Commun Transp Res

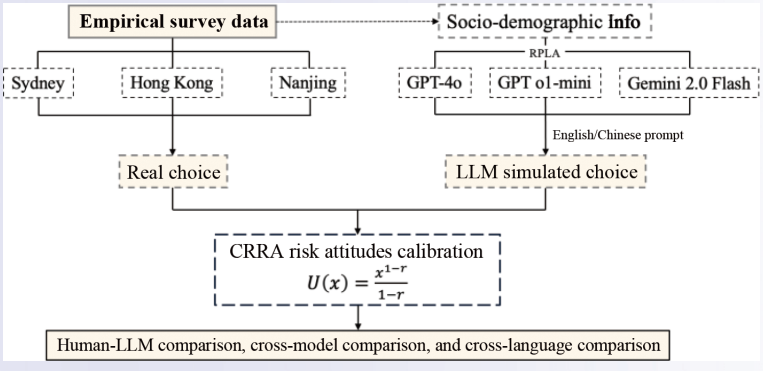
Research Article

Can large language models capture human risk preferences? A cross-cultural study

Jianing Liu[1,†], Bing Song[2,†], Vinayak Dixit[3], Chenyang Wu[4,5,✉], Sisi Jian[2,✉]

Cite this article: Liu J N, Song B, Dixit V, et al. *Commun Transp Res* 2026, **6**(2): 9640025. https://doi.org/10.26599/COMMTR.2026.9640025

ABSTRACT: Large language models (LLMs) are increasingly used as agents to simulate human behavior, yet their fidelity in complex decision-making under uncertainty remains insufficiently understood. To address this gap, we developed a comparative framework that benchmarked LLM-simulated risk preferences against empirical human behavior. Using demographic profiles from surveys conducted in Sydney, Hong Kong, and Nanjing, we constructed role-playing prompts and evaluated three LLMs on abstract lottery-choice tasks. We adopted the classical constant relative risk aversion (CRRA) framework as a domain-neutral "standard ruler" to compare risk attitudes. The analysis yielded three main findings. First, off-the-shelf LLMs do not exhibit a universal risk profile: The two GPT models are more risk-averse than human benchmarks, whereas Gemini is more risk-seeking. Second, prompt language systematically affects simulated risk attitudes, with English-to-Chinese switching inducing a more conservative shift in most cases. Third, LLMs do not reliably reproduce the empirical heterogeneity of human risk preferences, tending either to generate overly concentrated distributions or unrealistically large dispersion. Taken together, these findings show that off-the-shelf LLMs remain vulnerable to model-family-specific miscalibration, language-sensitive distortions, and failures in distributional fidelity. Rigorous empirical calibration is therefore necessary before off-the-shelf LLMs can be reliably deployed in computational social science and choice modeling.



KEYWORDS: large language model (LLM); risk attitude; stated preference survey; lottery choice games

1 Introduction

Large language models (LLMs) have undergone remarkable development in recent years (Luo et al., 2026; Nie et al., 2025a). Compared to earlier natural language processing tools, their capabilities have expanded far beyond traditional language processing tasks, extending to areas such as dialog systems (Qu et al., 2023; Yi et al., 2026), automated content creation (Ohde et al., 2025), and specialized domain applications, including legal, financial, and medical advisory (Chen et al., 2024c; Cheong et al., 2024; Liu et al., 2023). As a result, the trust placed in the outputs of LLMs is under increasing critical scrutiny, particularly regarding the underlying logic of the models and the reliability of the generated results.

As scrutiny of LLM outputs grows, researchers have begun to explore how LLMs perform in simulating human behavior, particularly in survey-based contexts (Liu et al., 2024c; Xu et al., 2025). In addition to traditional survey-style question-and-answer formats, role-playing language agents (RPLAs) have emerged as a promising approach. These agents prompt LLMs to embody specific social roles—such as a doctor, policymaker, or consumer—and respond from the perspective of that role within a given scenario. This method allows for more context-sensitive and socially grounded behavior simulations, enhancing the ecological validity of experiments involving human-like decision-making. Existing studies have examined various human-like tasks, including value judgments (Tjuatja et al., 2024), perceptual analysis (Li et al., 2024a), and intertemporal choices (Goli and Singh, 2024), providing insights into the extent to which LLM

† Jianing Liu and Bing Song contributed equally to this work.

[1] School of Economics and Management, Southwest Jiaotong University, Chengdu 611756, China. [2] Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Hong Kong 999077, China. [3] School of Civil and Environmental Engineering, The University of New South Wales, Sydney 2052, Australia. [4] School of Aeronautics, Northwestern Polytechnical University, Xi'an 710129, China. [5] National Key Laboratory of Aircraft Configuration Design, Xi'an 710129, China.

✉ Corresponding authors. E-mail: cywu@nwpu.edu.cn; cesjian@ust.hk

Received: January 15, 2026; Revised: January 24, 2026; Accepted: May 12, 2026

responses align with human preferences. However, this body of research remains focused on relatively direct tasks, providing limited insight into complex decision-making. Furthermore, the performance of LLMs and the conclusions drawn by researchers have shown significant variability across different tasks. This divergence underscores the urgent need to explore how LLMs operate in more demanding contexts, such as choice situations defined by risk and uncertainty.

Choice under risk and uncertainty involves selecting from alternatives with varying probabilistic outcomes. Unlike deterministic choice, it requires a comprehensive cognitive assessment of potential consequences, making it inherently more complex (Kahneman and Tversky, 1979). Such decisions are widespread and influence outcomes across diverse domains, ranging from personal finance (Eeckhoudt et al., 2011; Engelmann et al., 2009) to public policy (Chen et al., 2025; Greenspan, 2004; Lafont, 2015). Among the many domains influenced, transportation stands out as especially critical and complex. In everyday travel, this appears in the trade-off between a faster route with a higher risk of congestion and a slower but more reliable alternative. The complexity becomes even greater with emerging technologies such as autonomous vehicles (AVs) and urban air mobility (UAM), where passengers must weigh not only time and cost but also technological reliability and physical safety (Song and Jian, 2025a, 2025b). Understanding and simulating these decision-making preferences is therefore essential: It helps optimize today's transportation systems and steer the responsible integration of future systems. Accordingly, in recent years, researchers have begun to explore the use of LLMs to help address transportation problems (Nie et al., 2025b; Ren et al., 2024). However, evaluating LLMs directly in richly contextualized transportation scenarios would introduce substantial semantic confounding. In such settings, it becomes difficult to determine whether an observed choice reflects the model's underlying risk attitude or its associations with specific transportation concepts, attributes, and narratives. We therefore begin with abstract, domain-neutral lottery tasks to establish a controlled behavioral baseline.

Building on this controlled baseline, our empirical design prompts LLMs to act as RPLAs, where each agent's persona is meticulously defined by a real-world user profile. These profiles, sourced from three prior transportation studies in Sydney, Hong Kong, and Nanjing, provide a robust cross-cultural foundation and detail key demographic attributes (age, gender, education, and income). We then supplied these personas to three distinct LLMs (generative pre-trained transformer (GPT)-4o, GPT o1-mini, and Gemini 2.0 Flash) and instructed them to mimic the corresponding respondents by predicting their choices in a series of lottery-based games. Using the well-established constant relative risk aversion (CRRA) framework as a common "standard ruler", we compare simulated and empirical risk attitudes across human benchmarks, model families, and prompt languages. Given the opaque nature of LLM decision-making, adopting this classical framework provides a transparent and low-dimensional benchmark for evaluation. As CRRA relies on a single interpretable parameter, it minimizes additional disturbances introduced by richer model specifications and makes it easier to attribute observed discrepancies to LLMs' behavioral biases.

