# Prospect Theory Fails for LLMs: Revealing Instability of Decision-Making under Epistemic Uncertainty

Anonymous ACL submission


## Abstract

Prospect Theory (PT) models the human decision-making tendency under uncertainty. While recent studies have developed some questionnaires to elicit the PT parameters (features evaluating decision-making tendency) to describe decision behavior of Large Language Models (LLM), many of them did not report the performance (or explanatory power) of PT itself for LLMs. Additionally, although PT has been used in many LLM-related fields, few studies have tried to test its robustness under linguistic uncertainty, especially epistemic markers (e.g. "maybe"). To address these research gaps, we design an experiment workflow. We adopt a classic economic questionnaire and perform parameter estimation with performance metrics (e.g. McFadden $R^2$). We further let LLMs make binary choices which reflect the internal probability values of epistemic markers. We then incorporate epistemic markers into the questionnaire to examine the robustness of Prospect Theory parameters. Our findings suggest that modeling LLMs' decision-making with PT is not consistently reliable, and applying Prospect Theory to LLMs is likely not robust under epistemic uncertainty. [1]


## 1 Introduction

LLM is increasingly used in important decision-making tasks such as healthcare and finance (Keith and Stent, 2019; Lehman et al., 2022). While human decision-making under uncertainty has been extensively studied through both normative and descriptive frameworks, the decision patterns for LLMs remain underexplored. Among the human psychological model used in LLM field, Prospect Theory (PT) (Kahneman and Tversky, 1979; Tanaka et al., 2010; Rathi et al., 2025) stands out as particularly influential. It models a range of human behavioral patterns by introducing psychologically grounded parameters, such as loss aversion and probability weighting (P. and H., 2005; Gneezy et al., 2006; Rathi et al., 2025). Since LLM has gained more and more attention, Prospect Theory continuous to play an important role in its training, testing, and alignment (Cheng et al., 2025; Wang et al., 2025). However, although Prospect Theory is intensively used for LLMs, few studies report the performance or explanatory power of this human psychological model when

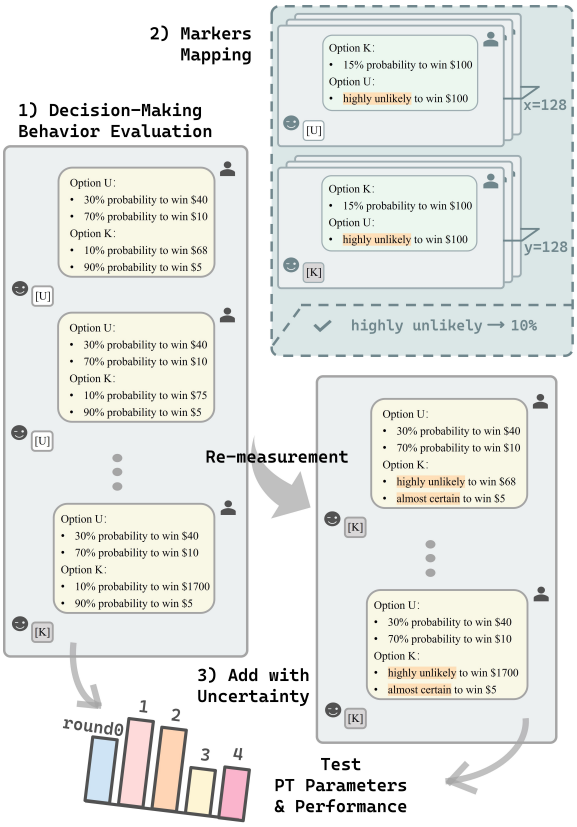


Figure 1: **An overview of our three-stage experiment.** Stage 1 fits PT parameters from binary choices with precise probabilities. Stage 2 estimates each markers internal probability from the point where options K and U are chosen equally. Stage 3 substitutes probabilities with markers to measure the effect of linguistic uncertainty markers.

[1] We will release our code and data upon acceptance.

applying it to LLM.

In addition, decision-making is a highly safety-critical task (Leyli-abadi et al., 2025), so we must ensure the robustness of explanation theory such as PT under uncertainty. Existing studies have examined how personality prompting, sociodemographic embedding, role-playing interventions, or post-training can systematically shift LLM's performance (Jia et al., 2024; Liu et al., 2025a; Wang et al., 2025), but whether and how epistemic uncertainty markers (Lee et al., 2025) shift LLM decision-making behaviour described under Prospect Theory remains underexplored.

Human uncertainty expressions are commonly expressed via epistemic markers to support flexible language use (Belem et al., 2024; Liu et al., 2025b; Zhou et al., 2023). Despite their ubiquity, these expressions are inherently ambiguous (Agata, 2017), and correctly interpreting them is essential for effective decision-making (Mahmoodi et al., 2020). To examine how LLMs handle decision-making under uncertainty, we design a lottery-based evaluation adapted from classic behavioral economics (Allais, 1953; Brandstätter et al., 2006; Tanaka et al., 2010). Our evaluation consists of three stages, illustrated in Figure 1. We begin by presenting binary choice tasks with precise numerical probabilities to estimate each models Prospect Theory parameters. In the second stage, we let LLM to make binary choices between a numerical probability value and a probability described by epistemic markers. This allows us to infer the models implicit probabilistic interpretation of those uncertainty expressions. Finally, we reassess model behavior on the original decision tasks, now framed using epistemic markers grounded in their inferred probability values, to directly evaluate the impact of epistemic markers on decision-making.

Our results reveal two key insights. First, Prospect Theory does not consistently performs well at explaining LLM decision behaviors, and it exhibits model-wise differences where larger scale models outperform the others. Second, introducing epistemic markers disrupts decision consistency and alters PT parameters, exposing the fragility of LLMs' decision-making under linguistic uncertainty. These findings point to fundamental gaps in the interpretability and robustness of current LLMs when processing ambiguous language, underscoring the need for improved epistemic calibration and more nuanced theoretical frameworks for modeling LLM behavior in uncertain contexts.

## 2 Related Work

**Prospect Theory in Economic Decision-Making.** Prospect Theory (Kahneman and Tversky, 1979) has long served as a foundational framework for modeling human decision-making under risk, capturing key behavioral patterns like loss aversion and probability distortion. Empirical studies extend PT to diverse populations and settings (Tanaka et al., 2010), and recent work explores whether LLMs exhibit similar patterns (Jia et al., 2024). Our work builds on these efforts but focuses on how language-based uncertainty impacts PT-consistent behavior in LLMs.

**LLM Decision-Making under Uncertainty.** Recent work evaluates how LLMs handle uncertain scenarios, including economic games, moral dilemmas, and ambiguous instructions (Liu et al., 2025a; Jia et al., 2024; Zong et al., 2025; Zheng et al., 2025d). These studies often probe whether LLMs mimic human cognitive biases or align with normative models. We extend this line of inquiry by testing whether linguistic uncertainty conveyed by epistemic markers causes instability of LLM behaviors in decision-making situations.

**Epistemic Markers in NLP.** Recent work investigates how LLMs interpret and respond to linguistic signals related to uncertainty and confidence. Zhou et al. (2023) study how different epistemic markers embedded in prompts affect model predictions. Belem et al. (2024) evaluate LLMs' interpretation of epistemic markers, finding partial human-like behavior but systematic biases. Liu et al. (2025b) further argue that such markers are often unreliable indicators of internal confidence in LLMs. These findings suggest that models may mimic surface linguistic patterns rather than demonstrate true epistemic reasoning. Our study builds on this line of work by examining how LLMs understand epistemic markers in economic contexts with uncertainty.

