Silicon-Based Sampling: Simulating Heterogeneous Decision Distributions via Zero-Shot Large Language Models

Wangpeiyi Chen; Hairuo Guo; Quan Hao; Yanan Sun*

Abstract

This paper proposes a zero-shot decision simulation method that does not require explicit behavioral rules. By inputting individual backgrounds and decision environments in natural language form, we generate "silicon-based samples" under zero-shot conditions to approximate the decision distribution of real human populations. Using insurance consumption decisions as an empirical scenario, the results of this paper indicate that after introducing complete semantic individual information, the population decision distribution generated by LLMs is highly consistent with the empirical human distribution, with the bidirectional KL divergence dropping significantly from 10.75 to 1.683. Further ablation analysis shows that psychological information plays a critical role in constraining the decision distribution, and its marginal contribution is significantly higher than that of other information. We also document boundaries: performance remains weaker for high-load multi-attribute trade-offs and is sensitive to prompt interventions that compress or "guide" the information. The paper clarifies when and how profile-conditioned, zero-shot LLM simulation can complement traditional surveys, discriminative prediction, and agent-based modeling for rapid distributional counterfactuals.

Keywords

large language models; decision simulation; silicon-based sampling; insurance choice; distributional fidelity

1 Introduction

Economic research and policy evaluation face a common challenge: when market institutions, information disclosure methods, or contracts change, researchers need to predict how heterogeneous individuals adjust their decisions, assessing the welfare and consequences of policy changes by constructing credible counterfactuals. Yet the relevant data are often missing at the decision point—because the policy has not been implemented, because outcomes lie in the tail, or because field experiments and surveys are too slow and costly to iterate(Manski 2004). In this paper, we propose that a zero-shot large language model (LLM), conditioned on respondent-specific natural-language profiles, can serve as a scalable tool for ex-ante distributional evaluation in behaviour research and policy settings.

* Chen: School of Information, Central University Of Finance And Economics,Beijing, China, Email: 2022312101@email.cufe.edu.cn. Guo: Center for Forecasting Science, Chinese Academy of Sciences, Beijing, China, Email: guohairuo@amss.ac.cn. Hao: School of Information Science and Technology, Beijing University of Technology, Beijing, 100124, China, Email: haoquan@emails.bjut.edu.cn. Sun： College of Economics and Management，Jining Normal University，Jining, Inner Mongolia, China，Email: antonsun2022@outlook.com.

With the advent of large language models (LLMs), researchers have proposed using LLMs to construct silicon-based samples to simulate social research (Toubia et al. 2025; Serapio-García et al. 2025). The simulation of human decision-making by generative AI based on the Transformer architecture is essentially prompt-based zero-shot prediction (Vaswani et al., 2017), while pre-trained large models can be viewed as a high-dimensional prior probability space containing massive human behavior patterns (Ye et al., 2025). Inputting a natural language Prompt with specific features as a constraint into the model, it essentially adds constraints to the high-dimensional space to form a conditional probability distribution, prompting the total probability space to compress toward a specific sub-probability space, and finally outputting a text with maximum expectation under that sub-probability space.

This paper proposes and tests a method: we view the large language model as a decision scenario simulator, using personal information and decision scenarios in natural language as its input, and constructing the decision choice distribution of that population by inputting the personal information of every real human in the group. Our method does not use algorithms to replace real human individuals, but provides a reproducible, verifiable "decision simulation model" for evaluating counterfactual distributions when real data is scarce or when the expected effects of alternative policies need to be quickly compared during the policy formulation stage. It should be emphasized that our method does not aim to identify individual preferences, but to obtain a distribution of decisions based on rich personal information.

We use an insurance decision task as a test scenario, which possesses classic economic decision problems such as risk assessment, time discounting, emotional anticipation, and multi-attribute trade-offs. We used real questionnaire data containing 204 respondents and designed five types of typical decision tasks, including intertemporal choice, willingness to pay, state dependence, regret emotion, and multi-attribute trade-offs. In terms of experimental design, we used the real human sample as the standard and constructed multiple sets of control scenarios: on the one hand, comparing the differences between the decision distribution generated by the LLM and the real distribution when providing no individual background information versus providing a complete natural language profile; on the other hand, comparing this method with traditional machine learning models based on structured features to assess the relative performance of the generative paradigm in macro-distribution fitting and micro-prediction. It implies that the LLM inference in this paper is completed under zero-shot conditions and does not involve any form of retraining or fine-tuning.

Our main finding is that by inputting personal information, the prediction accuracy of the large language model for individual decisions increases, and at the group level, the large language model with introduced personal information can achieve a distribution similar to the group decision distribution. We compared the baseline decision without individual information against the full-information decision containing demographics, financial information, and psychological traits. The results show that introducing personal information can significantly improve the fitting effect of the group distribution, with the bidirectional KL divergence dropping significantly from 10.75 to 1.683, and the JS divergence dropping from 0.430 to 0.117. Secondly, we conducted ablation experiments for different types of information. The results show that psychological characteristics are the most important factor in improving the fitted distribution; the bidirectional KL divergence of its single-factor model (2.025) is far superior to the model using only demographic information (KL=9.616) and the model using financial information (KL=3.505).

Among them, psychological information regarding risk perception and attitude, social influence and emotion, and intra-household dynamics has the most significant impact on simulation effects.

For the 5 decision tasks, we conducted simulation prediction analyses respectively. The results show that our method performs excellently on the four decision problems of intertemporal choice, willingness to pay, state dependence, and subsequent emotion, with small fitting differences from the real human distribution, while there is room for significant improvement in product trade-offs involving high cognitive load. Finally, we further explored the gain effect of existing human knowledge on the fitting of decision distributions. We found that this knowledge not only did not improve the fitting effect but actually reduced the degree of fit of the decision distribution. We believe that the reason lies in the fact that existing classical theories cannot completely describe all human behaviors; adding this part of information actually introduces a prior distribution into the model that is biased against real human behavior, thereby causing an increase in KL divergence.

Currently, mainstream economic analysis tools can be divided into four categories. First, econometric models. This method can quickly capture data-driven economic changes and provide short-term forecasts for policies and markets (Sims 1980). But the problem is also obvious: once the economic environment changes drastically, historical data will immediately become invalid, and model predictions will suffer shocks accordingly (Lucas 1976). The reason lies in overly strong assumptions leading to unavoidable model misspecification, spurious causality, and the famous Lucas critique (McCallum 2010). Second, General Equilibrium models and Dynamic Stochastic General Equilibrium (DSGE) models. This method examines the economy within a holistic framework: resource constraints and policy shocks across different sectors, regions, and times can be systematically included (Kaplan and Violante 2018). It starts from basic microeconomic principles—consumer utility maximization, producer profit maximization—and then glues together countless such individuals and sectors to construct a massive economic model with internally consistent logic (Kirman 1992). But under the assumption of equilibrium, people are oversimplified into representative agents with "rational expectations," and heterogeneity at the individual level is erased. However, real economic activity stems from individual heterogeneity, not homogeneity (Kaplan et al. 2018). Third, surveys and questionnaires. This method can directly control the way information is presented and obtain counterfactual results. But they are limited by scale and iteration speed (Manski 2004), and suffer from measurement problems such as classic stated preference bias and social desirability bias (Keillor et al. 2001; Taylor 2012). Fourth, ABM models. This method abandons the top-down path of traditional models based on preset macro laws and turns to a bottom-up simulation method (Schinckus 2019). It no longer tries to find a unified equation for the entire economy, but views the economy as an evolutionary system composed of a large number of autonomously interacting "agents" (Axtell and Farmer 2025). However, problems such as how to set the behavioral rules of agents without presetting the distribution, how to avoid creating black-box models that are too complex and difficult to explain, and how to effectively validate and calibrate the models empirically limit the further utility of this method (Dehkordi et al. 2023).

Our research contributes to this literature. We propose a zero-shot decision simulation method without explicit behavioral rules, providing an alternative approach to the rule setting and calibration problems long faced by ABM. However, we do not advocate replacing structural models or traditional ABM with LLMs, but position it as a distribution-level simulation tool

suitable for data scarcity, early-stage policy design, or scenarios requiring rapid comparison of multiple contexts.

In addition, this paper provides a line of thought in response to the questioning that LLMs "can only give average answers" (Gao et al., 2025). Our results show that the output of heterogeneous answers by large language models does not originate from their internal parameters (temperature, top_k), but from the heterogeneity of their input information. When the input information possesses sufficient personal characteristics, its output will also possess a considerable degree of personal characteristics; that is, the enhancement of input features can effectively mitigate the problem of the large model's answers presenting as homogenized or stereotyped. This point has been argued in Park et al. (2024) and Toubia et al. (2025). However, the questions tested in this class of literature were exposed to large language models during the pre-training stage, and their high accuracy for individual decisions mixes two components: the average human bias inherent in the model, and the private bias brought by prompt guidance. Our research combines classic problems with reality, adopting the method of replacing expressions to circumvent the "memory" and "search" capabilities of large language models. Our research provides clear usage boundaries and design principles for using large models for group behavior simulation.

The remainder of this paper is arranged as follows: Section 2 reviews relevant literature and defines the research background; Section 3 introduces the questionnaire, task design and evaluation metrics; Sections 4 and 5 report empirical results, robustness checks, and ablation analyses; Section 6 discusses prompt interventions, applicable boundaries, and responsible use; and Section 7 concludes.

2 Research Background

2.1 Traditional Consumer Decision Prediction and Limitations

Traditional consumer decision prediction has long been established on utility theory and the "rational man" assumption, viewing consumers as decision-making subjects capable of integrating information and maximizing expected utility, thereby providing a standardized analytical framework for demand prediction and preference measurement (Bray 2009). However, this idealized picture has been constantly revised in behavioral decision research: consumer judgment and choice are often limited by incomplete information, cognitive resources, and time pressure, and brand choice often presents characteristics of being intuition-driven rather than strictly calculated (Johnson and Puto 1987; Quevedo and Gopalakrishna 2021; Audia et al. 2024). At the empirical measurement level, Stated Preference (SP) methods have been recorded by a large number of studies as having systematic "hypothetical bias," that is, compared to Revealed Preference (RP) in real contexts, respondents in questionnaires/experiments tend to overestimate willingness to pay and purchase probability; relevant meta-analyses and experimental comparisons further indicate that this bias can reach significant magnitudes and is influenced by experimental protocols (Murphy et al. 2005; List and Gallet 2001; List 2001). To reduce bias, research has proposed procedural corrections such as "explicit warnings" and "certainty reporting/calibration," but these methods are more of a calibration of the measurement process rather than a reconstruction of the decision generation mechanism itself (List and Gallet 2001; Beck et al. 2016). Meanwhile, the intervention of social norms will further contaminate survey data through Social

Desirability Bias (SDB), especially in high-normativity issues such as ethical consumption, sustainable behavior, financial behavior, or piracy, where respondents are more likely to give answers "approved by society"; even using indirect questioning, scale-type self-reporting is still difficult to completely free from this problem (Keillor et al. 2001; Fisher 1993; King and Bruner 2000; Kelly et al. 2017; Taylor 2012; Bursztyn et al. 2025). Regarding data morphology, reliance on structured scales also compresses the semantic and contextual information of consumer experiences, while the use of user-generated content and unstructured data on the enterprise side is changing the way demand is understood and driving attention to richer data and modeling paradigms (Hauser et al. 2023; Shin et al. 2020; Adamopoulos et al. 2022; Bisaria et al. 2025).