Our findings indicate that off-the-shelf LLMs do not exhibit a universal human-like risk profile. Across the tested datasets, the two GPT models display systematically greater risk aversion than human respondents, whereas the Gemini 2.0 Flash shows a greater risk-seeking tendency. Within the OpenAI family, GPT o1-mini aligns more closely with the empirical benchmarks than GPT-4o, suggesting relatively higher cognitive fidelity in the simulation of risky choice. Beyond these directional differences, we also find that LLMs do not reliably reproduce the empirical heterogeneity of human risk preferences. Rather than matching the observed dispersion of human responses, the models tend either to generate overly concentrated distributions or to produce substantially wider variation than the human benchmarks. At the same time, prompt language systematically affects simulated risk attitudes. In most tested cases, switching from English to Chinese induces a more conservative shift, although the effect of this shift on simulation accuracy depends on each model family's baseline calibration bias. Taken together, these results suggest that although reasoning-enhanced models may improve alignment with the central tendency of human behavior, off-the-shelf LLMs remain vulnerable to language-sensitive distortions, model-family-specific miscalibration, and failures to capture the distributional fidelity of human decision-making in multilingual and cross-cultural applications.

The remainder of this paper is organized as follows: Section 2 provides a review of the relevant literature. Section 3 outlines the dataset employed in this study. Section 4 presents the modeling framework. Section 5 discusses the empirical findings and their policy implications. Section 6 concludes the study with key findings, discusses the underlying mechanisms driving these behavioral phenomena, and outlines directions for future research.

2 Literature review

2.1 Simulating human behavior with RPLAs

The rapid advancement of LLMs has significantly accelerated the development of RPLAs (Anil et al., 2025; OpenAI et al., 2024). RPLAs enable artificial intelligence to interact with humans in ways that resemble embodied intelligence (Chen et al., 2024c), making these agents increasingly capable of simulating complex social behaviors. Through alignment training, RPLAs can replicate the knowledge systems of specific individuals, imitate their linguistic styles and behavioral tendencies, and reproduce latent personal characteristics (Dai et al., 2023, 2025; Ge et al., 2025). When combined with contextual prompting techniques, RPLAs can be tailored to simulate specific personas or emulate social groups by drawing on internal parameterized knowledge.

Current applications generally fall into two categories: (1) An exploratory direction, which simulates scenarios that are difficult to examine in real-world settings, and (2) a validating direction, which compares RPLA outputs with empirical results to evaluate model fidelity.

The first category addresses domains where conducting real-world studies is challenging due to limitations such as long time frames and ethical concerns. For example, Zhao et al. (2024) proposed CompeteAI, a GPT-4-powered framework to study competitive dynamics in a virtual town, while Li et al. (2024a) developed EconAgent to simulate macroeconomic decisions. In healthcare, frameworks such as AgentClinic (Schmidgall et al., 2024) and Agent Hospital (Li et al., 2024b) use LLM agents to model complex human interactions in controlled, lifelike clinical settings.

Complementing this, the second line of research leverages RPLAs (human benchmark-based behavioral consistency evaluation methods) to systematically assess the alignment between LLMs and established human benchmarks. These studies typically adopt a "simulate-compare-validate" methodology, where model-generated responses are compared against human behavior to evaluate their fidelity and consistency. This approach

not only identifies which LLMs are most suitable for specific tasks but also provides insights into their internal decision-making mechanisms.

Notably, conclusions from this line of research vary significantly across different domains. In structured tasks such as travel planning (Xie et al., 2024) and recommendation systems (Lin et al., 2024), LLMs have shown performance on par with, or even surpassing, human capabilities, producing outputs that are logical, coherent, and closely tailored to user needs. However, in more complex or less structured domains, such as consumer choice trade-offs (Goli and Singh, 2024) and theory of mind (Xu et al., 2024a), LLMs struggle to perform at the same level. Research indicates that while these models can effectively mimic human decision-making through pattern recognition, their outputs often reflect only superficial alignment with human reasoning. They fail to grasp the deeper logic, contextual nuances, or psychological motivations underlying complex human decisions, highlighting significant limitations in their adaptability and understanding.

2.2 Modeling risk preferences under uncertainty in transportation

Uncertainty is an inherent feature of transportation systems, fundamentally shaping how individuals make travel decisions (He et al., 2025). The diverse responses to this uncertainty reveal underlying risk preferences, making their modeling essential for accurate travel demand forecasting and policy evaluation. Two major theoretical streams have been employed to capture these preferences. One stream builds upon expected utility theory (EUT) (Mongin and Baccelli, 2021), often specified using a CRRA utility function (Pratt, 1978). The CRRA model is valued for its mathematical tractability and parsimonious representation of risk aversion through a single coefficient, assuming that an individual's aversion to risk is proportional to their "wealth" (e.g., their available travel time budget). An alternative, influential stream is rooted in behavioral economics, led by cumulative prospect theory (CPT) (Tversky and Kahneman, 1992). CPT provides a more psychologically detailed account, incorporating concepts such as reference dependence, loss aversion, and probability weighting. Both frameworks are prominent in transportation research, offering a trade-off between the elegant simplicity of EUT and the descriptive richness of CPT.

The empirical application of these theoretical frameworks in transportation is extensive, providing clear evidence of risk preferences shaping traveler behavior. Within the EUT paradigm, recent studies continue to refine our understanding of the value of reliability. For example, Carrion and Levinson (2012) analyzed long-term panel data and confirmed that the value of reliability is a stable and significant impact factor in commuter decision-making. Furthermore, a meta-analysis by Shams et al. (2017) systematically reviewed studies and found a robust link between risk attitudes and the valuation of reliability, providing strong, aggregate evidence for the applicability of EUT-based models in quantifying risks. Lu et al. (2025) analyzed metro evacuation data and found that under metro service disruptions, activity urgency and travel distance significantly impact passengers' mode choices, reflecting their risk preferences under time and cost uncertainty. The results validate the applicability of the EUT framework in explaining heterogeneity in passenger decision-making.

In parallel, research grounded in CPT has offered more nuanced behavioral insights. For instance, Gao et al. (2010) demonstrated that CPT better explains how travelers process real-time information, particularly in how they overweight the small probabilities of high-consequence delays (e.g., due to accidents), a behavior EUT struggles to predict. Ghader et al. (2019) applied CPT to the complex Makkah road network and found that it effectively captured drivers' route choice behavior under uncertainty, particularly their asymmetric responses to potential travel time losses. More recently, Liu et al. (2024a) employed CPT in the mobility-as-a-service context to estimate travelers' risk perceptions and their valuations of gains and losses in travel time and cost. Building on this, Liu et al. (2025) integrated CPT with a latent class model to identify discrete heterogeneity in these perceptions.

Despite the maturity of theoretical models for risk preference, the conventional field survey methods used for their parameter estimation face severe practical challenges. First, acquiring sufficient and diverse samples is both time-consuming and prohibitively expensive. Second, even when these hurdles are overcome, data quality remains a concern. For example, stated preference surveys are susceptible to hypothetical bias, creating a gap between respondents' stated and actual choices. The revealed preference survey, on the other hand, suffers from recall errors. Together, these practical and methodological challenges create a significant bottleneck in advancing behavioral research.

2.3 Research gap

A critical research gap exists in the current literature: While LLM-based RPLAs have the potential to address the practical challenges faced by traditional field survey methods in transportation systems—such as the time-consuming and cost-prohibitive process of obtaining sufficiently large and diverse samples—there is limited understanding of their fidelity in replicating human respondents' behaviors, particularly in complex and uncertainty-laden scenarios. Although RPLAs have shown promise in mimicking human decision-making in well-structured tasks, their ability to accurately emulate human responses in highly uncertain or complex contexts remains underexplored.