## 3 Preliminary

In rational decision theory, an agent's preference over risky prospects follows the *von Neumann-Morgenstern expected utility framework*. For a prospect $P = (x_1, p_1; \cdots; x_n, p_n)$ yielding outcome $x_i$ with probability $p_i$, the *expected utility* is

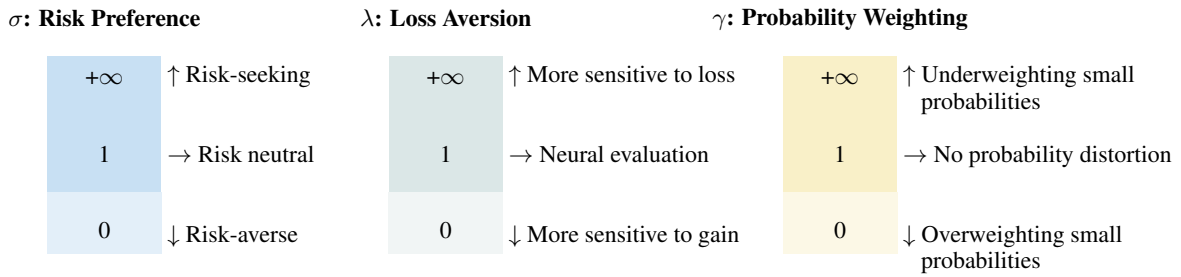


Figure 2: **Visual illustration of the three PT parameters ($\sigma$, $\lambda$, $\gamma$).** Each parameter is shown with its meaning and an interpretation of its directional significance.

computed as:

$$EU(P) = \sum_{i=1}^{n} p_i \cdot u(x_i), \tag{1}$$

where $u(\cdot)$ is a cardinal utility function mapping outcomes to real numbers. Human decisions are assumed to maximize $EU(P)$ under traditional Expected Utility Theory (EUT), given by von Neumann et al. (1944). However, empirical evidence systematically violates EUT assumptions. Prospect Theory (PT) addresses these anomalies through three psychological distortions of rational utility. It maintains a utility-based approach but fundamentally alters Expected Utility Theory through three key properties:

- **Risk Preference** ($\sigma$): Agents often exhibit varying degrees of risk aversion or risk-seeking behavior.
- **Loss Aversion** ($\lambda$): Losses psychologically outweigh equivalent gains.
- **probability weighting** ($\gamma$): Agents often exhibit systematic probability distortion.

To capture these characteristics, Prospect Theory introduce two specialized functions: the *value function* formalizes how the outcomes translate into subjective utility, while the *probability weighting function* captures non-linear probability perception. Together, these functions model the distorted utility calculations that characterize PT decision-making.

The **value function** v(x) quantifies subjective satisfaction from outcomes relative to a reference point (zero in this study). For outcomes $x \geq 0$ (gains) and $x < 0$ (losses), the value function is:

$$v(x) = \begin{cases} x^{\sigma} & \text{for } x \geq 0 \\ -\lambda(-x)^{\sigma} & \text{for } x < 0. \end{cases} \tag{2}$$

Key parameters here are loss aversion ($\lambda$) which is a multiplier of negative perceptions, and risk preference ($\sigma$) which controls curvature (sensitivity to values).

The **probability weighting function** formalizes the transformation of objective probabilities $p$ in subjective decision weights:

$$w(p) = \frac{p^{\gamma}}{(p^{\gamma} + (1-p)^{\gamma})^{1/\gamma}}, \tag{3}$$

where $\gamma$ controls the curvature of the function.

The final utility for any binary prospect in the form $P = (x, p; y, q)$ is defined as follows:

$$u(P) = \begin{cases} v(y) + w(p)(v(x) - v(y)) & [1] \\ w(p)v(x) + w(q)v(y) & [2], \end{cases} \tag{4}$$

where [1] is "if $x > y > 0$ or $x < y < 0$" and [2] is "if $x < 0 < y$".

All functions follows the standards set in (Kahneman and Tversky, 1979).

## 4 Decision-Making Behavior Evaluation

Based on the above economic frameworks, we mainly adopt the three-series lottery-choice experiment developed by Tanaka et al. (2010) to get a reliable PT parameter measurement. Series 1 and 2 are designed to elicit risk preference ($\sigma$) and probability weighting ($\gamma$), while series 3 is designed for loss aversion ($\lambda$). The prospect settings are shown in Appendix B, and the prompt design is in Appendix C.

Each lottery consists of two option: a relatively safer option K with lower risk, and a riskier option U. The agent is asked to directly choose from option K and U based on its risk preference.

After sampling 256 times for each question, we count the portion of choosing option K for each lottery. Then we define the predicted probability

of choosing option K for each lottery as follows:

$$P(\text{choose K}) = \frac{e^{\text{EUDIFF}}}{1 + e^{\text{EUDIFF}}}, \quad (5)$$

where EUDIFF $= u(\text{K}) - u(\text{U})$ is the difference in the distorted utility of prospect K and U under prospect theory. We choose the sigmoid function as it is a usual setting in economic studies, e.g., Chakravarty and Roy (2009). We add up the Bernoulli log-likelihood for all 35 lotteries as the negative log-likelihood function, and run MLE with this function to estimate $\sigma$, $\lambda$ and $\gamma$.

To get **confidence intervals**, we use a bootstrap method (Efron, 1979) by generating simulated datasets through binomial sampling from predicted probabilities derived from the original parameter estimates. Specifically, for each observation $i$, we sample $\tilde{y}_i \sim \text{Binomial}(n = 1, p = \hat{p}_i)$ where $\hat{p}_i$ is the predicted probability. The model parameter standard deviation $\sigma_{\hat{\theta}}$ is estimated from the bootstrap distribution, and the 95% confidence interval is constructed using the percentile method:

$$\text{CI}_{95\%} = \left[\hat{\theta}^*_{(0.025)}, \hat{\theta}^*_{(0.975)}\right], \quad (6)$$

where $\hat{\theta}^*_{(\alpha)}$ denotes the $\alpha$-quantile of the bootstrap parameter estimates. This approach accounts for parameter uncertainty in finite samples.

We then quantify model goodness-of-fit by computing **McFadden pseudo-$\mathbf{R^2}$** (McFadden, 1977), defined as:

$$R^2_{\text{McFadden}} = 1 - \frac{\mathcal{L}_{\text{PT}}}{\mathcal{L}_{\text{null}}}, \quad (7)$$

where $\mathcal{L}_{\text{PT}}$ is the log-likelihood of Prospect Theory and $\mathcal{L}_{\text{null}}$ represents the log-likelihood of the intercept-only model with uniform choice probabilities. This metric measures the improvement of our model over random guessing.

Finally, we calculate the **mean absolute error** (MAE) between the actual probability $p_{\text{actual}}$ and the predicted probability $p_{\text{pred}}$ of choosing option K derived from our PT model:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} \left| p^{(i)}_{\text{actual}} - p^{(i)}_{\text{pred}} \right|, \quad (8)$$

where $N$ denotes the number of observations. This provides a direct measure of prediction error.

Overall, the three PT parameters, their confidence intervals, the McFadden pseudo-$R^2$, and the MAE score together describe the revealed agent's risk attitude and our model's explanatory power itself.

## 5 Decision-making Scenarios with Uncertainty

Real world decisions are often made under vague linguistic epistemic uncertainty rather than precise numerical probability. This urges for experiments on how epistemic markers influence LLM's decision-making behaviour. We investigate how decision-making is affected when numerical probabilities are replaced by verbal probability expressions, or epistemic markers. Section 5.1 estimates their numerical equivalents via a controlled lottery experiment, and Section 5.2 applies these values in the PT framework to re-measure PT parameters. The prompt design is in Appendix C.