2.2 Evolution of Prediction Paradigms: From Discriminative to Generative AI

In the past two decades, the mainstream path in the fields of information systems and marketing has been predictive analytics centered on discriminative modeling: mapping demographics, historical behavior, and environmental characteristics to labels such as churn, clicks, or purchases by learning conditional relationships to improve prediction performance and management decision efficiency (Shmueli and Koppius 2011). The key constraint of this paradigm lies in its reliance on "feature engineering"—the relationships the model can learn are highly limited by the variables constructed and encoded by researchers in advance. Faced with massive, heterogeneous, and unstructured data (such as text, images, cross-platform interactions), without systematic preprocessing and representation construction, the scalability and information utilization rate of traditional methods are limited (Liu et al. 2016). The rise of deep learning and representation learning constitutes an important transition: on the one hand, research has begun to map text and multimodal content into vector representations to support tasks such as branding, content, and text exploration, thereby capturing semantic structures with less manual feature design (Dew et al. 2022; Lee et al. 2022; Hartmann and Netzer 2023); on the other hand, deep learning has also been used to mitigate privacy and scale constraints in data transfer and extended to more complex data structures such as graph learning (Anand and Lee 2023; Das et al. 2023). However, mainstream applications in this stage are still dominated by classification, scoring, and effect evaluation (e.g., sentiment/attribute scoring, recommendation strategy effect estimation), solving the problem of "predicting more accurately" rather than "generation process and subject simulation" (Chakraborty et al. 2022; Adamopoulos et al. 2022; Ma and Sun 2020; Alantari et al. 2022). Generative AI-related research further pushes the move from discriminative prediction to generative content and interactive personalization: models can not only output scores but also generate marketing content such as text and images and participate in personalized services, thereby providing a new technical foundation and research agenda for simulation and intervention "with behavior sequences as the unit" (Parker et al. 2025; Hartmann et al. 2025; Grewal et al. 2025).

2.3 Silicon-Based Sampling and LLM Simulation

In the context of enhanced generative model capabilities, "silicon-based sampling" is gradually becoming an operable methodology: researchers use language models to generate a set of "simulated samples" under given conditions (such as demographic cues, situational information, or task instructions) to reproduce the response distribution of specific groups in surveys or experiments, and conduct rapid prototype testing and hypothesis testing based on this (Argyle et al. 2023; Rytting 2023). Relevant research further proposes the "Homo Silicus" perspective, viewing LLMs as parametrizable economic agents: by endowing them with preferences, endowments, and

information sets, the models can not only answer questionnaire items but also exhibit behavior patterns similar to humans in more complex economic interactions, thereby providing a new toolbox for the simulation of market and institutional contexts (Horton 2023; Korinek 2023; Kitadai et al. 2025). regarding market research applications, existing work has systematically discussed and empirically tested the usability of LLMs in surveys, concept testing, and insight generation, emphasizing their potential complementary value to traditional questionnaire/interview processes in terms of speed, cost, and iterability (Brand et al. 2023; Goli and Singh 2024). At the same time, the key issue at the methodological level lies in "algorithmic fidelity": that is, it is necessary to explicitly assess the degree of consistency between model-generated samples and real populations in terms of distribution, heterogeneity, and context sensitivity, and define under what conditions extrapolation is reliable (Rytting 2023; Argyle et al. 2023).

2.4 Probability Space, Context Dependence, and Emergent Abilities

From a mechanistic view, using LLMs for decision simulation can be understood as conditional generation in a linguistic probability space: the word sequence distribution learned by the model not only encodes linguistic forms but also settles the statistical laws of organizational interaction, argumentation structure, and action selection to a certain extent; therefore, the process of "generating the next word" may correspond to "generating reasonable decisions and explanations in that context" (Faraj et al., 2025). Under this framework, prompt words and In-Context Learning (ICL) can be viewed as constraints and update mechanisms for the generation space: by providing the model with task instructions, role settings, and situational information, researchers can induce it to produce more consistent choices and reasons under specific conditions; Chain-of-Thought (CoT) prompting further enhances the model's ability to output step-by-step reasoning in complex tasks, providing an operable technical path for process-level simulation (Wei et al., 2022). At the same time, the literature emphasizes the importance of "emergent abilities" for moving from shallow pattern matching to zero-shot generalization and stronger reasoning: when model scale and training data reach a threshold, reasoning, long-range consistency, and complex task capabilities may significantly increase, making the simulation of longer-span decision sequences possible (Wei et al., 2022; Binz & Schulz, 2023; Hu & Tian, 2025). In directions closer to management decision-making and engineering optimization, recent research has also begun to explore the agentic performance and application boundaries of LLMs in tasks such as decision-making under uncertainty, online learning, and combinatorial optimization, providing new evidence bases and risk discussion spaces for embedding generative models into real decision processes (Park et al., 2024; Jiang et al., 2025; Csaszar et al., 2024; Ejnarsson Lundberg, 2025).

3 Model and Experimental Design

This section details the insurance consumption decision simulation framework constructed in this study based on large language models (LLMs). The core mechanism employs natural language-driven semantic constraints to simulate heterogeneous individual decision-making processes. The model treats pre-trained LLMs as high-dimensional prior probability spaces, where multi-dimensional profiles of specific individuals serve as Bayesian constraints. This approach guides the model to converge from general probability distributions to specific sub-probability spaces, enabling zero-shot generation of decision outputs with silicon-based sample characteristics.

Figure 1 illustrates the silicon-based sample construction workflow under zero-shot inference.

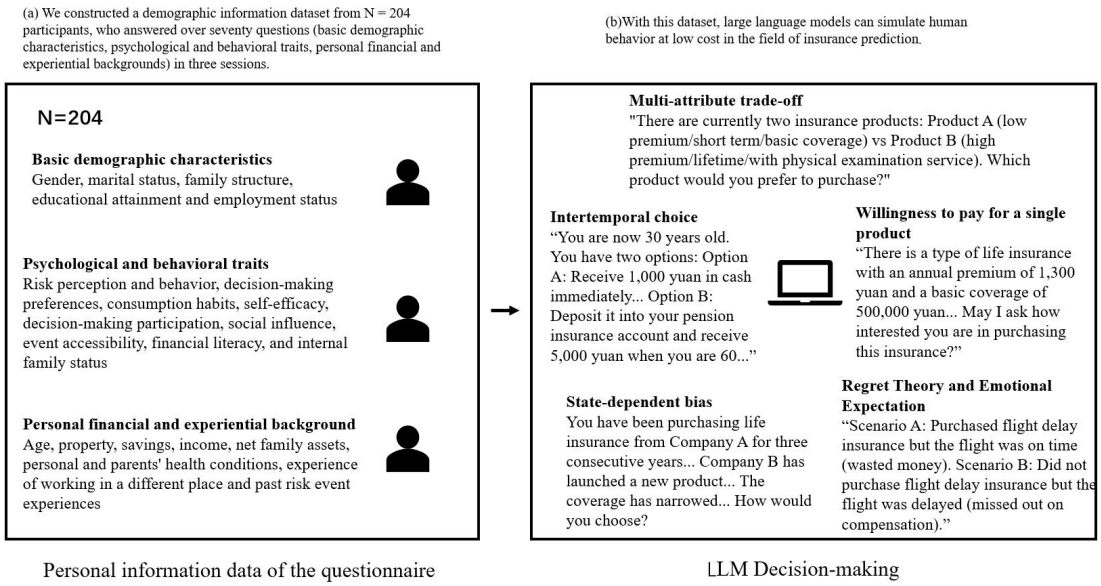


Figure 1 Silicon-Based Sample Construction Diagram

Note: This figure illustrates the pipeline for generating silicon-based samples. The process begins with collecting heterogeneous individual data (demographics, psychology, finance), transforming them into natural language profiles, and inputting them into a frozen LLM as a Bayesian prior. The model then performs zero-shot reasoning to output decision distributions compliant with the specific sub-probability space.

As shown in Figure 1, this study constructs an end-to-end silicon-based decision simulation framework aimed at realizing the mapping from real individual characteristics to decision behaviors. The operational logic of the model follows the path of "personal characteristics—probability space compression—zero-shot reasoning": First, the model transforms heterogeneous individual data obtained through questionnaires (including explicit demographic characteristics and financial status, as well as implicit psychological behavioral traits) into structured natural language prompt words (Prompt), thereby constructing a digital "user profile" with rich granularity; subsequently, this profile is injected into the large language model as a strong Bayesian prior, forcing the model to compress its general high-dimensional probability distribution to a subspace conforming to the specific individual characteristics; finally, in a specific insurance consumption scenario, the model performs zero-shot reasoning within the restricted probability subspace to generate the corresponding "silicon-based sample" decision sequence.

3.1 Questionnaire Design

Regarding questionnaire design, we integrated multiple papers and theoretical frameworks. regarding explicit factors, the questionnaire not only collected basic demographic characteristics such as gender, age, family composition, and marital status (Anderson 2018; Halek and Eisenhauer 2001), but also collected personal property (Flavin and Yamashita 2002), savings (Hubbard et al. 1995), income (Gropper and Kuhnen 2025), family net assets, health status of oneself and parents (Yogo 2016), non-local employment history (Dustmann et al. 2023), and past risk event experiences to supplement insurance decision information.

On the implicit psychological dimension, the construction of questionnaire items was strictly

based on theoretical definitions in the literature. Among them, risk illusion is defined as an individual's excessive worry about low-probability risks or relative underestimation of major risks (Dikmen et al. 2018), based on which measurement items regarding worry about the probability of serious illness or accidents were set. At the same time, studies indicate that marketing fraud perception and inertia also affect consumer decisions (Estelami and Liu 2024). Based on definitions in the literature, this questionnaire measured behaviors where marketers induce purchases through misleading quotes and situations where consumers habitually use institutions with poor service or uncompetitive rates (Estelami 2014).

