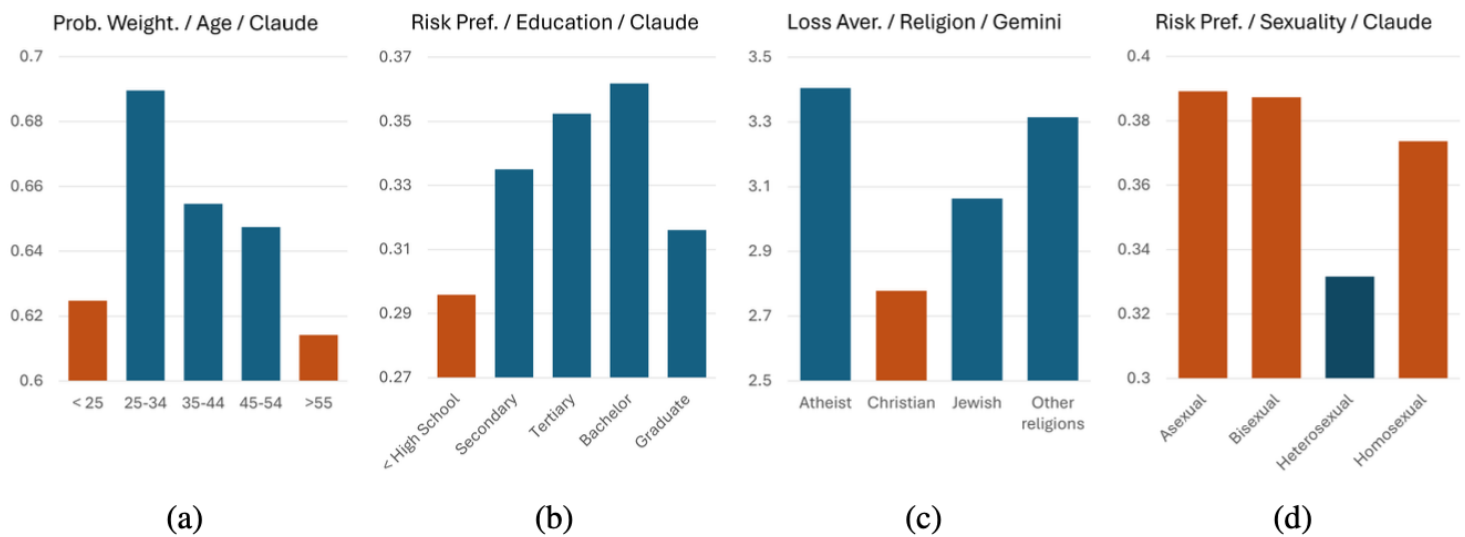


Figure: Example of Divergent Behaviors

Implications and Discussion

Some Open Questions

Fundamental question:

- Should LLMs be neutral knowledge processors or reflective of human-like behaviors?

Balancing ethical responsibility with usability:

- Should LLMs reflect human biases or aim to correct them?

Supplementary evidence in social science research:

- Can LLM outputs help overcome survey bias in social science research, while balancing accuracy and ethical considerations?

The End

Questions? Comments?

Please contact corresponding author:

Jingru Jia (jingruj3@illinois.edu)