### 5.1 Probability Mapping of Epistemic Markers

Epistemic markers are inherently vague and context-sensitive (Liu et al., 2025b; Bergqvist, 2015), yet they often substitute for precise numerical probabilities in practice. For LLMs, the ability to interpret epistemic markers consistently and meaningfully is critical if they are used as decision-support tools. However, there is little empirical understanding of how LLMs map epistemic markers to numerical probabilities, and whether this mapping is coherent across different models or aligned with human intuition.

| No. | Epistemic Marker | Probability Mapping by Human |
|---|---|---|
| 1 | *almost certain* | 95% |
| 2 | *highly likely* | 90% |
| 3 | *very likely* | 90% |
| 4 | *likely* | 80% |
| 5 | *probable* | 70% |
| 6 | *somewhat likely* | 70% |
| 7 | *possible* | 60% |
| 8 | *uncertain* | 50% |
| 9 | *somewhat unlikely* | 30% |
| 10 | *unlikely* | 25% |
| 11 | *not likely* | 20% |
| 12 | *doubtful* | 20% |
| 13 | *very unlikely* | 10% |
| 14 | *highly unlikely* | 10% |

Table 1: **Epistemic markers used in the experiment.** Human probability mapping comes from Belem et al. (2024)

To address this issue, we design a controlled lottery experiment in an economic decision-making context. Each trial presents the model with a

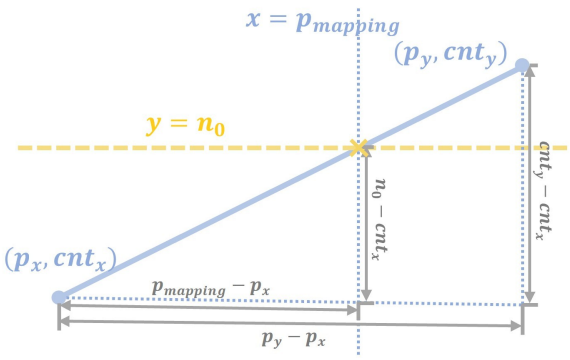


Figure 3: **An illustration of calculating** $p_{mapping}$**.** We use linear interpolation for fitting and calculate $p_{mapping}$ through the similarity ratio.

choice between two options. For option K, there is a fixed probability $p\%$ of winning $\$M$, while $p$ traverses all values in set $probs$ (see Appendix A.1). For option U, there is an unknown possibility of winning $\$M$ which is defined by the different markers. For the usage of epistemic markers, we manually select 14 commonly used markers from previous work in Belem et al. (2024) to ensure that they are suitable for this context (see Table 1).

For each marker, we record the number of times the model selects option K, denoted as $\text{NUM}_K$. The key idea is: when $\text{NUM}_K$ reaches $p_0 = 50\%$ of the total trials for a given probability $p$, the model considers the two options equally attractive. We define the mapping probability $p_{mapping}$ for that marker as the value of $p$ at which this occurs:

$$p_{mapping} = p. \tag{9}$$

Since $p$ is sampled at discrete points, the exact $p_0$ point may fall between two sampled probabilities. To estimate $p_{\text{mapping}}$, we perform linear interpolation between the two nearest points. Let $n_0 = p_0 \cdot sample_number$ be the target count corresponding to $p_0$ selection of option K, $(p_x, cnt_x)$ be the probability-count pair just **below** $n_0$, and $(p_y, cnt_y)$ be the pair just **above** $n_0$. As illustrated in Figure 3, by the slope formula

$$\frac{p_{mapping} - p_x}{p_y - p_x} = \frac{n_0 - cnt_x}{cnt_y - cnt_x}, \tag{10}$$

we solve for $p_{mapping}$:

$$p_{mapping} = \frac{(n_0 - cnt_x) \cdot p_y + (cnt_y - n_0) \cdot p_x}{cnt_y - cnt_x}. \tag{11}$$

Through this experiment, we obtain a list of 14 probability values (one per marker) for each model, capturing how it semantically interprets verbal uncertainty in economic terms.

| Model | $\sigma$ | $\lambda$ | $\gamma$ | MAE↓ | $R^2$↑ |
|---|---|---|---|---|---|
| *Human* | 0.670 | 2.630 | 0.685 | - | - |
| *Llama-3.1-8B-Instruct* | 0.585 | 0.010 | 0.753 | 0.332 | 0.092 |
| *Mistral-7B-Instruct-v0.3* | 0.534 | 0.570 | 0.577 | **0.155** | **0.132** |
| *Qwen2.5-7B-Instruct* | 0.429 | 0.010 | 3.645 | **0.047** | **0.116** |
| *Qwen2.5-14B-Instruct* | 0.503 | 1.909 | 0.896 | 0.257 | 0.067 |
| *Qwen2.5-32B-Instruct* | 0.598 | 1.213 | 0.867 | **0.161** | **0.225** |

Table 2: **Baseline PT parameter estimation across LLMs.** Human reference values are taken from Tanaka et al. (2010). Most LLMs exhibit human-like risk aversion and probability distortion, but show inconsistent and generally weaker loss aversion. We **bold** the results where MAE $\leq 0.20$ and $R^2 \geq 0.10$, indicating that PT provides a reliable fit.

### 5.2 Re-measurement of Prospect Theory Parameters

After we have established the probability mapping, we pick a pair of epistemic markers which has a close normalized probability to the original settings. This is to ensure that the probability setting does not change much. Detailed replacement rules are in Appendix D. For example, for a model with *"somewhat unlikely"* mapped to $p_1 = 32\%$ and *"highly likely"* mapped to $p_2 = 68\%$, we normalize them using the formula:

$$p_1' = \frac{p_1}{p_1 + p_2}, \quad p_2' = \frac{p_2}{p_1 + p_2} \tag{12}$$

to ensure they add up to $100\%$. Then we replace the probabilities in the original decision-making behavior evaluation framework with the closest epistemic marker pairs. We re-run the PT metrics measurement test using $p_1'$ and $p_2'$ and compare the result with the original study. There are four rounds in our study. We control the option to which epistemic markers are introduced. In round 1, we only introduce markers in option K, series 1 and 2. In round 2, all 3 series of option K are with epistemic markers. In round 3, all option U are with markers. In round 4, both option K and option U of all series are with epistemic markers.

## 6 Results and Findings

### 6.1 Risk Preference under Numerical Probabilities

For decision-making under exact probability, we obtain different PT parameters compared with previous works by Jia et al. (2024) and Liu et al.

| Model | Top 7 Epistemic Markers | | | | | | |
|---|---|---|---|---|---|---|---|
| | **almost certain** | **highly likely** | **very likely** | **likely** | **probable** | **somewhat likely** | **possible** |
| *Llama-3.1-8B-Instruct* | 87.92 | 56.04 | 58.00 | 41.80 | 44.23 | 36.71 | 36.29 |
| *Mistral-7B-Instruct-v0.3* | 96.80 | 67.89 | 63.10 | 57.50 | 87.22 | 48.98 | 52.16 |
| *Qwen2.5-7B-Instruct* | 82.71 | 67.00 | 67.06 | 4.78 | 3.44 | 8.93 | 4.51 |
| *Qwen2.5-14B-Instruct* | 91.56 | 55.00 | 54.10 | 42.38 | 26.51 | 32.47 | 38.38 |
| *Qwen2.5-32B-Instruct* | 97.50 | 95.08 | 82.82 | 65.00 | 54.74 | 46.08 | 55.00 |

Table 3: **Switching probabilities (%) for top 7 epistemic markers across models.** Models differ substantially in how they interpret high-certainty markers such as almost certain or very likely, with mapping values varying by over 50 percentage points across models, indicating inconsistent internal representations of certainty.