3 Data

This study utilized three datasets collected through surveys conducted with real respondents in Sydney, Hong Kong, and Nanjing. These three datasets originate from the following studies: Dixit et al. (2019b), Liu et al. (2024a, 2024b), and Guo et al. (2026a, 2026b). All three contain questions on respondents' sociodemographic features and include lottery-choice games designed to elicit respondents' risk attitudes[1]. While each dataset contains rich information about respondents, the specific attributes available vary slightly due to differences in survey design and local priorities. For example, some include detailed employment or household composition variables, while others do not. To ensure consistency and comparability in simulated agents, this study uses the set of sociodemographic variables common across all three datasets, age, gender, education level, and income.

3.1 Descriptive analysis of sociodemographics

A descriptive analysis of the sociodemographic variables shared across the three datasets is presented in Table 1. Substantial differences are observed across the three datasets: On average, the

[1] The authors acknowledge that risky behavior is context-dependent (e.g., the same individual may behave differently when facing uncertainty in finance versus transportation). Given that this study is a preliminary exploration and is constrained by data availability, we do not undertake more complex, context-specific analyses of risk behavior. Instead, we rely on a relatively context-independent, general uncertainty paradigm—the lottery-choice task.

Table 1 Descriptive statistics of sociodemographic variables used in this study across the three datasets

Datasets	Age (SD)	Male (%)	Bachelor or above (%)	Income (SD)
Sydney	28.7 (25.5)	79.70	50	3.8 (2.6)
Hong Kong	40.7 (13.1)	45.30	62.50	3.8 (1.4)
Nanjing	36.8 (9.2)	50.30	85.50	2.5 (1.0)

Note: Due to differences in exchange rates and local purchasing power, the "income" variable is not directly comparable across the three datasets. Instead, the values in this column should be interpreted as indicative of the relative relationship between the mean and standard deviation within each region.

Sydney sample is the youngest, while the Hong Kong sample is the oldest; the Nanjing dataset exhibits the highest average education level.

3.2 Overview of lottery choice games

A lottery game is a controlled experimental task in which participants choose between probabilistic outcomes. By systematically varying the probabilities and magnitudes of these outcomes, each individual's risk attitude—that is, their degree of aversion to or preference for risk—can be inferred. The setup of the lottery choice game in each survey is detailed below. It is worth noting that while the outcome values of the lotteries vary across the three surveys, the primary focus should be on the differences in expected values between the left and right options within each game.

The Sydney dataset consists of a series of nine lottery choice tasks. Each task presents participants with a binary decision: The right option offers a fixed outcome, while the left option involves risk, providing either a higher or lower reward. The probabilities associated with the left options are clearly communicated to participants and vary across tasks. Details of these lottery tasks are presented in Table 2. The terms expected value $(\mathrm{EV})_{\mathrm{Left}}$ and $\mathrm{EV}_{\mathrm{Right}}$ are the expected utility of selecting the left and right lotteries, respectively. The Sydney dataset includes responses from 64 valid participants.

The Hong Kong dataset also consists of nine lottery choice tasks. The probability settings are identical to those in Dixit et al. (2019a); however, the lottery values have been scaled by a factor of 100, as detailed in Table 3. This dataset includes responses from 997 valid participants.

The Nanjing dataset comprises ten lottery choice tasks. Each task presents participants with a binary decision, where both options involve risk. In the left lottery, participants could receive a lower outcome of 1.6 with probability p or a higher outcome of 2 with probability $1 - p$. Similarly, the right lottery offers a lower outcome of 0.1 with probability p or a higher outcome of 3.85 with probability $1 - p$. The value of p varies across tasks, with details provided in Table 4, where p corresponds to the "Prob. payoff 2". The Nanjing dataset includes responses from 145 valid participants.

Table 2 Lottery setup for the Sydney dataset

Lottery	Left lottery			Right lottery	$\mathrm{EV}_{\mathrm{Left}}$	$\mathrm{EV}_{\mathrm{Right}}$
	Low	High	P_{Low}	Fixed		
1	0.5	20	0.5	8	10.25	8
2	0.5	20	0.7	1	6.35	1
3	0.5	20	0.9	4	2.45	4
4	0.5	20	0.5	6	10.25	6
5	0.5	20	0.7	2	6.35	2
6	0.5	20	0.9	2	2.45	2
7	0.5	20	0.5	4	10.25	4
8	0.5	20	0.7	5	6.35	5
9	0.5	20	0.9	1	2.45	1

Table 3 Lottery setup for the Hong Kong dataset

Lottery	Left lottery			Right lottery	$\mathrm{EV}_{\mathrm{Left}}$	$\mathrm{EV}_{\mathrm{Right}}$
	Low	High	P_{Low}	Fixed		
1	50	2000	0.5	800	1025	800
2	50	2000	0.7	100	635	100
3	50	2000	0.9	400	245	400
4	50	2000	0.5	600	1025	600
5	50	2000	0.7	200	635	200
6	50	2000	0.9	200	245	200
7	50	2000	0.5	400	1025	400
8	50	2000	0.7	500	635	500
9	50	2000	0.9	100	245	100

Table 4 Lottery setup for the Nanjing dataset

Lottery	Prob. payoff 1	Prob. payoff 2	Left lottery Payoff 1	Left lottery Payoff 2	Right lottery Payoff 1	Right lottery Payoff 2	$\mathrm{EV}_{\mathrm{Left}}$	$\mathrm{EV}_{\mathrm{Right}}$
1	0.1	0.9	2	1.6	3.85	0.1	1.64	0.475
2	0.2	0.8	2	1.6	3.85	0.1	1.68	0.85
3	0.3	0.7	2	1.6	3.85	0.1	1.72	1.225
4	0.4	0.6	2	1.6	3.85	0.1	1.76	1.6
5	0.5	0.5	2	1.6	3.85	0.1	1.8	1.975
6	0.6	0.4	2	1.6	3.85	0.1	1.84	2.35
7	0.7	0.3	2	1.6	3.85	0.1	1.88	2.725
8	0.8	0.2	2	1.6	3.85	0.1	1.92	3.1
9	0.9	0.1	2	1.6	3.85	0.1	1.96	3.475
10	1	0	2	1.6	3.85	0.1	2	3.85

4 Methods

To compare human behavior with that of LLMs under uncertainty, we used a role-playing language-agents protocol in which the models simulated our participants. We presented the same uncertainty scenarios to the models and required them to respond as if they were the corresponding human respondents. An overview of the study design and comparative framework is provided in Fig. 1.

We evaluated three publicly accessible LLMs—Gemini 2.0 Flash, GPT-4o, and GPT o1-mini—for the following reasons: (1) At the time of the project (December 2024), these three models were among the more advanced models available to the public; and (2) given the rapid pace of model releases, our objective was not to identify a single model that perfectly simulates human behavior on a specific type of task but to test for systematic differences between current LLMs and human performance. Accordingly, we did not include a broader set of models in formal simulation. Other models that were evaluated during the testing phase but not formally adopted included ChatGPT-3.5 and o1-preview.

4.1 RPLAs architecture and construction

This subsection introduces the construction and structure of the RPLA used in this study. The RPLA is designed to simulate human respondents by assigning different roles and background profiles to an LLM through carefully crafted prompts (Shanahan et al., 2023). The RPLA consists of four key components: Profile, memory, planning, and action (Chen et al., 2024b), as detailed below.