Risk attitude also has a significant impact on consumer decisions (Lim et al. 2022); based on this, this questionnaire designed questions involving income reduction sensitivity and risk aversion willingness to pay. In addition, based on research mentioning that consumers are influenced by the financial decisions of those around them, peer effects were introduced to explore the decision pressure formed by the financial choices of surrounding groups on the individual (Lieber and Skimmyhorn 2018). Furthermore, studies point out that the financial literacy dimension covers multiple aspects such as financial knowledge, financial behavior habits, financial status, and information acquisition channels, all of which significantly affect consumers' financial decisions (van Rooij et al. 2024; Xiao and Tao 2020). Based on consumer empowerment theory, this questionnaire also constructed self-efficacy and perceived controllability modules, and integrated measurement items on decision authorization preference, purchase decision involvement, and perceived performance risk regarding insurance company claims fulfillment (Buehler and Maas 2018). Simultaneously, drawing on regret theory and decision content accessibility, this questionnaire was designed to measure the psychological pain anticipation consumers might experience (Bailey and Kinerson 2005) and the degree of influence of the activation level of vivid events in memory on personal decisions (Kusev et al. 2017).

We distributed the questionnaire on the "Jianshu" platform in three batches on December 24, 29, and 30 2025. By selecting participants with a past adoption rate of greater than or equal to 90% to ensure sample quality, and then screening samples based on attention checks and reverse validation questions, 46 invalid questionnaires were eliminated, and finally, 204 samples were collected.

The question list and measurement content are shown in Table 1. Please see Appendix A for the specific content of the questionnaire.

Table 1 Question List and Measurement Content

Task (Source)	Measurement Content
Gender	Gender may affect risk perception and insurance demand; differences may exist between males and females regarding insurance purchase types and amounts.
Marital Status (Halek and Eisenhauer 2001)	Marital status may affect the perception of family responsibility; married individuals are more likely to purchase family protection-oriented insurance such as life and health insurance.
Family Composition (Children Status) (Anderson 2018)	The influence of children structure on insurance allocation.
Family Composition (Number of Children) (Anderson 2018)	The number and age of children affect demands for education and medical insurance; the more or younger the children, the higher the related insurance demand may be.
Family Composition (Children Age) (Anderson 2018)	Children's age affects insurance purchasing decisions.
Education Level (Halek and Eisenhauer 2001)	Those with higher education levels typically have higher financial literacy and are more likely to understand and purchase complex insurance products.
Employment Status (Self-employed/Business	Self-employed individuals lack corporate protection and are more likely to purchase commercial insurance as risk compensation.

Task (Source)	Measurement Content
Owner) (Halek and Eisenhauer 2001)	
Risk Illusion (Estelami 2014)	Individuals with high risk illusion may underestimate true risks, leading to insufficient insurance purchasing. Consumers who are more worried about getting sick may be more inclined to purchase insurance.
Utility Function (Markowitz 2010)	The degree of concern for income affects risk decisions; different curve shapes may lead to differences in insurance purchase willingness.
Risk Tolerance (Lim et al. 2022)	Those with high risk tolerance may be more inclined toward investment-oriented insurance, while those with low tolerance prefer protection-oriented insurance.
Marketing Fraud Perception (Estelami 2014)	Those sensitive to marketing fraud may hold a skeptical attitude toward insurance sales pitches and be more cautious in purchasing decisions.
Information Presentation Bias (Anderson 2018)	Missing information may affect consumers changing their insurance purchase willingness. Brief or biased information may enhance consumers' willingness to change insurance purchases.
Overconfidence (Sandroni and Squintani 2007)	Overconfident individuals may believe they do not need insurance, leading to low purchase willingness.
Inertia (Estelami 2014)	Individuals with strong inertia may continue their existing insurance allocation and be unwilling to adjust or make new purchases.
Consumer Empowerment (Buehler and Maas 2018)	Those with high self-efficacy are more confident in making correct insurance decisions, and their purchase willingness may be stronger. Consumers who believe insurance claims are useful are more likely to purchase insurance.
Reverse Validation Question (N/A)	Serves as a reverse validation question for Question 29, used to test the consistency of self-efficacy measurement.
Decision Delegation Preference (Buehler and Maas 2018)	Those preferring delegation may entrust insurance decisions to others, resulting in low personal involvement.
Purchase Decision Involvement (Buehler and Maas 2018)	Those with high involvement are more likely to gain an in-depth understanding of insurance products, making purchasing decisions more rational.
Peer Effect (Lieber and Skimmyhorn 2018)	Those with strong peer effects may follow suit in purchasing insurance because their social circle does so. Making insurance consumption decisions by following friends' decisions.
Anticipated Emotion (Bailey and Kinerson 2005)	Those with positive anticipated emotions may be more willing to pay for future protection, resulting in high purchase willingness.
Decision Content Accessibility (Kusev et al. 2017)	Those who can easily access insurance-related information are more likely to make informed decisions.
Non-rational State Influence (Kusev et al. 2017)	Non-rational states (such as mood swings) may lead insurance purchase decisions to deviate from rationality.
Financial Literacy (Xiao and Tao 2020)	Those with high financial literacy better understand insurance functions, and their purchase willingness and allocation are more reasonable.
Attention Check Question (N/A)	Used to screen invalid samples; does not directly affect insurance purchasing.
Perceived Performance Risk (Buehler and Maas 2018)	Those with high trust in insurance companies are more willing to purchase their products. Overall trust in the insurance industry affects purchase willingness and product choice.
Sense of Control (Lim et al. 2022)	Those with a strong sense of control may believe they can cope with risks themselves, leading to low insurance demand. Consumers capable of setting and achieving long-term financial goals may be more inclined to purchase insurance.
Perceived Spousal Social Support (Lim et al. 2022)	Those with high spousal support may be more active in family insurance planning. It may be easier to make financial decisions when supported by a spouse.
Insurance Positioning in Family Finance (Li et al. 2021)	Those who view insurance as a necessity are more likely to continue purchasing and renewing.
Risk Attitude (Lim et al. 2022)	Risk-averse individuals are more inclined to purchase protection-oriented insurance.
Age (Hambel et al. 2017)	Age affects the type of insurance demand; for example, young people prefer health insurance, while middle-aged and elderly people focus on pension insurance.
Property Status (Flavin and Yamashita 2002)	Property owners are more likely to purchase property insurance and family liability insurance.
Regular Savings Status (Hubbard et al. 1995)	Those with saving habits are more likely to have financial planning awareness and stronger insurance purchase willingness.
Income Status (Gropper and Kuhnen 2025)	The higher the income, the higher the insurance purchasing power and demand typically are.
Family Net Assets	Those with high net assets are more likely to purchase high-value or diversified

Task (Source)	Measurement Content
	insurance products.
Parents' Income (Chew et al. 2018)	Parents' income may affect an individual's economic security and insurance demand.
Health Status (Yogo 2016)	Those with poor health are more likely to purchase health insurance and critical illness insurance.
Parents' Health Status (Ko 2022)	Parents' health issues may affect the individual's awareness of purchasing insurance.
Non-local Employment History (Dustmann et al. 2023)	Non-local employees may rely more on commercial insurance to make up for gaps in social security.
Social Basic Medical Insurance Participation	Those who already have basic medical insurance may have reduced demand for commercial health insurance or purchase it as a supplement.
Social Basic Pension Insurance Participation	Those who already have pension security may have reduced demand for commercial pension insurance or use it as a supplement.
Commercial Insurance Participation	Those who have purchased commercial insurance are more likely to continue purchasing or upgrade their insurance.
Major Event Experience (Gallagher 2014)	Those who have experienced major risk events value insurance protection more, and their purchase willingness is enhanced.
Claims Experience	Those with good claims experience trust insurance more and have high willingness to renew and purchase additional insurance; those with poor experience may stop purchasing.

Note: This table summarizes the variables collected in the survey. Source indicates the literature reference for the theoretical construct. Measurement Content describes the rationale behind including the specific variable for insurance decision profiling. Explicit factors (e.g., demographics) and implicit factors (e.g., risk illusion, psychological traits) are both included to construct the "Full Context" profile.

3.2 Insurance Purchase Scenarios and Decisions

To verify the expressive power of silicon-based samples in real consumption decisions, we selected the following 5 questions for testing. Table 2 reports the stem content of the questions. For more content, please see Appendix A.

Table 2 Insurance Purchase Scenarios and Decisions

No.	Decision	Content	Question
Q1	Intertemporal Choice	Tests time preference	You are now 30 years old. You have two choices: Choice A: Get 1,000 yuan cash immediately... Choice B: Deposit into a pension insurance account, get 5,000 yuan at age 60...
Q2	Single Product Willingness to Pay	Tests value perception	There is a life insurance policy with an annual premium of 1,300 yuan and a basic sum insured of 500,000 yuan... How willing are you to buy this insurance?
Q3	State-Dependent Bias	Tests inertia and switching costs	You have purchased life insurance from Company A for three consecutive years... Company B launches a new product... narrower coverage... How would you choose?
Q4	Regret Theory and Emotional Anticipation	Tests irrational emotional factors	Imagine this scenario: Scenario A: Bought delay insurance but flight was on time (wasted money). Scenario B: Didn't buy delay insurance but flight was delayed (missed compensation). Question: Which scenario would cause you stronger negative emotion?
Q5	Multi-Attribute Trade-off	Tests trade-offs under high cognitive load	Two insurance products available: Product A (Low premium/short term/basic coverage) vs Product B (High premium/lifetime/includes physical exam service). Which product do you prefer to buy?

Note: This table outlines the five specific decision-making tasks (Q1–Q5) used to evaluate the model. Each task corresponds to a classic economic or psychological theory (e.g., Intertemporal Choice, Regret Theory) to test the model's capability in simulating various dimensions of consumer behavior.

To verify the expressive power of silicon-based samples in real consumption decisions, this study designed five types of typical decision tasks (Q1-Q5). These scenarios not only cover typical links in insurance consumption but are also used to assess the fit between the model and

real decisions.

The intertemporal choice decision requires individuals to decide between immediate profit and deferred profit to measure their time bias and time discount rate. Insurance products naturally possess the intertemporal attribute of "current expenditure, future benefit." By setting a contrast between receiving cash at age 30 and receiving a pension at age 60, it is possible to probe the patience level of individuals in long-range expectations. As Manski (2004) pointed out, measuring such subjective expectations is crucial for understanding individuals' long-term precautionary savings motives in data-missing contexts.

The single product decision aims to test individuals' value perception of insurance and the boundaries of risk premium by setting specific annual premiums and insured amounts.

State-dependent bias focuses on whether individuals will change their long-term insurance habits when facing new and old product choices, used to test decision inertia and state dependence. Previous studies have pointed out that financial consumers show significant status quo bias in decisions; even if new products have price advantages, switching costs and dependence on the original institution will inhibit their adjustment behaviors (Estelami 2014).