| Model | Bottom 7 Epistemic Markers | | | | | | |
|---|---|---|---|---|---|---|---|
| | **uncertain** | **somewhat unlikely** | **unlikely** | **not likely** | **doubtful** | **very unlikely** | **highly unlikely** |
| *Llama-3.1-8B-Instruct* | 35.49 | 33.71 | 33.49 | 36.91 | 33.65 | 31.30 | 32.83 |
| *Mistral-7B-Instruct-v0.3* | 48.08 | 40.15 | 38.27 | 30.93 | 34.03 | 29.47 | 27.88 |
| *Qwen2.5-7B-Instruct* | 35.58 | 27.24 | 19.90 | 27.70 | 25.69 | 18.37 | 19.32 |
| *Qwen2.5-14B-Instruct* | 29.03 | 26.45 | 19.10 | 20.82 | 13.04 | 10.94 | 10.52 |
| *Qwen2.5-32B-Instruct* | 2.98 | 21.89 | 3.33 | 3.08 | 3.42 | 2.77 | 2.51 |

Table 4: **Switching probabilities (%) for bottom 7 epistemic markers across models.** Several models compress distinct low-certainty markers into similarly low probabilities, especially in larger models like *Qwen2.5-32B-Instruct*, suggesting a limited sensitivity to fine-grained uncertainty.

(2025a). This is likely because of using different models and prompt design. The key parameter values we obtained are shown in Table 2, and more detail can be found in Appendix E.

**Not all LLMs decision-making choices can be explained well by Prospect Theory.** Following human econometric model regression standards, we consider a MAE score $> 0.20$ means regression result not reliable, and a McFadden pseudo-$R^2$ score $< 0.10$ means using PT may be meaningless. (McFadden, 1977) *Llama-3.1-8B-Instruct* and *Qwen2.5-14B-Instruct* have too high MAE and too low $R^2$, which indicates they do not fit into Prospect Theory well.

**LLMs are consistent and show human-like risk preference, but not consistent and show significantly lower loss aversion than human.** All models show some extent of risk aversion $\sigma$ between 0.4 and 0.6 (a bit lower than human), but they are not consistent in $\lambda$ value and all significantly lower than human. Moreover, some models are more sensitive to loss ($\lambda > 1$), while others seems more sensitive to gain($\lambda < 1$). *Llama-3.1-8B-Instruct* and *Qwen2.5-7B-Instruct* have unusual boundary $\lambda$ values, which implies they have an irregular choice pattern.

**Majority of LLMs show human-like probability distortion.** We observe that the majority of human-like probability weighting in LLMs, which is overweighting small probabilities and underweighting large probabilities. This is the opposite result given by Jia et al. (2024).

### 6.2 Cross-Model Comparison of Marker Mappings

The experimental results are shown in Tables 3 and 4. We also intuitively present the probability mapping of different models for different markers in Figure 4.

**Different models assign different probabilities to the same epistemic markers.** Across the top 7 markers, the mapping probabilities vary significantly between models. For example, *"almost certain"* is mapped to over 97% in *Qwen2.5-32B-Instruct* but only less than 83% in *Qwen2.5-7B-Instruct*, while *"somewhat likely"* ranges from 8.9% (*Qwen2.5-7B-Instruct*) to 48.9% (*Mistral-7B-Instruct-v0.3*). This highlights that language models, even within the same family, do not share a consistent internal representation of uncertainty expressions.

**Despite the divergence in absolute values, there is a striking consistency in relative ordering among markers.** Nearly all models assign the highest probabilities to *"almost certain"*, followed by *"highly likely"* and *"very likely"*, suggesting that models broadly agree on the ordinal

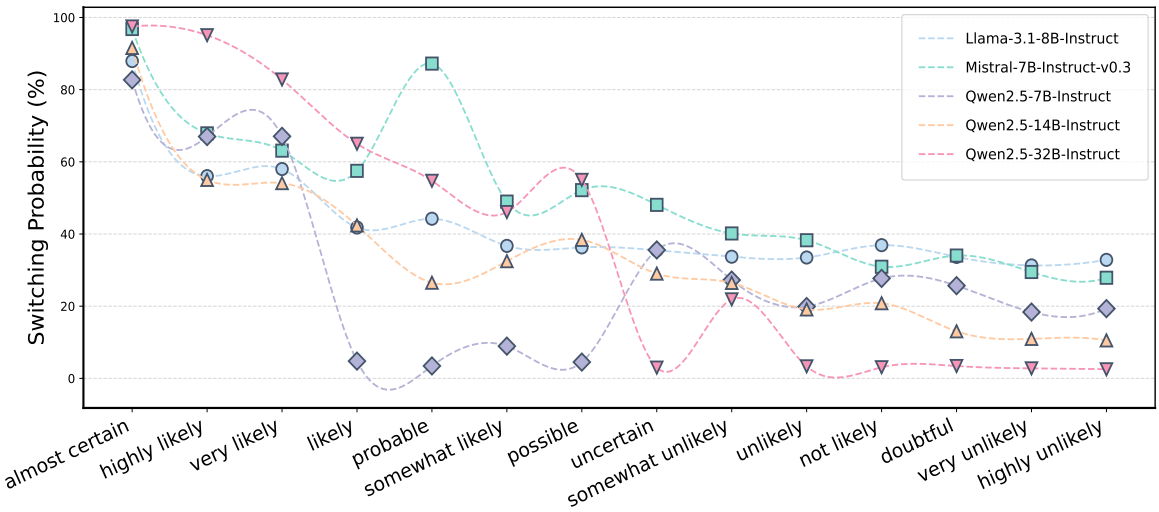


Figure 4: **Probability mapping of different models for different markers.** While different models assign divergent probability values to the same marker, the relative ordering of markers remains broadly consistent.

semantics of epistemic language even if they differ numerically. This partial consistency could be useful for comparative or ranking-based tasks, but it is insufficient for applications requiring calibrated probabilistic reasoning.

**Some models collapse multiple distinct markers into indistinguishably low probabilities.** Some models exhibit a compression effect, where multiple markers are mapped to similar probabilities. For instance, *Qwen2.5-32B-Instruct* maps 6 weakest markers to probabilities between 2.5% and 3.5%, effectively collapsing fine-grained uncertainty distinctions. Such behavior suggests an overconfident or categorical interpretation of ambiguous language, which undermines the expressiveness of epistemic markers. Similarly, *Llama-3.1-8B-Instruct* displays a narrow dynamic range for markers, most of them close to 30%, indicating limited sensitivity of epistemic markers.

**Larger model size does not guarantee human-aligned probability interpretations.** Interestingly, larger models do not always produce more calibrated mappings. Although *Qwen2.5-32B-Instruct* maps a wider range, with some close to 100% and some near 0%, this does not prevent it from having poor discrimination on low-probability markers. Meanwhile, smaller models such as *Mistral-7B-Instruct-v0.3* perform no worse than *Qwen2.5-32B-Instruct* across the entire mapping of markers. These patterns suggest that scale alone does not guarantee a nuanced probabilistic understanding of epistemic markers, and that other factors such as model architecture may play a more decisive role.

### 6.3 Changes of Decision-Making Behavior under Epistemic Markers

The estimation results under varying degrees of linguistic uncertainty are illustrated in Figure 5 and 6. We examine how replacing precise numerical probabilities with epistemic markers affects LLMs' internal decision-making parameters and performance under the framework of Prospect Theory. The analysis covers four experimental conditions, including marker substitution in either or both options of the lottery choice task.