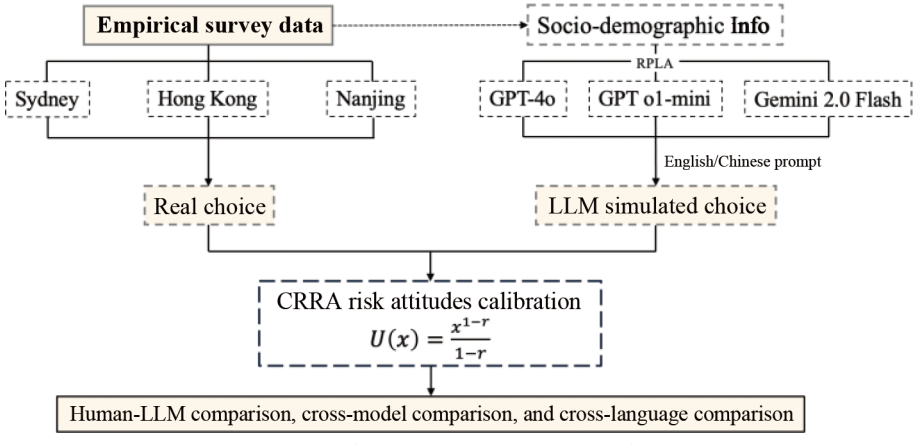


Fig. 1 Overview of the study design and comparative framework.

Profile: A profile defines the unique characteristics of a simulated individual, serving as the foundation for modeling realistic agent behaviors in RPLA systems (Chen et al., 2024a). It typically captures several elements, including the agent's inherent attributes (such as age, gender, or socioeconomic background), task-related specifications (such as role identity, goal orientation, or decision criteria), and external constraints (such as cultural norms, environmental factors, or task-specific limitations). A well-constructed profile enables the agent to behave in a consistent and plausible manner within the simulated environment (Takagi et al., 2025). In this study, we use the four attributes introduced in Section 3.1 (age, gender, education level, and personal income) for profile construction. Additionally, we treated the city of residence as a crucial contextual attribute, as it reflects broader sociocultural and geographical environments that influence agent behavior and language use.

The concept of a profile involves two key dimensions: Form and construction (Xu et al., 2024b). The form dimension refers to the structure and presentation of profile information—ranging from structured data formats (e.g., JavaScript object notation (JSON) schemas or attribute tables) to natural language descriptions or interactive prompts. The construction dimension, on the other hand, concerns how the information in the profile is obtained. This may involve manual creation by researchers, automatic generation via statistical or machine learning models, or derivation from empirical sources such as ethnographies or behavioral datasets.

To determine the optimal content and structure for our prompts, we conducted a preliminary pilot study using a small, purposively selected subset of the Hong Kong dataset. This iterative testing process had two primary objectives: (1) Identify the most salient respondent attributes to include, confirming that the combination of age, gender, education, and income was sufficient to elicit differentiated and plausible behaviors; (2) refine the narrative framing—specifically, how to best integrate the survey's background context with the agent's persona for the LLM. The insights from this pilot phase directly informed the construction of the final prompt template used across all simulations. A summary of the prompt iterations and the rationale for their refinement is provided below in Table 5.

In this study, the form dimension is implemented as interactive prompts that generate natural language profile descriptions by combining demographic and contextual information into coherent, human-readable inputs. The construction dimension is implemented by extracting, standardizing, and aligning selected attributes from the three datasets to build a unified input framework for simulation. A sample English-language prompt is provided below to illustrate the final profile form used in our study:

You will take on the role of a survey respondent. The purpose of this survey is to assess people's risk attitudes through nine consecutive lottery questions. For each question, you are required to carefully consider and compare the expected income with the potential risks; however, you do not need to disclose your reasoning; simply provide your answers. In this survey, you will portray a {age}-year-old {gender} from {city}, {country} who has completed the highest level of education: {education}. Your monthly income is: {income}.

Note that this prompt explicitly instructs the agent to focus on the decision criteria of comparing expected income with potential risks, rather than exploring broader, unconstrained decision heuristics.

In addition to generating profiles in English, we also created profiles in Chinese to better align with the native language of respondents from Hong Kong and Nanjing, China. Prior research suggests that language differences can influence both the reasoning processes of RPLAs and experimental outcomes (Goli and Singh, 2024). Therefore, we aim to investigate whether the language used in profile construction affects the RPLA's risk preferences. The sample prompt in Chinese was translated from the English version[2].

Memory: The second key component is memory. A well-known limitation of LLMs is their restricted context window, which constrains the amount of prior information they can retain during interactions. To maintain the consistency and coherence of RPLA behavior, it is therefore essential to design a memory mechanism capable of storing both received input and generated output from the agent (Alizadeh et al., 2024).

Memory involves two aspects: Memory types and memory operations. Memory can be broadly divided into short-term memory and long-term memory, with the former focusing on immediate, session-specific information, and the latter concerning persistent knowledge across sessions. Memory operations refer to the agents' continuous updating and use of their memory, including writing, retrieval, and reflection.

In this study, short-term memory was implemented by including the entire session's interaction history (comprising the initial profile prompt, all preceding lottery questions, and the

[2] Following consultation with a third-party language quality assessment agency, the Chinese translation was found to align well with the English original and to introduce no ambiguity, omissions, or other translation-induced comprehension problems.

Table 5 Iterative refinement of RPLA prompts

Prompt iteration	Key problem identified	Conclusion and refinement
Initial attempt "Please act as a survey respondent who is a female, aged 19–34, and currently residing in Hong Kong. I will present you with two options, and you must choose either "Option 1" or "Option 2" to maximize your benefits."	The persona lacked critical economic attributes (e.g., income and education). The instruction to "maximize your benefits" was ambiguous, leading to ungrounded and generic agent responses.	**Conclusion**: Personas without economic context cannot effectively simulate realistic decision-making. **Refinement**: Enrich the persona with key economic data and provide a clearer cognitive instruction (e.g., "weigh the risks and rewards").
Over-specification "You are a survey respondent who identifies as female, aged between 19 and 34 years (with an equal probability of being in the age group 19–24 or 25–34), currently residing in Hong Kong...."	Defining an attribute as a probability distribution was unnatural for a single agent's persona. This added unnecessary complexity and cognitive "noise," interfering with the core decision-making process.	**Conclusion**: Simplicity and relevance are key. A persona should represent a single, concrete individual, not a probabilistic ensemble. **Refinement**: Focus the persona on specific, salient attributes directly relevant to economic choices.
Integration of key data "You are a survey respondent who identifies as female, aged 25–34 without a driver's license, employed full-time, and holding a bachelor's degree. Your household consists of two members, your monthly personal income is 25,000 Hong Kong dollar (HKD), and your total household monthly income is 35,000 HKD....You are required to evaluate the associated risks and rewards of each option and select the one you consider most advantageous.... Option 1: You will have 50% chance of winning 50 HKD and the remaining 50% chance of winning 2000 HKD. Option 2: You will win 800 HKD for sure."	This version successfully integrated key economic data, resulting in plausible, context-aware reasoning. However, a fundamental structural flaw remained, as it combined the persona with the task, making it not scalable for multiquestion surveys where the persona need to persist.	**Conclusion**: A detailed persona with core economic variables is essential for effective simulation. **Refinement**: Address the final, structural problem of prompt design to enable sequential questioning.
Final solution		Decouple the **Persona prompt** (a one-time setup) from the **Task prompts** (the sequential questions). This allows the agent to maintain a consistent identity throughout the entire survey.

agent's previous choices) into the input for each new decision. This contextual persistence ensures that the agent maintains a consistent persona and remains aware of its past choices throughout the nine-question sequence. Long-term memory across sessions, however, is intentionally excluded. This approach aligns with the natural thought process of real human respondents in real-world scenarios, as they have memory of their previous options in the survey but lack access to others' experiences and are required to complete the survey independently.

Planning: Planning refers to the process by which an agent formulates a sequence of actions to achieve specific objectives. In this study, we adopt empathetic planning, which leverages prompt-guided chain-of-thought (CoT) reasoning to enable the agent to adjust its decisions based on character-related information (Mou et al., 2026). This approach enhances the agent's ability to anticipate and infer the behaviors and emotions of simulated human respondents before making decisions.