According to Bailey and Kinerson (2005), regret avoidance is an important determinant of individual risk tolerance, and many insurance purchase decisions are influenced by irrational emotions. When facing emotions generated by payouts, it can reflect the individual's irrational psychology. Q4 captures the emotional anticipation of the sample in the risk-avoidance process by comparing the negative feelings of "wasting money" and "missing compensation."

In the multi-product trade-off scenario, the questionnaire simulated the trade-off decisions of the sample in a high cognitive load environment by setting different premiums, insurance terms, and additional benefits.

3.3 Model Setting

In terms of specific architectural implementation, the model systematically integrates three dimensions measured by the questionnaire: basic demographic characteristics [Part1_Demo], implicit psychological and behavioral traits [Part3_psych], and personal financial and experience background [Part4_finance]. Finally, these profile traits and specific insurance decision test questions ([Question_Test]) are logically combined to construct the following standardized prompt processing sequence:

$$Prompt = [Part1_\text{Demo}] + [Part3_\text{psych}] + [Part4_\text{finance}] + [Question_\text{Test}]$$

Each respondent profile is constructed by mapping the questionnaire modules into a standardized natural-language template (demographics, financial background, psychological traits) and concatenating it with the corresponding decision-task stem. Appendix A reports the full questionnaire items and task stems.

And as a control baseline group, only the test question is input in the prompt.

$$Prompt = [Question\ Test]$$

3.4 Evaluation Indicator System

This study adopts a multi-level indicator system to evaluate model performance, comprehensively considering everything from the precision of individual predictions to the fit of group distributions.

This paper uses Accuracy, Weighted Kappa Coefficient (QWK), and Stability to measure the micro-prediction ability of silicon-based individuals.

(1) Accuracy: The proportion of samples correctly predicted by the model, serving as a basic

evaluation indicator.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} = \frac{\sum_{i=1}^{n} I(\text{True}_i = \text{Predict}_i)}{n}$$

(2) Weighted Kappa Coefficient: Used to measure the consistency between predicted values and true values in ordinal classification tasks, addressing the class imbalance problem through squared penalties.

$$\text{QWK} = \frac{\sum_{i,j} w(i,j) \cdot \left(P_o(i,j) - P_e(i,j)\right)}{1 - P_e(i,j)}$$

(3) Stability (Agree_Rate): The fraction of identical model outputs across repeated inferences under the same prompt (and, where applicable, under specified perturbations), used as a reliability diagnostic for stochastic decoding.

$$\text{Stability} = \frac{\text{Number of Stable Outputs}}{\text{Total Outputs}} = \frac{\sum_{i=1}^{n} I\left(\text{Predict}_i = \text{Predict}_i^{(\text{new})}\right)}{n}$$

For each decision task, we form empirical choice distributions for the human sample and the model outputs under the corresponding prompt. We compute Forward KL (Real||Pred), Reverse KL (Pred||Real), their sum (Bidirectional KL), and JS on these discrete distributions.

(4) KL Divergence:

Forward KL (Real||Pred): Measures the information loss of the model prediction distribution deviating from the real distribution.

$$D_{\text{Forward KL}}(\text{True} \parallel \text{Predict}) = \sum_x P_{\text{True}}(x) \log\left(\frac{P_{\text{True}}(x)}{P_{\text{Predict}}(x)}\right)$$

Reverse KL (Pred||Real): Measures the difference of the real distribution relative to the prediction distribution, used to detect mode collapse.

$$D_{\text{Reverse KL}}(\text{Predict} \parallel \text{True}) = \sum_x P_{\text{Predict}}(x) \log\left(\frac{P_{\text{Predict}}(x)}{P_{\text{True}}(x)}\right)$$

Bidirectional KL: Comprehensive assessment of the mutual difference between two distributions.

$$D_{\text{Bidirectional KL}}(\text{True} \parallel \text{Predict}) + D_{\text{KL}}(\text{Predict} \parallel \text{True})$$

(5) JS Divergence: A symmetric smoothed improvement version of KL divergence, with values in the interval [0, 1], solving the asymmetry problem of KL divergence, used to measure distribution differences more robustly.

$$M = \frac{1}{2}(\text{True} + \text{Predict})$$

$$D_{\text{JS}}(\text{True,Predict}) = \frac{1}{2} D_{\text{KL}}(\text{True} \parallel M) + \frac{1}{2} D_{\text{KL}}(\text{Predict} \parallel M)$$

4 Empirical Results

We proceed in three steps. First, we test whether profile-conditioned, zero-shot LLM inference can recover the population distribution of insurance decisions, using bidirectional KL and JS as primary distribution-fidelity outcomes. Second, we conduct robustness tests using placebo effects by modifying the temperature, replacing the large language model, and reordering the input information. Third, we compare our results with those from supervised learning to explore the performance gap between our method and supervised learning trained with data.

Finally, we perform ablation studies on the input information to analyze which information is most valuable for the simulation.

4.1 Full Sample Results

Based on the empirical data of the full sample (N=204), we conducted a comparative analysis between the "Baseline Group (Pure_Question2)" and the "Full Information Group (BG_1_3_4)" containing all personal information. Table 3 reports the average results of the relevant indicators for the two groups across 5 questions.

Table 3 Full Sample Results (n=204)

Condition	Pure_Question2	BG_1_3_4
Accuracy	0.304	**0.411**
Stability (Agree_Rate)	0.843	**0.847**
QWK	-0.004	**0.135**
KL_Forward (Real||Pred)	9.642	**1.401**
KL_Reverse (Pred||Real)	1.108	**0.282**
KL_Bidirectional	10.750	**1.683**
JS_Divergence	0.430	**0.117**

Note: Pure_Question2 represents the baseline condition with no background information. BG_1_3_4 represents the full-profile condition including demographic, psychological, and financial information. Accuracy measures individual-level prediction correctness. QWK (Quadratic Weighted Kappa) measures ordinal agreement. KL Divergence (Kullback-Leibler) and JS Divergence (Jensen-Shannon) measure the distance between the model-generated distribution and the empirical human distribution; lower values indicate better distributional fidelity.

In the full sample (Table 3), the full-information condition (BG_1_3_4) reduces bidirectional KL divergence from 10.750 to 1.683 and JS divergence from 0.430 to 0.117 relative to prompting with the question alone (Pure_Question2). Micro-level diagnostics also improve (Accuracy: from 0.304 to 0.411; QWK: from 0.004 to 0.135) while stability remains comparable (Agree_Rate: 0.843 vs 0.847). This indicates that introducing individual features via natural language acts as a strong Bayesian constraint, successfully guiding the large language model to converge from a general prior distribution to a probability subspace highly consistent with the real population. In terms of accuracy, the full information group also showed improvement.

4.2 Robustness

To verify the robustness of the model results, this section employs replacement of large language model internal parameters, large language model architectural heterogeneity, and methodological paradigm comparison to complete robustness testing.

4.2.1 Temperature

The output of large language models possesses randomness, usually controlled by the Temperature parameter. To verify that the model's fit to the real population distribution does not stem from coincidences brought by high randomness, we set the Temperature parameter to 0.0, 0.4, and 0.8 respectively, and repeated the above experiments under each setting. Table 4 shows the experimental results under different parameters.

Table 4 Experimental Results with Different Parameters

Temperature	0.8		0.4		0.0	
Condition	Pure_Question2	BG_1_3_4	Pure_Question2	BG_1_3_4	Pure_Question2	BG_1_3_4

Accuracy	0.307	0.412	0.302	0.406	0.306	0.405
Stability (Agree_Rate)	0.852	0.854	0.849	0.849	0.856	0.855
QWK	0.010	0.138	0.002	0.140	0.004	0.136
KL_Forward (Real‖Pred)	9.475	1.196	9.892	1.339	10.869	1.197
KL_Reverse (Pred‖Real)	1.103	0.283	1.113	0.279	1.123	0.286
KL_Bidirectional	10.577	1.479	11.005	1.618	11.992	1.483
JS_Divergence	0.429	0.117	0.438	0.116	0.440	0.118

Note: This table reports model performance under varying temperature settings ($T \in \{0.0, 0.4, 0.8\}$). Results indicate that the fit of the full-information group (BG_1_3_4) remains stable and superior to the baseline across different randomness levels, suggesting the results are driven by profile conditioning rather than stochastic noise.

Data corresponds to the source text. For BG_1_3_4, KL_Bidirectional values are 1.479 (Temp 0.8), 1.618 (Temp 0.4), and 1.483 (Temp 0.0).

The experimental results indicate that regardless of how Temperature changes, the bidirectional KL divergence of the full information group (BG_1_3_4) consistently stabilizes in the low range of 1.479 to 1.618, far lower than the baseline group's 10.577 to 11.992. This suggests that the constraint power of natural language Prompts on the model probability space is dominant, effectively resisting the noise interference brought by sampling parameter changes. The model's restoration of human decision distributions is based on effective reasoning of user features, not randomly generated artifacts.

4.2.2 Replacing Large Language Models

To rule out the influence of specific model training data bias on the results and verify the universality of the paradigm, we replaced the underlying inference model with GPT-4o, DeepSeek-v3, and Gemini-3.0-flash, conducting cross-model validation. Table 5 reports the performance of each model in the baseline group and the full information group.

Table 5 Cross-Model Validation Results

Condition	GPT-4o		DeepSeek-v3		Gemini-3.0-flash	
	Pure_Question2	BG_1_3_4	Pure_Question2	BG_1_3_4	Pure_Question2	BG_1_3_4
Accuracy	0.281	0.390	0.291	0.409	0.380	0.428
Stability (Agree_Rate)	0.804	0.815	0.927	0.902	0.927	0.890
QWK	0.000	0.048	-0.007	0.049	0.003	0.154
KL_Forward (Real‖Pred)	8.866	5.914	12.848	3.566	10.560	1.798
KL_Reverse (Pred‖Real)	1.249	0.623	1.173	0.571	0.916	0.268
KL_Bidirectional	10.115	6.538	14.021	4.137	11.476	2.066
JS_Divergence	0.421	0.259	0.454	0.244	0.382	0.118

Note: Comparison of simulation performance across different LLM architectures (GPT-4o, DeepSeek-v3, Gemini-3.0-flash). BG_1_3_4 consistently outperforms Pure_Question2 across all

models, demonstrating the generalizability of the proposed method. Lower KL and JS values indicate a closer fit to human data.

The results show that for all models, after introducing background information, both KL divergence and JS divergence dropped substantially. For example, DeepSeek-v3's bidirectional KL divergence dropped from 14.021 to 4.137, and Gemini-3-flash dropped significantly from 11.476 to 2.066. In this task, Gemini-3-flash demonstrated fitting capabilities closest to the main experiment model (GPT-5), with its full information group JS divergence being only 0.118, highly consistent with the main experiment results.