**Linguistic uncertainty causes moderate shifts in risk preference but affects other decision parameters more profoundly.** For the majority of models, the estimated risk preference remains relatively stable despite replacing numeric probabilities with epistemic markers, indicating that risk attitudes are only mildly perturbed by uncertainty. However, models differ considerably in their loss aversion and probability weighting under these conditions. Larger models tend to maintain more consistent loss sensitivity resembling human behavior, whereas smaller models show larger fluctuations. Similarly, the increase in $\gamma$ after adding disturbance shows that most models exhibit increasingly conservative probability distortions, reflecting cautious decision patterns under ambiguity.

**Epistemic markers sometimes activate a more human-aligned decision-making behavior for**

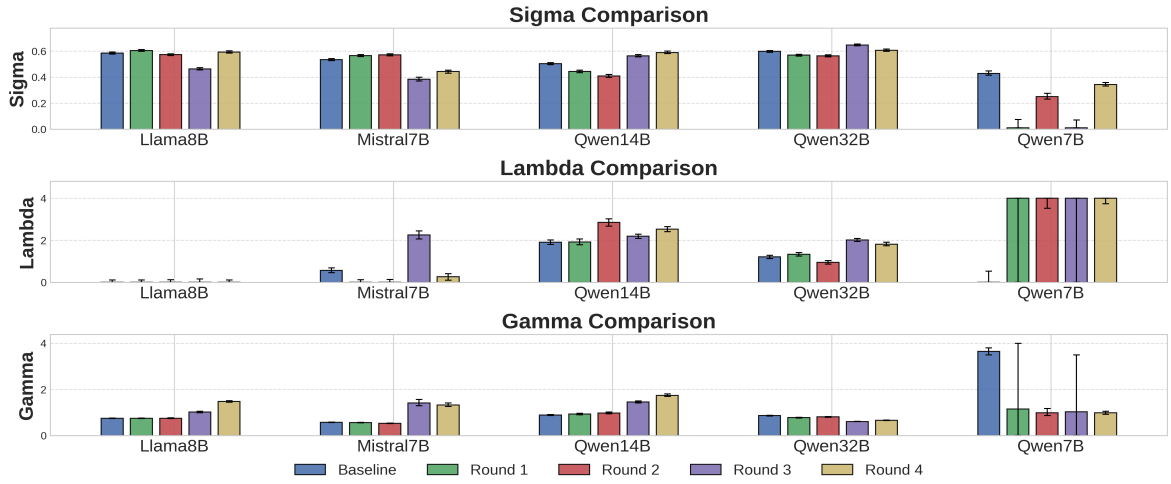


Figure 5: **The PT parameters estimated across rounds.** The "Baseline" represents results from the first stage of our experiment. In subsequent rounds, parameters fluctuate relative to the baseline, indicating instability in the models PT alignment under epistemic perturbations.

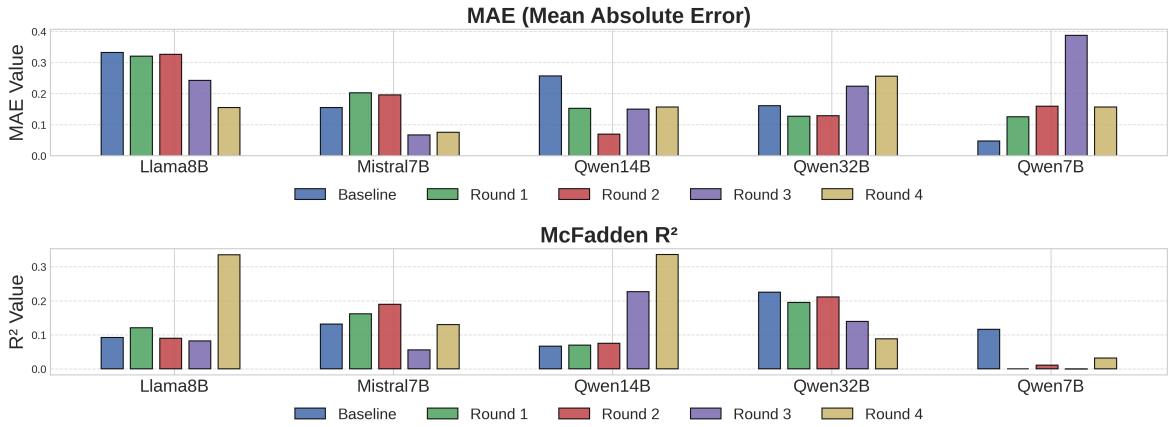


Figure 6: **The PT performance across rounds.** Similar to Table 5, the "Baseline" displays each model's original PT fit. Across later rounds, both MAE and McFadden $R^2$ values vary considerably, further suggesting that LLM behavior under uncertainty is not reliably captured by human cognitive models.

**LLMs.** This may because LLMs are more sensitive to linguistic uncertainty than numeric probability value, showing a different probability interpretation of epistemic markers as we obtain in stage 2. However, this finding does not mean epistemic markers can guarantee a more human-aligned decision-making behavior. It further strengthen our claim that LLMs' decision-making is not stable and explainable.

**Larger size LLMs seem to have more stable decision-making behavior under linguistic uncertainty.** While smaller language models (e.g. *Mistral-7B-Instruct-v0.3*) exhibit drastically varying PT parameters under epistemic markers, sometimes even touch boundary values (e.g. close to zero loss aversion), which indicates a fail regression. Larger models (e.g. *Qwen2.5-32B-Instruct*) are more robust to linguistic uncertainty.

## 7 Conclusion

This study challenges the direct application of human-centric Prospect Theory to LLMs, especially in linguistically uncertain contexts. Through a novel three-stage framework combining behavioral economics and semantic analysis, we show that LLMs (1) decisions may not always explained by Prospect Theory (2) exhibit widely divergent interpretations to epistemic markers under decision-making context (3) fail to maintain stable and reasonable PT parameters under epistemic markers despite scale advantages. We also present the first mapping between markers and probability values in the context of economic decisions.

## Limitations

We deliberately choose non-reasoning models and ask them to directly make their choice, because many models tend to directly calculate the expected utility and choose the option with higher EU if they are allowed to reason. This contradicts with our purpose, which is to obtain their risk attitude under quick decisions. However, avoiding reasoning models may limit the applicability of our results. Another limitation is that the probability mapping for epistemic markers are obtained in a slightly different context with the decision-making evaluation test. Meanwhile, we do not find a universal pattern for PT parameter changes under epistemic markers. This makes it hard to figure out a more powerful decision theory for LLM than PT. Future work may explore this kind of decision theories, and better calibration methods for epistemic markers.

## Ethics Statement

This paper utilize a lottery-based economic questionnaire developed by (Tanaka et al., 2010) published on America Economic Association, which allows usage with appropriate citations. The questionaire is done by LLMs, so there is no privacy issues. Our experiment discuss the risk attitude measured under Prospect Theory, which does not contain offensive expressions. The questionnaire is intended to test risk attitude measured by Prospect Theory, and it is used as intended in our paper.

Our experiment involves the usage of Qwen2.5 series models (7B, 14B, 32B) (Qwen et al., 2025) with Apache 2.0 license, Mistral-7B-Instruct-v0.3 with Apache 2.0 license, and Llama3.1-8B-Instruct with Llama3.1 license. They run on an 8x RTX 3090 GPU cluster.

Our paper mainly tests the robustness of PT under epistemic uncertainty, which points out risk of using PT in LLM-related fields. This does not introduce extra risks. Our research focuses on financial decision-making within the English language domain.

## References


Rozumko Agata. 2017. Adverbial markers of epistemic modality across disciplinary discourses: A contrastive study of research articles in six academic disciplines. *Studia Anglica Posnaniensia*, 52(1):73–101.