We aim for the RPLA to simulate the decision-making processes of respondents when faced with sequential lottery-related questions. Specifically, we expect the agent to construct its internal reasoning chain by integrating information from its individual profile, local average income, its financial status, and the expected returns of different lottery choices. This ensures that the RPLA's responses align more closely with the cognitive processes observed in human participants in real-world settings. This process is implemented by instructing the agent to internally deliberate before generating a constrained output, as detailed in the Action section below.

Action: Action refers to the direct interaction between the RPLA and the real world. As an interface for simulating human behavior, action enables the RPLA to effectively execute tasks and generate responses. We employ a closed-domain approach and regulate the RPLA's actions through prompt-based instructions. Specifically, when responding to lottery-related questions, the RPLA is required to provide only its final choice without revealing the underlying reasoning process. This design serves three primary purposes. First, it minimizes the influence of linguistic variability and noise that often emerge during open-ended reasoning generation. Since LLMs are highly sensitive to prompt phrasing and contextual cues, exposing step-by-step thought processes may introduce inconsistencies unrelated to the core decision logic (Xu et al., 2022). Second, by limiting the output to a discrete choice, we standardize the response format, which facilitates reliable and efficient modeling of risk preferences across different scenarios and agents. Third, it mirrors real-world surveys, where researchers can only observe respondents' final choices, not their reasoning processes. A complete example of a prompt combining the planning and action instructions is provided below:

Based on your profile, first think step-by-step about the expected values and your financial situation. After your internal deliberation, you do not need to explain your reasoning process—simply indicate whether you choose "Option 1" or "Option 2".

4.2 Risk attitude estimation

A decision maker's risk attitude refers to their consistent pattern of ranking and choosing among uncertain outcomes, reflecting their willingness to accept variability in results (Pratt, 1978). Specifically, when presented with two alternatives of equal expected value (e.g., a guaranteed gain of $50 versus a 50% chance of gaining $100), an individual who prefers the former is considered risk averse (avoiding potential risks), while one who prefers the latter is risk seeking (pursuing higher potential returns despite the risks). If the individual is indifferent between the two, they are considered risk neutral. In the expected utility framework, risk attitudes are represented by the curvature of an individual's utility function over wealth: Concave utility indicates risk aversion, linear utility indicates risk neutrality, and convex utility indicates risk seeking.

In this study, we employ the CRRA model (Dixit et al., 2019a) to quantify respondents' risk preferences. The utility function in the CRRA model is defined as

$$U(x) = \frac{x^{1-r}}{1-r} \quad (1)$$

where x is the reward, and r is the risk preference ($r > 0$: Risk-averse; $r = 0$: Risk-neutral; $r < 0$: Risk-seeking).

5 Empirical analysis

This section first compares risk attitudes estimated from the real and simulated data and then examines the impact of using different prompt languages. To mitigate the impact of randomness in LLM-generated outputs, each lottery choice task faced by each respondent in each dataset was simulated three times. The majority output—i.e., the option selected in at least two out of three simulations—was adopted as the final simulated response.

5.1 Comparison of real and simulated risk attitudes

Table 6 summarizes the risk attitudes estimated from real and LLM-simulated data. Clear and systematic differences emerge across the three model families. Across all three datasets, the risk attitudes estimated from GPT-4o-simulated data are consistently higher than the corresponding real risk attitudes, indicating that GPT-4o tends to produce more risk-averse choices than human respondents. The estimates derived from GPT o1-mini-simulated data are also higher than the real benchmarks, but they are notably closer to the empirical values than those generated by GPT-4o, suggesting improved fidelity in the simulation of risky choice. In contrast, the Gemini 2.0 Flash-simulated estimates are consistently lower than the real risk attitudes across all three datasets, indicating a systematic tendency toward more risk-seeking behavior relative to human respondents. Paired t tests further confirm that the differences between real and simulated risk attitudes are statistically significant at the 1% level across all datasets and model families. Taken together, these results show that LLM-based simulations do not deviate from human behavior in a uniform direction.

Table 6 Estimation of risk attitudes based on real and LLM-simulated data

Datasets	Real	GPT-4o	GPT o1-mini	Gemini 2.0 Flash
Sydney	0.420	1.387	0.641	0.234
Hong Kong	0.765	1.620	0.509	0.157
Nanjing	0.130	0.357	0.229	−0.277

To provide a more intuitive view of these directional biases and to further examine the heterogeneity of simulated risk preferences, Fig. 2 presents grouped boxplots of the estimated CRRA coefficients across the three datasets. The figure complements Table 6 by illustrating not only the directional differences across model families but also their distributional characteristics.

Beyond the median patterns already reported in Table 6, the boxplots reveal substantial differences in the extent to which the models reproduce the empirical dispersion of human risk attitudes. The human benchmarks (gray boxes) exhibit an observed spread that reflects considerable heterogeneity across respondents. Relative to this benchmark, Gemini 2.0 Flash appears markedly underdispersed in some settings, most notably in the Sydney dataset, where the interquartile range is highly compressed and the whiskers are short. This pattern suggests that the model tends to produce overly concentrated risk-attitude estimates and therefore fails to capture the observed variation in human responses.

In contrast, GPT o1-mini shows the opposite tendency in several datasets, particularly Sydney and Hong Kong, where the boxes and whiskers are substantially wider than those of the human benchmarks. This indicates that although GPT o1-mini often produces medians that are closer to the empirical values, it does not consistently reproduce the empirical spread of risk attitudes. In other words, improved alignment in central tendency does not necessarily imply improved alignment in distributional shape or variance.

Overall, these patterns suggest that off-the-shelf LLMs struggle to match both the level and the dispersion of human risk preferences. Their limitations therefore extend beyond simply overestimating or underestimating risk aversion and include difficulty in reproducing the heterogeneity observed in real respondent populations.

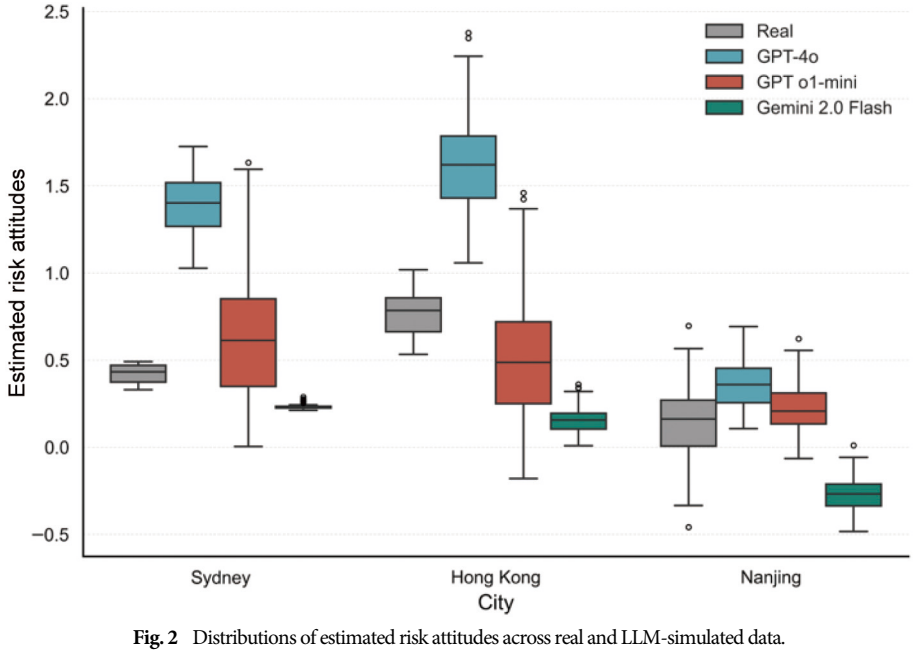


Fig. 2 Distributions of estimated risk attitudes across real and LLM-simulated data.