4.2.3 Random Permutation of Input Information

To further verify that the improvement in model performance indeed stems from logical inference of "specific individual features," rather than simply from random gains due to increased Prompt length or general background noise, we designed a "placebo test" based on information misalignment.

We constructed a shuffled matching mechanism. While keeping the total amount of questionnaire questions (Part 2) and all background information (Part 1, 3, 4) unchanged, by rotationally displacing the sample list, we forced a pairing of user background information U_i with user decision problem D_j (where $i \neq j$). At this time, although the model received equally rich and real natural language descriptions, the input features did not match the decision subject. If the paradigm proposed in this study is effective, breaking the "person-decision" correspondence should lead to a significant decline in micro-prediction performance.

Table 6 shows the comparison results between the original matching group and the shuffled matching group.

Table 6 Input Information Random Permutation Experimental Results

Condition	BG_1_3_4_Shuffled	BG_1_3_4_Original
Accuracy	0.373	0.422
Stability (Agree_Rate)	0.856	0.852
QWK	0.013	0.134
KL_Forward (Real||Pred)	1.673	1.659
KL_Reverse (Pred||Real)	0.294	0.288
KL_Bidirectional	1.968	1.947
JS_Divergence	0.123	0.120
McNemar_p_exact	0.000***	

Note: BG_1_3_4_Original uses correctly matched user profiles. BG_1_3_4_Shuffled randomly permutes profiles (mismatched user-prompt pairs). The significant drop in individual-level metrics (Accuracy, QWK) in the shuffled group confirms that the model utilizes specific individual information for prediction. *** indicates $p<0.001$ based on the McNemar test.

The experimental results revealed two issues. First, the significant decline in micro-level prediction suggests that profile–person matching matters for individual alignment. When background information was misaligned with the decision maker, the model's weighted kappa (QWK) fell from 0.135 to 0.013, and accuracy also declined (McNemar $p < 0.001$). This pattern is consistent with micro-level gains being driven by matched conditioning information rather than

prompt length alone. When conditioned on mismatched profiles, the model outputs can remain coherent but are less aligned with the intended individual benchmark.

Second, the stability of the macro-distribution reveals the source of robustness for group simulation. It is worth noting that the KL divergence and JS divergence of the shuffled group are very close to the original group (KL_Bidirectional: 1.968 vs 1.947). This result does not conflict with the conclusions of this paper but further reveals the internal mechanism of LLM-based ABM simulation: the fit of the group distribution depends on the overall distribution characteristics of the input group profile set, rather than the one-to-one correspondence of individuals. Since the shuffling mechanism only changed the arrangement order of the samples without changing the total probability density distribution P(X)P(X)P(X) of the input information, the aggregated decision distribution P(Y)P(Y)P(Y) output by the model can still restore the overall preference shape of the real population.

In summary, the random permutation experiment further clarifies the applicable boundaries of the method in this study: if the goal is to predict the behavior of specific individuals, sufficient profile features must be ensured; whereas if the goal is to simulate market macro-distributions (such as policy evaluation), it is necessary to construct a silicon-based group profile library that is representative in statistical characteristics.

4.3 Comparison with Machine Learning Models

This subsection aims to verify the superiority of the "natural language + generative model" paradigm relative to the traditional "structured features + discriminative model" paradigm. We input questionnaire background information into machine learning models in the form of structured features and adopted a unified data preprocessing pipeline to ensure reproducibility and cross-model comparability.

In the data preprocessing stage, for numerical features, median imputation was used to handle missing values, and standardization was performed to satisfy zero mean and unit variance in the training set; for categorical features, mode imputation was used to handle missing values, and one-hot encoding was used to transform them into sparse indicator variables. We used 5-fold cross-validation to test the results of machine learning models on this task. The comparative experimental results are shown in Table 7.

Table 7 Comparison Results of Random Forest Model and Large Language Model

Condition	Random Forest	Extra Trees	Gradient Boosting	Ada Boost	Logistic Regression	BG_1_3_4
Accuracy	0.471	0.454	0.422	0.416	0.375	0.411
Stability (Agree_Rate)	0.939	0.954	0.956	0.999	1.000	0.847
QWK	0.155	0.165	0.106	0.091	0.159	0.135
KL_Forward (Real‖Pred)	2.054	2.150	0.818	1.223	0.580	1.401
KL_Reverse (Pred‖Real)	0.317	0.298	0.371	0.246	0.372	0.282
KL_Bidirectional	3.080	2.448	1.189	1.469	0.952	1.683
JS_Divergence	0.134	0.117	0.063	0.066	0.058	0.117

Note: Comparison between the zero-shot LLM approach (BG_1_3_4) and traditional supervised learning models (trained with 5-fold cross-validation). While Random Forest achieves higher micro-accuracy, the LLM demonstrates superior performance in capturing macro-distributional heterogeneity (lowest JS Divergence and Bidirectional KL), avoiding the "mode collapse" common in discriminative models.

Observing Table 7, compared to discriminative models exemplified by Random Forest, the large language model demonstrated a stronger ability to capture group heterogeneity. Although the Random Forest model achieved the highest 0.471 in prediction accuracy for micro-individuals, the cost was a severe distortion of the macro-distribution, with its bidirectional KL divergence reaching as high as 3.080, significantly worse than the LLM full information group's 1.683. This phenomenon reveals the limitations of traditional ensemble tree models when dealing with social science data: to minimize the loss function, the model tends to filter out "noise" in the data and converge toward the mode. In contrast, the LLM, based on a generative probabilistic sampling mechanism, does not attempt to eliminate uncertainty in human decision-making, but preserves the possibility of long-tail choices.

It should be pointed out that Gradient Boosting and Logistic Regression were slightly superior to LLM in distribution fitting indicators (KL divergence 1.189 and 0.952 respectively), but this actually reflects an asymmetric competition between different training paradigms. Machine learning models run within a fully supervised learning framework with 5-fold cross-validation, meaning the models have fully fitted the feature distribution of historical data before inference; whereas the LLM is under zero-shot conditions, reasoning solely on semantic understanding of the natural language Prompt and general priors. In other words, the LLM, without exposure to any training data, can obtain fitting results approaching supervised learning models solely through semantic inference.

4.4 Analysis of Decisions by Scenario

To deeply explore the performance of large language models in different types of insurance decision tasks, we further conducted a detailed analysis of the five specific decision scenarios (Q1-Q5).

4.4.1 Intertemporal Choice

Intertemporal choice (Q8) aims to test individuals' time preferences. Table 8 reports the indicator results for the baseline group and the full information group.

Table 8 Intertemporal Choice Results

Condition	Pure_Question2	BG_1_3_4
Accuracy	0.333	0.395
Stability (Agree_Rate)	0.89	0.818
QWK	0.001	0.138
KL_Forward (Real||Pred)	10.046	1.83
KL_Reverse (Pred||Real)	1.07	0.333
KL_Bidirectional	11.116	2.163
JS_Divergence	0.443	0.139

Note: Comparison of distributional fit for time preference decisions. The introduction of full profiles (BG_1_3_4) significantly reduces distribution divergence compared to the baseline, indicating effective capture of individual time discounting factors.

Experimental results show that in the absence of background information, the baseline group's KL divergence was as high as 11.116, indicating that the general model cannot accurately capture individuals' patience levels. By contrast, the experimental group introducing full information profiles significantly reduced the KL divergence to 2.163 by integrating age, income, and psychological characteristics, and raised QWK to 0.138. This indicates that natural language profiles effectively constrained the model's inference of time discount rates.

4.4.2 Single Product Willingness to Pay

Single product willingness to pay aims to test individuals' value perception ability. Table 9 reports the indicator results for the baseline group and the full information group.

Table 9 Single Product Decision Results

Condition	Pure_Question2	BG_1_3_4
Accuracy	0.216	0.472
Stability (Agree_Rate)	0.78	0.869
QWK	0.016	0.251
KL_Forward (Real||Pred)	7.425	1.177
KL_Reverse (Pred||Real)	1.266	0.267
KL_Bidirectional	8.691	1.443
JS_Divergence	0.446	0.113

Note: Comparison of distributional fit for single product willingness to pay. The introduction of full profiles (BG_1_3_4) significantly reduces distribution divergence compared to the baseline, indicating effective capture of individual time discounting factors.

Notes: Accuracy doubles from 0.216 to 0.472; KL_Bidirectional drops from 8.691 to 1.443.

Test results show that the effectiveness of the natural language paradigm is particularly outstanding. After introducing background information, the model's prediction accuracy doubled from 0.216 to 0.472, and QWK increased from 0.016 to 0.251. More importantly, the bidirectional KL divergence plummeted from 8.691 to 1.443, and JS divergence was only 0.113. This result means that as long as sufficient user background is provided, the large language model can highly restore the acceptance distribution of specific prices by specific populations.

4.4.3 State-Dependent Bias

State-dependent bias aims to test the decision inertia of individuals when facing choices between new and old products. Table 10 reports the indicator results for the baseline group and the full information group.

Table 10 Inertia Decision Results

Condition	Pure_Question2	BG_1_3_4
Accuracy	0.402	0.485
Stability (Agree_Rate)	0.877	0.834
QWK	-0.006	0.14
KL_Forward (Real||Pred)	3.66	1.723
KL_Reverse (Pred||Real)	0.814	0.221
KL_Bidirectional	4.475	1.944
JS_Divergence	0.341	0.092

Note: Comparison of distributional fit for state-dependent biasy. The introduction of full profiles (BG_1_3_4) significantly reduces distribution divergence compared to the baseline, indicating effective capture of individual time discounting factors.

The results show that the baseline group exhibited extremely high randomness, with QWK close to 0 and a large distribution deviation (KL_Bidirectional 4.475). However, after introducing the complete user profile, the bidirectional KL divergence significantly dropped to 1.944, and JS divergence dropped from 0.341 to 0.092.

4.4.4 Regret Emotion

Regret emotion aims to test irrational emotional factors. Table 11 reports the indicator results

for the baseline group and the full information group.

Table 11 Regret Decision Results

Condition	Pure_Question2	BG_1_3_4
Accuracy	0.407	0.43
Stability (Agree_Rate)	0.978	0.878
QWK	0	0.07
KL_Forward (Real||Pred)	12.327	1.111
KL_Reverse (Pred||Real)	0.899	0.215
KL_Bidirectional	13.227	1.326
JS_Divergence	0.39	0.085

Note: Comparison of distributional fit for regret emotion. The introduction of full profiles (BG_1_3_4) significantly reduces distribution divergence compared to the baseline, indicating effective capture of individual time discounting factors.