M. Allais. 1953. Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'ecole americaine. *Econometrica*, 21(4):503–546.

Catarina G Belem, Markelle Kelly, Mark Steyvers, Sameer Singh, and Padhraic Smyth. 2024. Perceptions of linguistic uncertainty by language models and humans. *Preprint*, arXiv:2407.15814.

Henrik Bergqvist. 2015. Epistemic marking and multiple perspective: An introduction. *STUF - Language Typology and Universals*, 68(2):123–141.

Eduard Brandstätter, Gerd Gigerenzer, and Ralph Hertwig. 2006. The priority heuristic: Making choices without trade-offs. *Psychological Review*, 113(2):409–432.

Sujoy Chakravarty and Jaideep Roy. 2009. Recursive expected utility and the separation of attitudes towards risk and ambiguity: an experimental study. *Theory and Decision*, 66(3):199–228.

Zehua Cheng, Manying Zhang, Jiahao Sun, and Wei Dai. 2025. On weaponization-resistant large language models with prospect theoretic alignment. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10309–10324, Abu Dhabi, UAE. Association for Computational Linguistics.

Bradley Efron. 1979. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26.

Uri Gneezy, John A. List, and George Wu. 2006. The uncertainty effect: When a risky prospect is valued less than its worst possible outcome. *The Quarterly Journal of Economics*, 121(4):1283–1309.

Jingru Jia, Zehua Yuan, Junhao Pan, Paul E. McNamara, and Deming Chen. 2024. Decision-making behavior evaluation framework for llms under uncertain context. *Preprint*, arXiv:2406.05972.

Daniel Kahneman and Amos Tversky. 1979. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291.

Katherine Keith and Amanda Stent. 2019. Modeling financial analysts' decision making via the pragmatics and semantics of earnings calls. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 493–503, Florence, Italy. Association for Computational Linguistics.

Dongryeol Lee, Yerin Hwang, Yongil Kim, Joonsuk Park, and Kyomin Jung. 2025. Are llm-judges robust to expressions of uncertainty? investigating the effect of epistemic markers on llm-based evaluation. *Preprint*, arXiv:2410.20774.

Eric Lehman, Vladislav Lialin, Katelyn Y. Legaspi, Anne Janelle R. Sy, Patricia Therese S. Pile, Nicole Rose I. Alberto, Richard Raymund R. Ragasa, Corinna Victoria M. Puyat, Isabelle Rose I. Alberto, Pia Gabrielle I. Alfonso, Marianne Taliño, Dana Moukheiber, Byron C. Wallace, Anna Rumshisky, Jenifer J. Liang, Preethi Raghavan, Leo Anthony Celi, and Peter Szolovits. 2022. Learning to ask like a physician. *Preprint*, arXiv:2206.02696.

Milad Leyli-abadi, Ricardo J. Bessa, Jan Viebahn, Daniel Boos, Clark Borst, Alberto Castagna, Ricardo Chavarriaga, Mohamed Hassouna, Bruno Lemetayer, Giulia Leto, Antoine Marot, Maroua Meddeb, Manuel Meyer, Viola Schiaffonati, Manuel Schneider, and Toni Waefler. 2025. A conceptual framework for ai-based decision systems in critical infrastructures. *Preprint*, arXiv:2504.16133.

Jiaxin Liu, Yi Yang, and Kar Yan Tam. 2025a. Evaluating and aligning human economic risk preferences in llms. *Preprint*, arXiv:2503.06646.

Jiayu Liu, Qing Zong, Weiqi Wang, and Yangqiu Song. 2025b. Revisiting epistemic markers in confidence estimation: Can markers accurately reflect large language models' uncertainty? *Preprint*, arXiv:2505.24778.

Neda Mahmoodi, Georgina L. Jones, Tom Muskett, and Sally Sargeant. 2020. Exploring shared decision making in breast cancer care. *Communication & Medicine*, 16(1):40–53.

Daniel McFadden. 1977. Quantitative methods for analyzing travel behaviour of individuals: Some recent developments. *Cowles Foundation Discussion Papers*, (474).

Brooks P. and Zank H. 2005. Loss averse behavior. *J Risk Uncertainty*, 31:301325.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Neil Rathi, Dan Jurafsky, and Kaitlyn Zhou. 2025. Humans overrely on overconfident language models, across languages. *Preprint*, arXiv:2507.06306.

Tomomi Tanaka, Colin F. Camerer, and Quang Nguyen. 2010. Risk and time preferences: Linking experimental and household survey data from vietnam. *American Economic Review*, 100(1):55771.

John von Neumann, Oskar Morgenstern, and Ariel Rubinstein. 1944. *Theory of Games and Economic Behavior*, 60th anniversary commemorative edition edition. Princeton University Press.

Yikai Wang, Xiaocheng Li, and Guanting Chen. 2025. Risk profiling and modulation for llms. *Preprint*, arXiv:2509.23058.

Tianshi Zheng, Yixiang Chen, Chengxi Li, Chunyang Li, Qing Zong, Haochen Shi, Baixuan Xu, Yangqiu Song, Ginny Y. Wong, and Simon See. 2025a. The curse of cot: On the limitations of chain-of-thought in in-context learning. *CoRR*, abs/2504.05081.

Tianshi Zheng, Jiayang Cheng, Chunyang Li, Haochen Shi, Zihao Wang, Jiaxin Bai, Yangqiu Song, Ginny Y. Wong, and Simon See. 2025b. Logidynamics: Unraveling the dynamics of logical inference in large language model reasoning. *CoRR*, abs/2502.11176.

Tianshi Zheng, Zheye Deng, Hong Ting Tsang, Weiqi Wang, Jiaxin Bai, Zihao Wang, and Yangqiu Song. 2025c. From automation to autonomy: A survey on large language models in scientific discovery. *CoRR*, abs/2505.13259.

Tianshi Zheng, Weihan Li, Jiaxin Bai, Weiqi Wang, and Yangqiu Song. 2025d. KnowShiftQA: How robust are RAG systems when textbook knowledge shifts in k-12 education? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 183–195, Vienna, Austria. Association for Computational Linguistics.

Kaitlyn Zhou, Jena D. Hwang, Xiang Ren, and Maarten Sap. 2024. Relying on the unreliable: The impact of language models' reluctance to express uncertainty. *Preprint*, arXiv:2401.06730.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. 2023. Navigating the grey area: How expressions of uncertainty and overconfidence affect language models. *Preprint*, arXiv:2302.13439.

Qing Zong, Zhaowei Wang, Tianshi Zheng, Xiyu Ren, and Yangqiu Song. 2025. Comparisonqa: Evaluating factuality robustness of llms through knowledge frequency control and uncertainty. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 4101–4117. Association for Computational Linguistics.

# A Hyperparameters

## A.1 Probability Mapping Experiment

In the probability mapping experiments in 5.1, the monetary reward is fixed at $M = 100$. This value is chosen to provide a clear and intuitive payoff magnitude without introducing excessive numerical complexity.

The probability parameter $p$ takes values from the set $probs = \{5, 15, 25, 35, 45, 55, 65, 75, 85, 95\}$. These values are selected to uniformly cover the range of possible probabilities from low to high in increments of 10 percentage points, enabling systematic analysis of internal probability values of epistemic markers.

### A.2 Model Generation

Temparature is set to 0.7 for all experiments. Other hyperparameter are default values in huggingface.

| Hyperparameter | Value |
|---|---|
| Generation method | Sampling |
| Temperature | 0.7 |
| Maximum new tokens | 8 |
| Padding token | EOS token |
| EOS token | EOS token |
| Batch size | 16 |
| History length | 10 |
| Number of lottery rounds | 35 |

Table 5: Key hyperparameters for model generation

## B Lottery Design

Table 7, 8 and 9 shows the lottery design for PT parameter estimation. The values here are from Tanaka et al. (2010). They are specially designed to get best PT parameters.