5.2 Comparison of simulated risk attitudes across different language prompts

Given that the Nanjing and Hong Kong datasets were collected in non-English languages, we further examine whether prompt language affects the simulated risk attitudes from different LLMs. Specifically, we compare the CRRA estimates obtained under English and Chinese prompts for GPT-4o, GPT o1-mini, and Gemini 2.0 Flash. The results are reported in Table 7.

Table 7 Estimation of risk attitudes using different language prompts

Model	City	Real	English	Chinese
GPT-4o	Hong Kong	0.765	1.620	1.654
	Nanjing	0.130	0.357	0.951
GPT o1-mini	Hong Kong	0.765	0.509	0.390
	Nanjing	0.130	0.229	0.250
Gemini 2.0 Flash	Hong Kong	0.765	0.157	0.314
	Nanjing	0.130	−0.277	−0.027

Table 7 reveals that linguistic framing has a substantial and systematic impact on simulated risk attitudes. A notable pattern emerges: In five of the six cases, switching the prompt language from English to Chinese yields a higher estimated CRRA value. This indicates that Chinese prompts induce a conservatism shift, driving the LLMs toward greater risk aversion regardless of their underlying architecture.

The effect of this language-induced shift on simulation accuracy is determined by the model's baseline bias. For GPT-4o and GPT o1-mini, estimates generated under English prompts are already more risk averse than the empirical benchmarks. As a result, the additional conservative shift induced by Chinese prompts further amplifies this inherent bias, pushing the simulated values even farther from the observed data. In contrast, Gemini 2.0 Flash exhibits a systematic risk-seeking tendency under English prompts. In this case, the conservatism induced by Chinese prompts paradoxically counteracts the model's original bias, bringing the estimates closer to the empirical benchmarks. This apparent improvement, however, reflects a mathematical coincidence rather than any genuine cross-cultural competence.

Overall, these findings show that prompt language does not simply enhance or degrade simulation fidelity. Instead, it perturbs the internal calibration of risky choice in a predominantly conservative direction, while the final degree of alignment with human behavior depends on the model's baseline bias. The mechanisms underlying this pattern and its implications for the use of off-the-shelf LLMs are discussed in the next section.

6 Discussion and conclusions

Although LLMs have made significant advancements in recent years, their ability to simulate complex decision-making behavior, such as risky decision-making, remains unvalidated. This study explored the capacity of LLMs, specifically GPT-4o, GPT o1-mini, and Gemini 2.0 Flash, to replicate human risk-related decision-making behavior under uncertainty using lottery-choice scenarios across three geographically and socioeconomically distinct datasets. Respondents' risk attitudes were estimated using the classical CRRA model, functioning as a "standard ruler" to benchmark both real survey data and simulated responses generated by the LLMs.

6.1 Key results

The results show that off-the-shelf LLMs do not share a universal risk profile. Instead, different model families exhibit distinct baseline calibration biases in risky choice. The two ChatGPT models systematically generate more risk-averse choices than human respondents, whereas the Gemini 2.0 Flash tends to be more risk-seeking. This divergence suggests that behavioral fidelity is shaped by model-family-specific training, posttraining, and alignment processes. Within the GPT family, GPT o1-mini nevertheless produces estimates that are consistently closer to the empirical benchmarks than GPT-4o. We interpret this relative improvement as evidence of higher cognitive fidelity: As a reasoning-oriented model, GPT o1-mini appears better able to approximate the probabilistic trade-offs required in lottery-choice tasks, thereby reducing reliance on shallow heuristics. However, stronger reasoning does not eliminate the baseline biases embedded in the models.

Building on the empirical finding that Chinese prompting predominantly induces a conservative shift in estimated risk attitudes, we argue that this language-dependent pattern reflects structural mechanisms rather than simple semantic noise. We propose three underlying drivers for this phenomenon. First, one issue is the Anglocentric bias in multilingual model representations (AlKhamissi et al., 2024; Guo et al., 2025). As English heavily dominates most pretraining corpora, models develop a more robust and finely calibrated baseline for probabilistic trade-off reasoning in English. Second, another mechanism concerns cross-lingual mapping within the model's latent space, which may not operate neutrally. Prompting in Chinese may activate specific cultural priors embedded within Chinese-language training distributions. If Chinese texts historically frame risky or probabilistic choices more cautiously, the model will inherently adopt this conservative stance when generating Chinese responses, shifting away from its English-calibrated baseline. Third, this baseline conservatism is likely amplified by language-imbalanced safety fine-tuning. The human feedback data used in reinforcement learning from human feedback (RLHF) pipelines are unevenly distributed across languages (Askell et al., 2021; Verma and Bharadwaj, 2025). This imbalance can cause models to overcompensate when processing non-English languages, relying on coarser and more cautious heuristics to avoid generating unsafe content. This effectively imposes an "alignment tax" that reduces behavioral consistency across languages (Ouyang et al., 2022).

Under this interpretation, whether the conservative shift happens to improve accuracy, as in the case of Gemini, or worsen it, as in the GPT models, is secondary. In both cases, it reveals a fundamental vulnerability in cross-cultural simulation: Off-the-shelf LLMs do not preserve behavioral calibration consistently across languages.

6.2 Implications

Our findings highlight an important limitation of using off-the-shelf LLMs as tools for behavioral prediction. Since different model families exhibit different baseline calibration biases, the choice of model can materially affect the inferred pattern of public risk preferences. In practice, one model family may overstate conservatism, whereas another may overstate willingness to accept risk. Without empirical calibration, such biases can distort inference and lead researchers to draw policy conclusions that do not accurately reflect observed human behavior.

This limitation is particularly consequential in transportation research, where risk perception is central to decision-making under uncertainty. Travel behavior frequently involves probabilistic trade-offs, including route choice under unreliable travel times, mode switching during service disruptions, and the

adoption of emerging mobility technologies under safety and performance uncertainty. If the underlying risk preference parameters are systematically miscalibrated, demand forecasts, welfare evaluation, and policy design may all be biased. For example, using uncalibrated LLM-generated data to infer willingness to adopt safety-critical systems such as AVs or low-altitude mobility services could yield either overly conservative or overly optimistic projections, depending on the model family used.

The multilingual results add a further layer of caution. In linguistically diverse settings, prompt language is not a neutral implementation choice: It can systematically perturb the behavioral calibration of the model. This is especially relevant for transportation systems serving multilingual populations, where researchers may be tempted to use native-language prompting as a straightforward way to improve realism. Our results suggest that such an assumption is unwarranted unless the model has first been validated against human benchmarks in the relevant linguistic context.

Overall, the implications of this study are methodological as much as substantive. Off-the-shelf LLMs should not be treated as direct substitutes for human respondents in risk-sensitive behavioral applications. Instead, they should be regarded as tools whose outputs require domain-specific and language-sensitive calibration. Rigorous validation against human ground truth remains a necessary prerequisite for deploying these models in transportation and other cross-cultural social science settings.

6.3 Future research directions

Our findings point to several promising directions for future research. First, this study examined risk attitudes using a context-free approach via lottery choice games. Future work should evaluate the predictive performance of LLMs in a broader range of decision-making scenarios under uncertainty, as well as other types of complex decision-making processes, to enable broader validation of these models. Second, exploring individual-level heterogeneity and implementing personalized calibration of LLM outputs based on respondents' sociodemographic and psychological characteristics may improve the accuracy of simulated decision-making and enhance the practical utility of these models. In conclusion, this study represents an initial attempt to examine how LLMs simulate human behavior under uncertainty. While LLMs show promise as tools for behavioral modeling, there is still a considerable distance to go before they can fully and reliably replicate the complexity of human decision-making.