The results show that the baseline group's KL divergence was as high as 13.227, the group with the largest distribution difference among all questions. This indicates that the LLM's general training data has a serious disconnect regarding the cognition of "regret" compared to the risk preferences of the specific experimental population in this study. The model showed "blind confidence," with a Stability indicator of 0.978, but the distribution deviated seriously from real decisions. However, the experimental group drastically pulled the KL divergence down to 1.326, and JS divergence to 0.085.

4.4.5 Multi-Attribute Trade-off

Multi-attribute trade-off aims to test individuals' trade-off choices under high cognitive load. Table 12 reports the indicator results for the baseline group and the full information group.

Table 12 Multi-Attribute Decision Results

Condition	Pure_Question2	BG_1_3_4
Accuracy	0.162	0.273
Stability (Agree_Rate)	0.692	0.837
QWK	-0.029	0.076
KL_Forward (Real||Pred)	14.752	1.166
KL_Reverse (Pred||Real)	1.491	0.372
KL_Bidirectional	16.243	1.539
JS_Divergence	0.529	0.155

Note: Comparison of distributional fit for multi-attribute trade-off. The introduction of full profiles (BG_1_3_4) significantly reduces distribution divergence compared to the baseline, indicating effective capture of individual time discounting factors.

The results show that because the choice decision of "low premium/low coverage vs. high premium/high coverage" involves complex trade-offs with high cognitive load, the individual prediction accuracy of both the baseline group and the full information group remained at a very low level. Despite this, the experimental group's KL divergence still dropped significantly from the baseline group's 16.243 to 1.539. This result reveals the core value of LLMs as "market simulators": even when unable to precisely predict the specific choice of every individual, the model can still extremely accurately reproduce the preference distribution of the group.

5. Ablation Experiments

In the empirical analysis of the previous chapter, we confirmed that complete user profiles constructed based on natural language (Full Context) can significantly improve the fitting ability of large language models for group decision distributions. However, a core theoretical question arises: Does this improvement in fitting ability stem from simple alignment with explicit demographic characteristics, or from a deep understanding of implicit psychological mechanisms? Furthermore, is there redundancy or synergistic effects among information of different dimensions in decision simulation?

To deconstruct the independent contributions and interaction effects of features of different dimensions on LLM simulation capabilities, this section designed systematic ablation experiments (Ablation Study). We decoupled the natural language prompt words constituting the "silicon-based sample" into three independent feature modules: Demographic Characteristics (BG_1): containing explicit labels such as age, gender, education; Psychological Behavioral Traits (BG_3): containing implicit variables such as risk attitude, sense of control, financial literacy; Financial and Experience Background (BG_4): containing constraints such as income, assets, and past claims history.

This section will proceed following the logic of "single-factor efficacy analysis—multi-factor interaction effects—psychological factor dismantling." First, we independently test the benchmark performance of each feature module to identify the dominant factors affecting decision distribution; second, we test the complementarity between different information through feature combinations; finally, we conduct fine-grained dismantling of the psychological trait module, which showed the most significant performance, to precisely locate the key psychological mechanisms driving the simulation.

5.1 Single-Factor Efficacy Analysis

Table 13 shows the single-factor experimental results.

Table 13 Single-Factor Experimental Results

Condition	Pure_Question2	BG_1	BG_3	BG_4
Accuracy	0.304	0.356	**0.402**	0.394
Stability (Agree_Rate)	0.843	**0.900**	0.861	0.858
QWK	-0.004	0.038	**0.127**	0.04
KL_Forward (Real||Pred)	9.642	8.823	**1.712**	3.036
KL_Reverse (Pred||Real)	1.108	0.794	**0.312**	0.47
KL_Bidirectional	10.75	9.616	**2.025**	3.505
JS_Divergence	0.43	0.334	**0.128**	0.196

Note: Comparison of model performance using only one type of background information. BG_1: Demographics only; BG_3: Psychological traits only; BG_4: Financial background only. Results show that psychological traits (BG_3) provide the most significant contribution to distributional fidelity (lowest KL/JS divergence).

Observing Table 13, the BG_3 model containing only psychological trait factors performed best among all single-factor models, with a bidirectional KL divergence reaching 2.025. By contrast, the BG_1 model relying only on demographic characteristics performed the worst, with a KL divergence as high as 9.616, yet exhibiting the highest stability (Stability 0.900). This indicates that relying solely on demographic labels leads the model into "deterministic errors," that

is, producing stereotype-like predictions. For insurance consumption, which involves risk perception, the information content of "implicit psychological preferences" is significantly higher than "explicit population labels." The performance of the BG_4 model using financial background information lies between the two, with a KL_Bidirectional of 3.505, indicating that specific financial constraints are more informative than mere age and gender, but still cannot replace the explanatory power of psychological traits for irrational behavior.

5.2 Interaction Effect Analysis

Table 14 shows the interaction effect experimental results.

Table 14 Interaction Effect Experimental Results

Condition	Pure_Question2	BG_1_3	BG_1_4	BG_3_4
Accuracy	0.304	0.421	0.379	0.405
Stability (Agree_Rate)	0.843	0.861	0.852	0.858
QWK	-0.004	0.134	0.022	0.135
KL_Forward (Real‖Pred)	9.642	1.512	3.434	1.583
KL_Reverse (Pred‖Real)	1.108	0.298	0.499	0.291
KL_Bidirectional	10.75	1.809	3.933	1.874
JS_Divergence	0.43	0.122	0.205	0.121

Note: Evaluation of two-factor combinations. BG_1_3 (Demographics + Psychology) yields the best performance among combinations, approaching the full-information model, highlighting the critical role of psychological variables in economic simulation.

In multi-factor combinations, the BG_1_3 model containing demographic information and psychological information showed the best prediction effect, with the KL divergence indicator at the macro level dropping to 1.809, close to the 1.683 of the full information group. It is worth noting that the KL divergence of the BG_1_4 (demographic + financial) combination was as high as 3.933, significantly worse than combinations containing psychological features. This explains that pure demographic information, financial information, etc., cannot effectively simulate human decisions; human decisions are mainly influenced by psychological characteristics.

5.3 Psychological Factor Dismantling

Table 15 shows the ablation experiment results of introducing different psychological feature modules (Mod A-J) respectively on the basis of retaining demographic and financial backgrounds (BG_1+BG_4).

Table 15 Psychological Module Experimental Results

Condition	KL_Forward (Real‖Pred)	KL_Reverse (Pred‖Real)	KL_Bidirectional	JS_Divergence
BG_1_ModA_4	**1.1368**	**0.2434**	**1.3802**	**0.1010**
BG_1_ModB_4	2.8728	0.4201	3.2929	0.1721
BG_1_ModC_4	3.6834	0.4647	4.1481	0.1887
BG_1_ModD_4	1.8042	0.3623	2.1665	0.1503
BG_1_ModE_4	3.9292	0.4945	4.4238	0.2003
BG_1_ModF_4	**1.5572**	**0.2489**	**1.8061**	**0.1066**
BG_1_ModG_4	3.2702	0.3708	3.6410	0.1529
BG_1_ModH_4	2.9752	0.4471	3.4223	0.1809
BG_1_ModI_4	1.2871	0.3594	1.6465	0.1489
BG_1_ModJ_4	**1.5807**	**0.2739**	**1.8545**	**0.1153**

Note: Analysis of specific psychological sub-modules added to the base profile (BG_1+BG_4). Mod A represents Risk Perception/Attitude. Lower KL values indicate higher contribution to model fit. Risk perception and attitude features prove to be the most critical for simulating insurance decisions.

The data show that the model introducing "Risk Perception and Attitude" (Mod A) performed best in macro-distribution fitting, with its bidirectional KL divergence (KL_Bidirectional) dropping to 1.3802 and JS divergence (JS_Divergence) to 0.1010. In addition, "Intra-Household Dynamics" (Mod J) and "Social Influence" (Mod F) modules also showed lower divergence values, 1.8545 and 1.8061 respectively, indicating that these two types of features have high information gain when simulating decision distributions. By contrast, models introducing only "Self-Efficacy" (Mod D) or "Emotional Anticipation" (Mod E) had bidirectional KL divergences exceeding 2.0, indicating significant differences in the contribution of psychological features of different dimensions to decision simulation, with risk preference information playing a decisive role in restoring the insurance consumption distribution.

6. Further Research

In the previous analysis, we confirmed the effectiveness of complete user profiles based on natural language in simulating group decision distributions. However, the output of large language models relies heavily on the way prompts are constructed. To explore the deep influence mechanism of prompt engineering on simulation fidelity, this chapter will further launch in-depth research from two dimensions: "information processing strategy" and "observation context." We aim to answer: Is manual intervention on original information (such as theoretical guidance, summary compression) superior to raw input? And, can the model distinguish and simulate decision differences of humans under different social presence?

6.1 Influence Mechanism of Prompt Enhancement Strategies

To verify the influence of different prompt word enhancement strategies on model simulation capabilities, on the basis of the full information group (BG_1_3_4), we designed four sets of control experiments to examine the impact of "theoretical guidance," "explanatory constraints," and "information compression" on the prediction distribution.

The experimental groups were set as follows:

a. Pure Theory Guidance Group (Pure_With_Guidance): Adding behavioral finance theory and academic analysis without background information.

Prompt=[Question Test] + “Corresponding explanation for each question”

b. Full Information + Theory Guidance Group (BG_1_3_4_With_Guidance): Adding theoretical guidance on the basis of full background information.

Prompt=[Full_Context] + “Corresponding explanation for each question”

c. Full Information + Explanatory Constraint Group (BG_1_3_4_With_Explain): Requiring the model to explain questionnaire measurement content and results on the basis of full background information.

Prompt=[Full_Context] + “Corresponding explanation for questionnaire factors”

d. Information Summary Group (BG_1_SummarizedModules_4): Using LLM to summarize the psychological measurement part (Part 3) into a descriptive text instead of inputting the original questionnaire.

Prompt=[Part1_Demo]+[Summarize]+[Part4_context]+[Question_Test]

The experimental results are shown in Table 16.

Table 16 Prompt Enhancement Experimental Results

Condition	Accuracy	Stability	KL_Bidirectional	JS_Divergence
Pure_Question2	0.304	0.843	10.750	0.430
Pure_With_Guidance	0.381	0.958	12.593	0.414
BG_1_3_4_With_Guidance	0.413	0.865	2.200	0.153
BG_1_3_4_With_Explain	0.411	0.852	1.928	0.126
BG_1_SummarizedModules_4	0.380	0.873	5.488	0.245
BG_1_3_4	0.411	0.847	1.683	0.117

Note: Comparison of raw profile inputs against manually engineered prompts. With_Guidance: Adds theoretical explanations. Summarized: Uses LLM-compressed summaries of profiles. The degradation in performance for the "Summarized" group suggests that raw, high-granularity natural language descriptions provide richer constraints for the model than compressed or theoretically guided prompts.