## C Prompt Design

Figure 7 shows prompt design for decision-making evaluation test and probability mapping test. All lotteries are sampled 256 times. To simulate human decision-making, while keeping the model directly output its answer, up to 15 history decisions are maintained. Meanwhile, we use a random order for the 35 lotteries to relieve positional bias. An introduction is provided at the very beginning. NUMBER will be replaced by the values stated in table 7, 8 and 9 .

## D Marker Replacement Rules

For the details of how we replace probabilities with markers, see Table 6.

## E Detailed Experimental Results

In the main text, we presented selected key experimental results and visualizations. To provide a more comprehensive view of model performance across different rounds, we include in this appendix the full set of parameter estimates and model fit metrics.

Specifically, Tables 10, 11, and 12 report the estimates of parameters $\sigma$, $\lambda$, and $\gamma$ with their 95% confidence intervals for each model and round. Table 13 summarizes the models mean absolute errors (MAE) and McFadden $R^2$ values across rounds.

These additional data offer deeper insights into model behavior and the dynamics observed throughout the experiments.

## F Discussion and Implications

Our findings reveal challenging difficulties in applying human-centric cognitive frameworks, especially Prospect Theory (PT), to LLM decision-making. Different models display distinct interpretations of epistemic uncertainty markers, leading to divergent decision behaviors. Introducing these markers into the decision-making framework substantially alters LLM choices.

Our results suggest that LLMs may not inherently understand risk in human-like ways; their responses often reflect statistical training artifacts rather than cognitively grounded reasoning. **We recommend conducting regression analyses and goodness-of-fit tests before applying human cognitive models to LLMs.**

In real-world applications (e.g., medical diagnosis or financial advice), LLMs may give inconsistent recommendations when probabilistic language varies, posing reliability concerns. **We recommend establishing consistent standards for expressing uncertainty in LLM-driven decision systems.**

Furthermore, larger LLMs tend to exhibit more PT-like decision behavior, with PT parameters more closely aligned to human estimates. **We recommend using LLMs with at least 14B parameters when integrating PT into decision-making systems.**

## G Failure Cases Analysis

In our initial implementation of the marker mapping experiment, we adopted the set of epistemic markers from Table 6 (Human Judgements of Templates Based on Reliability) in Zhou et al. (2024). These markers were originally designed to test both human and LMs judgments of the reliability conveyed by these expressions. However, when directly applied to our economic decision-making setting, the resulting mappings for LLMs were unexpectedly unstable and, in some cases, counterintuitive.

We summarize two major issues observed in the experimental outcomes:

| Model | 30% | 70% | 10% | 90% |
|---|---|---|---|---|
| Qwen2.5-7B-Instruct | uncertain | almost certain | somewhat likely | highly likely |
| Llama3.1-8B-Instruct | likely | almost certain | very unlikely | almost certain |
| Mistral-7B-Instruct-v0.3 | very unlikely | highly likely | highly unlikely | almost certain |
| Qwen2.5-14B-Instruct | somewhat unlikely | highly likely | very unlikely | almost certain |
| Qwen2.5-32B-Instruct | somewhat unlikely | probable | somewhat likely | almost certain |

Table 6: Marker Replacement Rules for Different Models. This is determined for introducing the least numeric differences, and balancing model and human interpretations.

**Prompt Templates**

**Beginning Instruction**
You are invited to participate in an experiment.
Your task is to choose between option K and
option U for each of the following lotteries.
Here is lottery {i}:

**a) Used in Probability Mapping**
For option K:
p% probability to win $100.
For option U:
MARKER to win $100.

**b) Used in PT Estimation**
For option U:
NUMBER% probability to win $NUMBER
NUMBER% probability to win/lose $NUMBER
For option K:
NUMBER% probability to win $NUMBER
NUMBER% probability to win/lose $NUMBER

**End Instruction**
Please DO NOT REASON and
DIRECTLY output your choice,
by ONLY returning
one of the following two labels:
"[K]", "[U]".
The answer is:

Figure 7: **Templates for prompts used in the probability mapping and Prospect Theory estimation tasks.** The design includes an initial instruction, task-specific lottery descriptions, and a fixed closing instruction to ensure direct model responses without reasoning.

**(1) Highly oscillatory choice patterns.** Ideally, the number of times the model selects option K should increase monotonically with $p$, yielding a single and well-defined switching point. In practice, the selection curves were often non-monotonic, with multiple apparent switching points, which made the mapping probability $p_{\text{mapping}}$ ill-defined. An example is shown in Figure 8.

**(2) Severe semantic mismatches for high-**

| | Option K | | Option U | |
|---|---|---|---|---|
| **Lottery** | 30% | 70% | 10% | 90% |
| 1 | 40 | 10 | 68 | 5 |
| 2 | 40 | 10 | 75 | 5 |
| 3 | 40 | 10 | 83 | 5 |
| 4 | 40 | 10 | 93 | 5 |
| 5 | 40 | 10 | 106 | 5 |
| 6 | 40 | 10 | 125 | 5 |
| 7 | 40 | 10 | 150 | 5 |
| 8 | 40 | 10 | 185 | 5 |
| 9 | 40 | 10 | 220 | 5 |
| 10 | 40 | 10 | 300 | 5 |
| 11 | 40 | 10 | 400 | 5 |
| 12 | 40 | 10 | 600 | 5 |
| 13 | 40 | 10 | 1000 | 5 |
| 14 | 40 | 10 | 1700 | 5 |

Table 7: Series 1 both options are gains.

| | Option K | | Option U | |
|---|---|---|---|---|
| **Lottery** | 90% | 10% | 70% | 30% |
| 1 | 40 | 30 | 54 | 5 |
| 2 | 40 | 30 | 56 | 5 |
| 3 | 40 | 30 | 58 | 5 |
| 4 | 40 | 30 | 60 | 5 |
| 5 | 40 | 30 | 62 | 5 |
| 6 | 40 | 30 | 65 | 5 |
| 7 | 40 | 30 | 68 | 5 |
| 8 | 40 | 30 | 72 | 5 |
| 9 | 40 | 30 | 77 | 5 |
| 10 | 40 | 30 | 83 | 5 |
| 11 | 40 | 30 | 90 | 5 |
| 12 | 40 | 30 | 100 | 5 |
| 13 | 40 | 30 | 110 | 5 |
| 14 | 40 | 30 | 130 | 5 |

Table 8: Series 2 both options are gains.

| | Option K | | Option U | |
|---|---|---|---|---|
| | **50%** | **50%** | **50%** | **50%** |
| **Lottery** | **Win** | **Lose** | **Win** | **Lose** |
| 1 | 25 | 4 | 30 | 21 |
| 2 | 4 | 4 | 30 | 21 |
| 3 | 1 | 4 | 30 | 21 |
| 4 | 1 | 4 | 30 | 16 |
| 5 | 1 | 8 | 30 | 16 |
| 6 | 1 | 8 | 30 | 14 |
| 7 | 1 | 8 | 30 | 11 |

Table 9: Series 3 both options have gains and losses.

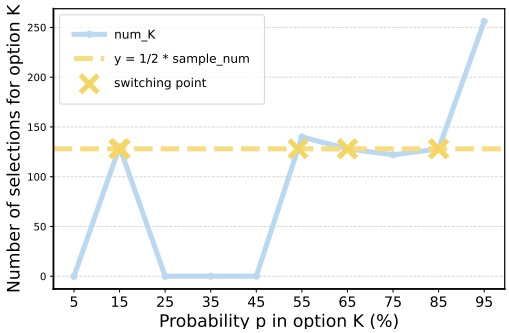


Figure 8: **An example of non-monotonic result with multiple switching points.** This result comes from marker *"Its undoubtedly to"* with *Qwen2.5-7B-Instruct* model over 256 trials. The blue line shows counts of option K selections, and the yellow dashed line marks half the trials. The crosses denote switching points.