Author contributions

Jianing Liu: Conceptualization, Data acquisition, Investigation, Methodology, Writing – original draft; **Bing Song**: Conceptualization, Investigation, Methodology, Writing – original draft; **Vinayak Dixit**: Data acquisition, Writing – review & editing; **Chenyang Wu**: Conceptualization, Data acquisition, Investigation, Methodology, Funding acquisition, Writing – review & editing; **Sisi Jian**: Conceptualization, Data acquisition, Investigation, Methodology, Supervision, Funding acquisition, Writing – review & editing.

Replication and data sharing

The data and codes used in this study are available at https://doi.org/10.26599/ETSD.2026.9190012. Due to privacy constraints, we have provided anonymized samples with randomly shuffled data distributions to illustrate the data structure used in our experiments.

Acknowledgements

This study is supported by the National Natural Science Foundation of China (Nos. 52472330 and 52102389), the Hong Kong Research Grants Council-General Research Fund (No. 16201423), the Hong Kong Research Grants Council-Early Career Scheme (No. 26205921), the Northwestern Polytechnical University Start-up funding (No. D5000230159), the Young Elite Scientists Sponsorship Program by Chinese Association for Science and Technology (No. YESS20220482), the Research Fund for Young Start of Science and Technology in Shaanxi Province (No. 2024ZC-KYXX-035), and the Young Talent Support Program of Shaanxi Province University (No. D5113240045). This research project (No. 2024.A7.032.24B) is also funded by the Public Policy Research Funding Scheme of The Government of the Hong Kong Special Administrative Region.

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

References

Alizadeh, K., Mirzadeh, S. I., Belenko, D., Khatamifard, S., Cho, M., Del Mundo, C. C., et al., 2024. LLM in a flash: Efficient large language model inference with limited memory. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, 12562–12584.

AlKhamissi, B., ElNokrashy, M., AlKhamissi, M., Diab, M., 2024. Investigating cultural alignment of large language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, 12404–12422.

Anil, R., Borgeaud, S., Alayrac, J. B., Yu, J., Soricut, R., et al., 2025. Gemini: A Family of Highly Capable Multimodal Models. https://arxiv.org/abs/2312.11805

Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., et al., 2021. A General Language Assistant as a Laboratory for Alignment. https://arxiv.org/abs/2112.00861

Carrion, C., Levinson, D., 2012. Value of travel time reliability: A review of current evidence. *Transp Res Part A Policy Pract*, **46**, 720–741.

Chen, H., Chen, H., Yan, M., Xu, W., Xing, G., Shen, W., et al., 2024a. SocialBench: Sociality evaluation of role-playing conversational agents. In: Findings of the Association for Computational Linguistics ACL 2024, 2108–2126.

Chen, J., Wang, X., Xu, R., Yuan, S., Zhang, Y., Shi, W., et al., 2024b. From Persona to Personalization: A Survey on Role-Playing Language Agents. https://arxiv.org/abs/2404.18231

Chen, L., Dong, T., Li, X., Xu, X., 2025. Logistics engineering management in the platform supply chain: An overview from logistics service strategy selection perspective. *Engineering*, **47**, 236–249.

Chen, S., Guevara, M., Moningi, S., Hoebers, F., Elhalawani, H., Kann, B. H., et al., 2024c. The effect of using a large language model to respond to patient messages. *Lancet Digit Health*, **6**, e379–e381.

Cheong, I., Xia, K., Feng, K. J. K., Chen, Q. Z., Zhang, A. X., 2024. (A)I Am not a lawyer, but...: Engaging legal experts towards responsible LLM policies for legal advice. In: The 2024 ACM Conference on Fairness, Accountability, and Transparency, 2454–2469.

Dai, S., Shao, N., Zhao, H., Yu, W., Si, Z., Xu, C., et al., 2023. Uncovering ChatGPT's capabilities in recommender systems. In: Proceedings of the 17th ACM Conference on Recommender Systems, 1126–1132.

Dai, Y., Hu, H., Wang, L., Jin, S., Chen, X., Lu, Z., 2025. Mmrole: A Comprehensive Framework for Developing and Evaluating Multimodal Role-playing Agents. https://arxiv.org/abs/2408.04203

Dixit, V., Jian, S., Hassan, A., Robson, E., 2019a. Eliciting perceptions of travel time risk and exploring its impact on value of time. *Transp Policy*, **82**, 36–45.

Dixit, V., Xiong, Z., Jian, S., Saxena, N., 2019b. Risk of automated driving: Implications on safety acceptability and productivity. *Accid Anal Prev*, **125**, 257–266.

Eeckhoudt, L., Gollier, C., Schlesinger, H., 2011. *Economic and Financial Decisions under Risk*. Princeton NJ, USA: Princeton University Press, 55–89.

Engelmann, J. B., Capra, C. M., Noussair, C., Berns, G. S., 2009. Expert financial advice neurobiologically "offloads" financial decision-making under risk. *PLoS One*, **4**, e4957.

Gao, S., Frejinger, E., Ben-Akiva, M., 2010. Adaptive route choices in risky traffic networks: A prospect theory approach. *Transp Res Part C Emerg Technol*, **18**, 727–740.

Ge, T., Chan, X., Wang, X., Yu, D., Mi, H., Yu, D., 2025. Scaling Synthetic Data Creation with 1,000,000,000 Personas. https://arxiv.org/abs/2406.20094

Ghader, S., Darzi, A., Zhang, L., 2019. Modeling effects of travel time reliability on mode choice using cumulative prospect theory. *Transp Res Part C Emerg Technol*, **108**, 245–254.

Goli, A., Singh, A., 2024. Frontiers: Can large language models capture human preferences? *Mark Sci*, **43**, 709–722.

Greenspan, A., 2004. Risk and uncertainty in monetary policy. *Am Econ Rev*, **94**, 33–40.

Guo, M., Liu, J., Jian, S., Jang, S., Wu, C., 2026a. Travel dynamics under MaaS bundles: The role of uncertainty in intertemporal choices and learning behavior. *Transp Res Part A Policy Pract*, **208**, 104954.

Guo, M., Liu, J., Jian, S., Li, Z., Ren, G., Wu, C., 2026b. How virtual experience reshapes commuters' MaaS subscription and mode choice: Insights from an economic experiment. *Travel Behav Soc*, **43**, 101208.

Guo, Y., Conia, S., Zhou, Z., Li, M., Potdar, S., Xiao, H., 2025. Do large language models have an English accent? Evaluating and improving the naturalness of multilingual LLMs. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, 3823–3838.

He, Z., Yang, Y., Mu, Y., Qu, X., 2025. Public acceptance of driverless buses: An extended UTAUT2 model with anthropomorphic perception and empathy. *Commun Transp Res*, **5**, 100167.

Kahneman, D., Tversky, A., 1979. Prospect theory: An analysis of decision under risk. *Econometrica*, **47**, 263.

Lafont, C., 2015. Deliberation, participation, and democratic legitimacy: Should deliberative mini-publics shape public policy? *J Polit Philos*, **23**, 40–63.

Li, P., Castelo, N., Katona, Z., Sarvary, M., 2024a. Frontiers: Determining the validity of large language models for automated perceptual analysis. *Mark Sci*, **43**, 254–266.

Li, Y., Huang, Y., Wang, H., Zhang, X., Zou, J., Sun, L., 2024b. Quantifying AI Psychology: A Psychometrics Benchmark for Large Language Models. https://arxiv.org/html/2406.17675v1

Lin, X., Wang, W., Li, Y., Yang, S., Feng, F., Wei, Y., et al., 2024. Data-efficient fine-tuning for LLM-based recommendation. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, 365–374.

Liu, J., Jian, S., Wu, C., Dixit, V., 2024a. Risky choice and diminishing sensitivity in MaaS context: A nonlinear logit analysis of traveler behavior. *Transp Res Part C Emerg Technol*, **162**, 104603.

Liu, J., Wen, X., Jian, S., 2024b. Toward better equity: Analyzing travel patterns through a neural network approach in mobility-as-a-service. *Transp Policy*, **153**, 110–126.