Observing Table 16, we can draw the following conclusions: First, information compression led to significant degradation of fidelity. The bidirectional KL divergence of the information summary group (BG_1_SummarizedModules_4) soared from 1.683 in the full information group to 5.488, and JS divergence deteriorated to 0.245. This indicates that the process of summarizing unstructured text reduced the resolution and effective mutual information of background information, weakening the constraint force of the conditional probability space, causing the model posterior distribution to degenerate to a state close to the general prior.

Second, in scenarios with sufficient information, manual intervention may be counterproductive. After introducing "theoretical guidance" or "explanatory constraints," although individual micro-indicators Accuracy and QWK remained basically level, the KL divergence and JS divergence of macro-distribution consistency showed varying degrees of decline. We believe this stems from "prior mismatch." Existing classical behavioral theories cannot yet fully cover complex irrational human behaviors; forcing the model to follow specific theoretical logic is essentially introducing a structured prior that deviates from the real data distribution. When background information has already provided sufficient Bayesian constraints, additional instructional constraints will cause systematic re-weighting of the response distribution, thereby deviating from the real population distribution.

Finally, in scenarios with information scarcity, guidance strategies only improve micro-discriminability. Although the accuracy of the Pure_With_Guidance group increased from 0.304 to 0.381, its KL divergence actually worsened from 10.750 to 12.593. This shows that although the theoretical framework standardized the model's reasoning format, in the absence of individual feature input, the model was more prone to falling into certain specific theoretical biases, leading to further widening of the difference between the group distribution and the real world.

In summary, for group distribution simulation tasks, retaining high-granularity original natural language input is superior to prompt strategies that have undergone manual "optimization" or "compression."

6.2 Observation Effect: private vs public

Social science research has long faced the challenge of "Social Desirability Bias," where respondents tend to give answers that conform to social norms rather than their true inner thoughts.

To explore whether LLMs can simulate this difference in psychological fields, we introduced a "situation priming" mechanism, placing the same group of silicon-based individuals under two distinctly different observation modes:

a. Public Mode (BG_1_3_4_Public_Mode): The prompt explicitly injects the instruction "You are undergoing an insurance survey, please answer...", activating the sense of being observed.

b. Private Mode (BG_1_3_4_Private_Mode): The prompt emphasizes "No one is watching you. This is your true inner thought. Please decide based on your true financial anxiety...", simulating a psychological state without observation pressure.

6.2.1 Full Sample Analysis

The experimental results are shown in Table 17.

Table 17 Observation Mode Full Sample Analysis Results

Condition	BG_1_3_4	BG_1_3_4_Private_Mode	BG_1_3_4_Public_Mode
Accuracy	0.411	0.4121	0.4232
Stability (Agree_Rate)	0.847	0.8439	0.8542
QWK	0.135	0.1392	0.1278
KL_Forward (Real||Pred)	1.401	1.1268	1.6826
KL_Reverse (Pred||Real)	0.282	0.2583	0.3086
KL_Bidirectional	1.683	1.3851	1.9912
JS_Divergence	0.117	0.1054	0.1285

Note: Private Mode simulates a decision environment without social observation, while Public Mode implies a survey context. Private Mode achieves a slightly better distributional fit (lower KL/JS), suggesting it may better capture authentic preferences by mitigating social desirability bias.

The experimental results indicate that Private Mode is slightly superior to Public Mode in both individual prediction and group distribution. This reveals a profound mechanism: when the silicon-based agent is endowed with a context of "inner monologue" or "no observation pressure," it can more precisely restore the true cognitive state of humans when facing complex financial decisions. By contrast, the "sense of being gazed at" introduced by Public Mode may have caused the model to overfit social norms, thereby deviating from human real decision data.

6.2.2 Analysis of Decisions by Scenario

To rigorously quantify and compare the prediction effects of the two modes, this study adopted a paired sample design for testing. Table 18 reports the detailed statistical results. In the table, indicators are defined as follows: Acc. Private and Acc. Public represent prediction accuracy under Private Mode and Public Mode respectively; Difference δ [95% CI] represents the difference value of accuracy of Public Mode compared to Private Mode(i.e.,$\delta=\text{AccPublic}-\text{AccPrivate}$), with the 95% confidence interval of this difference value in brackets, used to measure the stability of the effect size; p-value is calculated based on the McNemar test, which is specifically used to assess the significance of differences between paired nominal data, to judge whether the prediction divergence of the model under the two contexts is statistically significant.

Table 18 Observation Mode Paired Sample Analysis Results

Condition	Acc. Private	Acc. Public	Difference(δ) [95% CI]	McNemar p
Total	0.417	0.410	-0.6% [-1.9%, +0.5%]	0.318

Q1	0.405	0.369	-3.6% [-6.6%, -0.5%]	0.033**
Q2	0.432	0.468	+3.6% [+1.0%, +6.1%]	0.009***
Q3	0.537	0.511	-2.6% [-5.8%, +0.5%]	0.124
Q4	0.352	0.349	-0.3% [-3.1%, +2.5%]	0.852
Q5	0.359	0.354	-0.5% [-3.5%, +2.5%]	0.784

Note: Statistical comparison of accuracy between Private and Public modes. Difference (δ) is calculated as $Acc_{Public} - Acc_{Private}$. 95% CI denotes the confidence interval of the difference. McNemar p tests the significance of the divergence. Results show distinct "private" (e.g., Q1) vs. "public" (e.g., Q2) preference patterns consistent with psychological theories. *p<0.05, **p<0.01, *** p < 0.001..

As shown in Table 18, although the difference between the two modes did not reach statistical significance at the full sample level (p=0.318), the model showed complex hierarchy consistent with human psychological mechanisms in specific decision sub-tasks.

We found that instinctual preferences such as intertemporal choice are more suitable for private processing. The data shows that the accuracy of Private Mode on the intertemporal choice task was significantly higher than Public Mode (Difference = -3.6%, p<0.05). Intertemporal choice essentially involves the individual's impulse control mechanism. Public survey contexts often activate respondents' social norm preferences regarding "delayed gratification," leading the model to tend to overestimate the individual's patience level. Conversely, Private Mode effectively avoided this normative deviation, more precisely capturing the "Hyperbolic Discounting" characteristics of real individuals in the absence of supervision, thereby restoring more short-sighted but real instinctual preferences.

Normative preferences involving product payment are more suitable for socialized processing. Public Mode performed significantly better than Private Mode on insurance willingness to pay (Difference(δ\deltaδ) = +3.6%, p<0.01). Insurance purchase is usually viewed as a responsible social contract behavior, embedded within the framework of family responsibility and social expectations. Public Mode successfully induced the model to incorporate social norms into the operation of the utility function, thereby more accurately reconstructing the real-world scenario where consumers increase their hedging willingness due to perceived social responsibility or family pressure.

7. Conclusion

This paper proposes and evaluates a zero-shot insurance decision simulation approach that treats an off-the-shelf large language model (LLM) as a distribution-level simulator. Conditioning on respondent-specific natural-language profiles substantially improves distributional fidelity: relative to prompting with the question alone, full-profile prompting reduces bidirectional KL divergence from 10.750 to 1.683 and JS divergence from 0.430 to 0.117. We interpret the improvement as profile-based conditioning that shifts the model's output distribution toward the human benchmark under pre-specified metrics. Compared with demographic information and financial information, psychological traits including risk perception, family dynamics and social impact showed stronger simulation power in our sample.

Furthermore, we found that retaining high-granularity original natural language narratives is superior to inputs that have undergone manual theoretical guidance or summary compression, indicating that in information-sufficient contexts, excessive manual intervention will instead

introduce inductive bias and weaken the model's adaptive reasoning ability. Observation effect experiments also showed that "Private Mode" can better restore the true state of individuals in instinctual decisions, such as intertemporal choice, compared to the "Public Mode" by eliminating social desirability bias. This provides a new methodological path for using LLMs to solve the "stated preference bias" problem in traditional questionnaire surveys.

Although the model performed excellently on tasks such as intertemporal choice and willingness to pay, it performed poorly in scenarios involving highly irrational emotions (regret theory) and high cognitive loads (multi-attribute trade-offs). We believe there may be two reasons: on the one hand, current large language models still have capability boundaries when simulating human extreme irrational emotions and complex logical trade-offs; on the other hand, the natural language features we input could not provide sufficient information for simulating this class of problems.

In summary, this study not only provides a low-cost, scalable simulation tool for policy evaluation and market deduction in data-scarce contexts, but also provides an empirical basis for understanding the application potential and boundaries of generative AI in social science research. Future research can further explore breaking through current simulation bottlenecks in complex trade-off tasks through Chain-of-Thought (CoT) fine-tuning or Multi-Agent Interaction.

References

Adamopoulos P, Ghose A, Tuzhilin A (2022) Heterogeneous demand effects of recommendation strategies in a mobile application: Evidence from econometric models and machine-learning instruments. MIS Quart. 46(1):101–150.

Alantari HJ, Currim IS, Deng Y, Singh S (2022) An empirical comparison of machine learning methods for text-based sentiment analysis of online consumer reviews. Internat. J. Res. Marketing 39(1):1–19.

Ale Ebrahim Dehkordi M, Lechner J, Ghorbani A, et al. (2023) Using machine learning for agent specifications in agent-based models and simulations: A critical review and guidelines. J. Artificial Societies Social Simulation 26(1).

Anand P, Lee C (2023) Using deep learning to overcome privacy and scalability issues in customer data transfer. Marketing Sci. 42(1):189–207.

Anderson LR, Freeborn BA, Hulbert JP (2018) Behavioral factors in equity allocation decisions: A large-scale experimental study with context. J. Behavioral Finance 19(3):334–348.

Argyle LP, Busby EC, Fulda N, Gubler JR, Rytting C, Wingate D (2023) Out of one, many: Using language models to simulate human samples. Political Analysis 31(3):337–351.

Audia PG, Martinez DL, Newark DA, et al. (2024) Decision making and problem solving in organizations: Assessing and expanding the Carnegie perspective. Frontiers in Psych. 15:1455794.

Axtell RL, Farmer JD (2025) Agent-based modeling in economics and finance: Past, present, and future. J. Econom. Lit. 63(1):197–287.

Bailey JJ, Kinerson C (2005) Regret avoidance and risk tolerance. J. Financial Counseling Planning 16(1).

Beck MJ, Fifer S, Rose JM (2016) Can you ever be certain? Reducing hypothetical bias in stated choice experiments via respondent reported choice certainty. Transportation Res. Part B: Methodological 89:149–167.

Binz M, Schulz E (2023) Using cognitive psychology to understand GPT-3. Proc. National Acad. Sci. 120(6):e2218523120.