**certainty markers.** Some markers, such as *"It's extremely certain to"*, convey very high certainty in human interpretation, but were mapped by the model to surprisingly low numerical probabilities. An example is shown in Figure 9.

We hypothesize that these issues may stem from several factors:

**(1) Marker length and syntactic complexity.** The selected markers were often not single words or short phrases, but full clausal structures. This may introduce additional semantic and syntactic cues unrelated to uncertainty, thereby interfering with probability interpretation.

**(2) Shift from first-person to third-person framing.** The original markers in Zhou et al. (2024) were presented in the first person (e.g., *"I am not confident, maybe it's..."*), whereas our experiment reformulated them into third-person expressions (e.g., *"It's not condent, maybe can..."*).

**(3) Intrinsic instability of epistemic markers.** Even for human interpretation, such markers are context-dependent and inherently imprecise (Liu et al., 2025b; **?**; **?**; Zheng et al., 2025a,b). Their probability mapping by LLMs in economic decision-making contexts may exhibit fundamental reliability flaws (Zheng et al., 2025c).

These limitations motivated the redesign of our marker set and prompt formulation in subsequent experiments.

| Model | $\sigma$ (95% CI) | | | | |
|---|---|---|---|---|---|
| | baseline | round1 | round2 | round3 | round4 |
| *Llama-3.1-8B-Instruct* | 0.585 (0.578, 0.592) | 0.605 (0.599, 0.612) | 0.573 (0.566, 0.580) | 0.463 (0.455, 0.473) | 0.593 (0.585, 0.602) |
| *Mistral-7B-Instruct-v0.3* | 0.534 (0.526, 0.543) | 0.566 (0.558, 0.574) | 0.571 (0.564, 0.580) | 0.384 (0.370, 0.399) | 0.444 (0.431, 0.454) |
| *Qwen2.5-7B-Instruct* | 0.429 (0.415, 0.445) | 0.010 (0.010, 0.074) | 0.250 (0.232, 0.275) | 0.010 (0.010, 0.071) | 0.344 (0.332, 0.359) |
| *Qwen2.5-14B-Instruct* | 0.503 (0.495, 0.511) | 0.444 (0.434, 0.454) | 0.409 (0.398, 0.421) | 0.563 (0.555, 0.573) | 0.589 (0.581, 0.600) |
| *Qwen2.5-32B-Instruct* | 0.598 (0.591, 0.605) | 0.569 (0.561, 0.576) | 0.564 (0.556, 0.572) | 0.647 (0.640, 0.655) | 0.607 (0.599, 0.615) |

Table 10: $\sigma$ estimates with 95% confidence intervals across different rounds for each model.

| Model | $\lambda$ (95% CI) | | | | |
|---|---|---|---|---|---|
| | baseline | round1 | round2 | round3 | round4 |
| *Llama-3.1-8B-Instruct* | 0.010 (0.010, 0.125) | 0.010 (0.010, 0.117) | 0.010 (0.010, 0.130) | 0.010 (0.010, 0.168) | 0.010 (0.010, 0.112) |
| *Mistral-7B-Instruct-v0.3* | 0.570 (0.453, 0.688) | 0.010 (0.010, 0.132) | 0.010 (0.010, 0.135) | 2.260 (2.060, 2.445) | 0.267 (0.105, 0.414) |
| *Qwen2.5-7B-Instruct* | 0.010 (0.010, 0.584) | 4.000 (0.010, 4.000) | 4.000 (3.526, 4.000) | 4.000 (0.010, 4.000) | 4.000 (3.736, 4.000) |
| *Qwen2.5-14B-Instruct* | 1.909 (1.801, 2.013) | 1.919 (1.784, 2.070) | 2.851 (2.675, 3.023) | 2.191 (2.094, 2.295) | 2.531 (2.409, 2.648) |
| *Qwen2.5-32B-Instruct* | 1.213 (1.133, 1.295) | 1.340 (1.250, 1.423) | 0.953 (0.866, 1.036) | 2.013 (1.945, 2.090) | 1.815 (1.736, 1.905) |

Table 11: $\lambda$ estimates with 95% confidence intervals across different rounds for each model.

| Model | $\gamma$ (95% CI) | | | | |
|---|---|---|---|---|---|
| | baseline | round1 | round2 | round3 | round4 |
| *Llama-3.1-8B-Instruct* | 0.753 (0.740, 0.767) | 0.750 (0.737, 0.762) | 0.755 (0.741, 0.768) | 1.020 (0.994, 1.053) | 1.478 (1.443, 1.517) |
| *Mistral-7B-Instruct-v0.3* | 0.577 (0.565, 0.590) | 0.564 (0.554, 0.574) | 0.535 (0.527, 0.545) | 1.413 (1.300, 1.564) | 1.332 (1.266, 1.414) |
| *Qwen2.5-7B-Instruct* | 3.645 (3.501, 3.783) | 1.156 (0.010, 4.000) | 0.985 (0.870, 1.174) | 1.035 (0.010, 3.498) | 0.987 (0.926, 1.053) |
| *Qwen2.5-14B-Instruct* | 0.896 (0.875, 0.919) | 0.933 (0.905, 0.967) | 0.983 (0.944, 1.023) | 1.455 (1.414, 1.503) | 1.742 (1.693, 1.803) |
| *Qwen2.5-32B-Instruct* | 0.867 (0.851, 0.884) | 0.778 (0.762, 0.793) | 0.812 (0.797, 0.828) | 0.614 (0.605, 0.623) | 0.664 (0.653, 0.674) |

Table 12: $\gamma$ estimates with 95% confidence intervals across different rounds for each model.

| Model | MAE | | | | | McFadden $R^2$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | baseline | round1 | round2 | round3 | round4 | baseline | round1 | round2 | round3 | round4 |
| *Llama-3.1-8B-Instruct* | 0.332 | 0.320 | 0.326 | 0.242 | 0.155 | 0.092 | 0.121 | 0.090 | 0.082 | 0.335 |
| *Mistral-7B-Instruct-v0.3* | 0.155 | 0.202 | 0.196 | 0.067 | 0.075 | 0.132 | 0.162 | 0.190 | 0.056 | 0.130 |
| *Qwen2.5-7B-Instruct* | 0.047 | 0.125 | 0.159 | 0.387 | 0.157 | 0.116 | 0.000 | 0.011 | -0.001 | 0.032 |
| *Qwen2.5-14B-Instruct* | 0.257 | 0.152 | 0.069 | 0.150 | 0.157 | 0.067 | 0.070 | 0.075 | 0.227 | 0.336 |
| *Qwen2.5-32B-Instruct* | 0.161 | 0.127 | 0.129 | 0.224 | 0.256 | 0.225 | 0.195 | 0.211 | 0.139 | 0.088 |

Table 13: Mean absolute error (MAE) and McFadden $R^2$ across different rounds for each model.

**Failure Case of Markers Mapping**

**user:**

Here is lottery 2:

For option K: 15% probability to win $100.
For option U: It's extremely certain to win $100.

Please DO NOT REASON and DIRECTLY output your choice,
by ONLY returning one of the following two labels: "[K]", "[U]".

The answer is:

---

**assistant:**

K

Figure 9: **An example of mapping high-certainty markers to low probablity.** In this case, *Qwen2.5-7B-Instruct* preferred option K when $p = 15\%$ and option U was described using *"Its extremely certain to"*, implying a far lower internal probability than expected.