Liu, J., Wu, C., Jian, S., 2025. Heterogenous responses to risky behavior involving monetary and non-monetary variables in MaaS context: A latent class choice approach. *Travel Behav Soc*, **41**, 101097.

Liu, T., Li, M., Yin, Y., 2024c. Can Large Language Models Capture Human Travel Behavior? Evidence and Insights on Mode Choice. https://ssrn.com/abstract=4937575

Liu, Y., Wu, F., Liu, Z., Wang, K., Wang, F., Qu, X., 2023. Can language models be used for real-world urban-delivery route optimization? *Innovation*, **4**, 100520.

Lu, Q. C., Zuo, X. Y., Chen, C., Dong, Z., Xu, P. C., Li, J., 2025. The heterogeneity of travel mode choice behavior under unplanned metro service disruptions. *Transp Res Part D Transp Environ*, **142**, 104683.

Luo, S., He, S. Y., Song, L., Jian, S., Yao, Y., 2026. Exploring the potential of large language models (LLMs) in analyzing passengers' perceptions of transit service quality. *Environ Plan B Urban Anal City Sci*, **53**, 90–106.

Mongin, P., Baccelli, J., 2021. Expected utility theory, Jeffrey's decision theory, and the paradoxes. *Synthese*, **199**, 695–713.

Mou, X., Ding, X., He, Q., Wang, L., Liang, J., Zhang, X., et al., 2026. From individual to society: A survey on social simulation driven by large language model-based agents. *ACM Comput Surv*, **58**, 1–41.

Nie, T., He, J., Mei, Y., Qin, G., Li, G., Sun, J., et al., 2025a. Joint estimation and prediction of city-wide delivery demand: A large language model empowered graph-based learning approach. *Transp Res Part E Logist Transp Rev*, **197**, 104075.

Nie, T., Sun, J., Ma, W., 2025b. Exploring the roles of large language models in reshaping transportation systems: A survey, framework, and roadmap. *Artif Intell Transp*, **1**, 100003.

Ohde, J. W., Rost, L. M., Overgaard, J. D., 2025. The burden of reviewing LLM-generated content. *NEJM AI*, **2**, 1–3.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., et al., 2024. Gpt-4 Technical Report. https://arxiv.org/abs/2303.08774

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., et al., 2022. Training language models to follow instructions with human feedback. *Adv Neural Inf Process Syst*, **35**, 27730–27744.

Pratt, J. W., 1978. Risk aversion in the small and in the large. In: Uncertainty in Economics, 59–79.

Qu, X., Lin, H., Liu, Y., 2023. Envisioning the future of transportation: Inspiration of ChatGPT and large models. *Commun Transp Res*, **3**, 100103.

Ren, Y., Chen, Y., Liu, S., Wang, B., Yu, H., Cui, Z., 2024. TPLLM: A traffic prediction framework based on pretrained large language models. *Applied Soft Computing*, **184**, 113840.

Schmidgall, S., Ziaei, R., Harris, C., Reis, E., Jopling, J., Moor, M., 2024. Agentclinic: A Multimodal Agent Benchmark to Evaluate AI in Simulated Clinical Environments. https://arxiv.org/html/2405.07960v1

Shams, K., Asgari, H., Jin, X., 2017. Valuation of travel time reliability in freight transportation: A review and meta-analysis of stated preference studies. *Transp Res Part A Policy Pract*, **102**, 228–243.

Shanahan, M., McDonell, K., Reynolds, L., 2023. Role play with large language models. *Nature*, **623**, 493–498.

Song, B., Jian, S., 2025a. Balancing privacy and revenue: A differentially private dynamic pricing algorithm for ride-hailing. *Transp Res Part E Logist Transp Rev*, **203**, 104310.

Song, B., Jian, S., 2025b. Privacy-preserving personalized pricing and matching for ride hailing platforms. *Commun Transp Res*, **5**, 100205.

Takagi, H., Moriya, S., Sato, T., Nagao, M., Higuchi, K., 2025. A framework for efficient development and debugging of role-playing agents with large language models. In: Proceedings of the 30th International Conference on Intelligent User Interfaces, 70–88.

Tjuatja, L., Chen, V., Wu, T., Talwalkwar, A., Neubig, G., 2024. Do LLMs exhibit human-like response biases? A case study in survey design. *Trans Assoc Comput Linguist*, **12**, 1011–1026.

Tversky, A., Kahneman, D., 1992. Advances in prospect theory: Cumulative representation of uncertainty. *J Risk Uncertainty*, **5**, 297–323.

Verma, N., Bharadwaj, M., 2025. The Hidden Space of Safety: Understanding Preference-Tuned Llms in Multilingual Context. https://arxiv.org/abs/2504.02708

Xie, J., Zhang, K., Chen, J., Zhu, T., Lou, R., Tian, Y., et al., 2024. TravelPlanner: A benchmark for real-world planning with language agents. *Proc Mach Learn Res*, **235**, 54590–54613.

Xu, H., Zhao, R., Zhu, L., Du, J., He, Y., 2024a. OpenToM: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, 8593–8623.

Xu, L., Chen, Y., Cui, G., Gao, H., Liu, Z., 2022. Exploring the universal vulnerability of prompt-based learning paradigm. In: Findings of the Association for Computational Linguistics: NAACL 2022, 1799–1810.

Xu, R., Wang, X., Chen, J., Yuan, S., Yuan, X., Liang, J., et al., 2024b. Character is Destiny: Can Role-Playing Language Agents Make Per-

sona-Driven Decisions. https://arxiv.org/abs/2404.12138

Xu, Z., Sengar, N., Chen, T., Chung, H., Oviedo-Trespalacios, O., 2025. Where is morality on wheels? Decoding large language model (LLM)-driven decision in the ethical dilemmas of autonomous vehicles. *Travel Behav Soc*, **40**, 101039.

Yi, Z., Ouyang, J., Xu, Z., Liu, Y., Liao, T., Luo, H., et al., 2026. A survey on recent advances in LLM-based multi-turn dialogue systems. *ACM Comput Surv*, **58**, 1–38.

Zhao, Q., Wang, J., Zhang, Y., Jin, Y., Zhu, K., Chen, H., et al., 2024. CompeteAI: Understanding the competition dynamics of large language model-based agents. *Proc Mach Learn Res*, **235**, 61092–61107.

Jianing Liu received the Ph.D. degree in civil engineering from The Hong Kong University of Science and Technology, Hong Kong, China, in 2024. He is currently an Assistant Professor at Southwest Jiaotong University, China. His research focuses on travel behavior analysis and discrete choice modeling.

Bing Song received the B.S. degree in electronic and information engineering from Beihang University (BUAA) in 2019, and the M.S. degree in transportation engineering from Shanghai Jiao Tong University (SJTU) in 2022. He is currently pursuing the Ph.D. degree in transportation engineering with The Hong Kong University of Science and Technology. His current research interests include privacy protection in transportation, transportation data analysis, and transportation safety.

Vinayak Dixit is the Associate Vice President for Global Research and Innovation at the University of New South Wales (UNSW). He is currently a Professor of Transport Systems in the School of Civil and Environmental Engineering. His research focused on developing and implementing new technologies in transport systems ranging from connected automated vehicles and electric vehicles to the use of quantum technologies in transportation systems.

Chenyang Wu received the Ph.D. degree in civil and environmental engineering from Imperial College London, UK, in 2020. She is currently an Associate Professor at Northwestern Polytechnical University, China. Her research focuses on the planning of emerging mobility services such as urban air mobility, mobility-as-a-service, and shared mobility.

Sisi Jian received the Ph.D. degree in transport engineering from the University of New South Wales in 2017. She is currently an Associate Professor with the Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology. Her research interests include transportation network modeling, game theory, and bilevel optimization.