Bisaria C, et al. (2025) The evolution of marketing strategies in the age of AI: A review of adaptive and predictive models. J. Marketing Social Res. 2:213–222.

Brand J, Israeli A, Ngwe D (2023) Using LLMs for market research. Working paper, Harvard Business School, Boston.

Bray JP (2009) Consumer Behaviour Theory: Approaches and Models (Bournemouth University, Bournemouth, UK).

Buehler P, Maas P (2018) Consumer empowerment in insurance: Effects on performance risk perceptions in decision making. Internat. J. Bank Marketing 36(6):1073–1097.

Bursztyn L, Gonzalez A, Misra S, et al. (2025) The social desirability atlas. Working paper, National Bureau of Economic Research, Cambridge, MA.

Chakraborty I, Kim M, Sudhir K (2022) Attribute sentiment scoring with online text reviews: Accounting for language structure and missing attributes. J. Marketing Res. 59(3):600–622.

Csaszar FA, Ketkar H, Kim H (2024) Artificial intelligence and strategic decision-making: Evidence from entrepreneurs and investors. Strategy Sci. 9(4):322–345.

Das S, et al. (2023) Credit risk modeling with graph machine learning. INFORMS J. Data Sci. 2(2):197–217.

Dew R, Ansari A, Toubia O (2022) Letting logos speak: Leveraging multiview representation learning for data-driven branding and logo design. Marketing Sci. 41(2):401–425.

Dikmen I, Budayan C, Talat Birgonul M, et al. (2018) Effects of risk attitude and controllability assumption on risk ratings: Observational study on international construction project risk assessment. J. Management Engineering 34(6):04018037.

Dustmann C, Fasani F, Meng X, et al. (2023) Risk attitudes and household migration decisions. J. Human Resources 58(1):112–145.

Ejnarsson Lundberg A (2025) Beyond the Hype Augmenting Knowledge Work with Agentic AI: A Service Design Approach to Enterprise Transformation.

Estelami H (2014) An ethnographic study of consumer financial sophistication. J. Consumer Behaviour 13(5):328–341.

Estelami H, Liu K (2024) Content analysis of American consumers' credit card fraud complaints filed with the Consumer Financial Protection Bureau. J. Financial Crime 31(3):618–628.

Faraj S, et al. (2025) A time for monsters: Organizational knowing after LLMs. arXiv preprint arXiv:2511.15762.

Fisher RJ (1993) Social desirability bias and the validity of indirect questioning. J. Consumer Res. 20(2):303–315.

Flavin M, Yamashita T (2002) Owner-occupied housing and the composition of the household portfolio. Amer. Econom. Rev. 92(1):345–362.

Gallagher J (2014) Learning about an infrequent event: Evidence from flood insurance take-up in the United States. Amer. Econom. J.: Applied Econom. 6(3):206–233.

Goli A, Singh A (2024) Frontiers: Can large language models capture human preferences? Marketing Sci. 43(4):709–722.

Grewal D, et al. (2025) How generative AI is shaping the future of marketing. J. Acad. Marketing Sci. 53(3):702–722.

Gropper MJ, Kuhnen CM (2025) Wealth and insurance choices: Evidence from US households. J. Finance 80(2):1127–1170.

Halek M, Eisenhauer JG (2001) Demography of risk aversion. J. Risk Insurance 68(1):1–24.

Hambel C, Kraft H, Schendel LS, et al. (2017) Life insurance demand under health shock risk. J. Risk Insurance 84(4):1171–1202.

Hartmann J, Exner Y, Domdey S (2025) The power of generative marketing: Can generative AI create superhuman visual marketing content? Internat. J. Res. Marketing 42(1):13–31.

Hartmann J, Netzer O (2023) Natural language processing in marketing. Ma L, Sun B, eds. Artificial Intelligence in Marketing (Emerald Publishing Limited, Bingley, UK), 191–215.

Hauser JR, Li Z, Mao C (2023) Artificial intelligence and user-generated data are transforming how firms come to understand customer needs. Ma L, Sun B, eds. Artificial Intelligence in Marketing (Emerald Publishing Limited, Bingley, UK), 147–167.

Hong Chew S, Yi J, Zhang J, et al. (2018) Risk aversion and son preference: Experimental evidence from Chinese twin parents. Management Sci. 64(8):3896–3910.

Horton JJ (2023) Large language models as simulated economic agents: What can we learn from homo silicus? Working paper, National Bureau of Economic Research, Cambridge, MA.

Hu Y, Tian X (2025) Evaluating reasoning large language models on rumor generation, detection, and debunking tasks. iScience 28(11).

Hubbard RG, Skinner J, Zeldes SP (1995) Precautionary saving and social insurance. J. Political Economy 103(2):360–399.

Jiang S, Xie M, Luo J (2025) Large language models for combinatorial optimization of design structure matrix. Proc. Design Society 5:2201–2210.

Johnson MD, Puto CP (1987) A review of consumer judgment and choice. Rev. Marketing 1987:236–292.

Kaplan G, Moll B, Violante GL (2018) Monetary policy according to HANK. Amer. Econom. Rev. 108(3):697–743.

Kaplan G, Violante GL (2018) Microeconomic heterogeneity and macroeconomic shocks. J. Econom. Perspectives 32(3):167–194.

Keillor BD, Owens D, Pettijohn C (2001) A cross-cultural/cross national study of influencing factors and socially desirable response biases. Internat. J. Market Res. 43(1):1–19.

Kelly N, Baker ST, Bauer JW, et al. (2017) An examination of social desirability bias in measures of college students' financial behavior. College Student J. 51(1):115–128.

King MF, Bruner GC (2000) Social desirability bias: A neglected aspect of validity testing. Psych. Marketing 17(2):79–103.

Kirman AP (1992) Whom or what does the representative individual represent? J. Econom. Perspectives 6(2):117–136.

Kitadai A, et al. (2025) Can AI with high reasoning ability replicate human-like decision making in economic experiments? Group Decision Negotiation 1–24.

Ko A (2022) An equilibrium analysis of the long-term care insurance market. Rev. Econom. Stud. 89(4):1993–2025.

Korinek A (2023) Generative AI for economic research: Use cases and implications for economists. J. Econom. Lit. 61(4):1281–1317.

Kusev P, Purser H, Heilman R, et al. (2017) Understanding risky behavior: The influence of cognitive, emotional and hormonal factors on decision-making under risk. Frontiers in Psych. 8:102.

Lee D, Manzoor E, Cheng Z (2022) Focused concept miner (FCM): Interpretable deep learning for text exploration. SSRN 4280590.

Li Z, Liu L, Shi J, et al. (2021) Health insurance, risk attitudes, and household financial behavior. Health Econom. 30(5):1239–1246.

Lieber EMJ, Skimmyhorn W (2018) Peer effects in financial decision-making. J. Public Econom. 163:37–59.

Lim HN, Shin SH, Wilmarth MJ, et al. (2022) Who decides? Financial decision-making among older couples. J. Family Econom. Issues 43(2):310–337.

List JA, Gallet CA (2001) What experimental protocol influence disparities between actual and hypothetical stated values? Environmental Resource Econom. 20(3):241–254.

Liu X, Singh PV, Srinivasan K (2016) A structured analysis of unstructured big data by leveraging cloud computing. Marketing Sci. 35(3):363–388.

Lucas RE Jr (1976) Econometric policy evaluation: A critique. Carnegie-Rochester Conf. Ser. Public Policy 1:19–46.

Ma L, Sun B (2020) Machine learning and AI in marketing–Connecting computing power to human insights. Internat. J. Res. Marketing 37(3):481–504.

Manski CF (2004) Measuring expectations. Econometrica 72(5):1329–1376.

Markowitz HM (2010) Portfolio theory: As I still see it. Annu. Rev. Financ. Econ. 2(1):1–23.

McCallum BT (2010) Is the spurious regression problem spurious? Econom. Lett. 107(3):321–323.

Murphy JJ, Stevens TH, Weatherhead D (2005) A meta-analysis of hypothetical bias in stated preference valuation. Environmental Resource Econom. 30(3):313–325.

Park JS, Zou CQ, Shaw A, et al. (2024) Generative agent simulations of 1,000 people. arXiv preprint arXiv:2411.10109.

Parker L, Hassan O, Davis C (2025) The role of generative AI in the future of commercial personalization. J. Electronic Commerce 1(1):1–16.

Quevedo FJ, Gopalakrishna P (2021) Rationality is overrated: Brand choice is largely intuitive. Rutgers Bus. Rev. 6(3):312–332.

Robertson C, Safta C, Collier N, et al. (2025) Econometric policy evaluation: A critique. Statist. Medicine 44(6):e70029.

Rytting CM (2023) Algorithmic fidelity and the use of large language models in social science research. Unpublished doctoral dissertation, Brigham Young University, Provo, UT.

Sandroni A, Squintani F (2007) Overconfidence, insurance, and paternalism. Amer. Econom. Rev. 97(5):1994–2004.

Schinckus C (2019) Agent-based modelling and economic complexity: A diversified perspective. J. Asian Bus. Econom. Stud. 26(2):170–188.

Serapio-García G, Safdari M, Crepy C, et al. (2025) A psychometric framework for evaluating and shaping personality traits in large language models. Nature Machine Intelligence (Forthcoming).

Shin D, He S, Lee GM, et al. (2020) Enhancing social media analysis with visual data analytics: A deep learning approach. MIS Quart. 44(4):1459–1492.

Shmueli G, Koppius OR (2011) Predictive analytics in information systems research. MIS Quart. 35(3):553–572.

Sims CA (1980) Macroeconomics and reality. Econometrica 48(1):1–48.

Taylor SA (2012) Evaluating digital piracy intentions on behaviors. J. Services Marketing 26(7):472–483.

Toubia O, Gui GZ, Peng T, et al. (2025) Database report: Twin-2k-500: A data set for building digital twins of over 2,000 people based on their answers to over 500 questions. Marketing Sci. 44(6):1446–1455.

van Rooij M, Alessie R, Lusardi A (2024) Financial literacy in the DNB Household Survey: Insights from innovative data collection. J. Financial Literacy Wellbeing 1–17.

Vaswani A, Shazeer N, Parmar N, et al. (2017) Attention is all you need. Proc. Advances in Neural Information Processing Systems 30.

Wei J, Wang X, Schuurmans D, et al. (2022) Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems 35:24824–24837.

Xiao JJ, Tao C (2021) Consumer finance/household finance: The definition and scope. China Finance Rev. Internat. 11(1):1–25.

Ye Z, Yoganarasimhan H, Zheng Y (2025) Lola: LLM-assisted online learning algorithm for content experiments. Marketing Sci. (Forthcoming).

Yogo M (2016) Portfolio choice in retirement: Health risk and the demand for annuities, housing, and risky assets. J. Monetary Econom. 80:17–34.