Decoding Latent Decision Strategies from Think-Aloud Protocols

Jiaxuan Li[1*], Yufan Tian[2*], Kangjuan Lv (lvkangjuan@shu.edu.cn)[1#] & Lisheng He (hlisheng@shu.edu.cn)[1#]
[1]SILC Business School, Shanghai University, Shanghai, China
[2]Institute for Sustainable Futures, University of Technology Sydney, Sydney, Australia

Abstract

Latent decision strategies are central to human behavior and cognition. They may differ across persons and fluctuate over time within a single person. This paper introduces a framework that extracts latent, trial-level decision strategies from think-aloud verbalizations, with the resulting strategy estimates mapped onto formal strategy identification via quantitative model selection. We validated this approach with two intertemporal choice experiments, in which participants verbalized their thoughts while making binary choices between smaller-sooner and larger-later options in free-choice (Experiment 1) or instructed (Experiment 2) conditions. We used three popular LLMs to rate participants' alternative-based versus attribute-based evaluations based on their think-aloud protocols at the trial level. Results suggest that think-aloud protocols effectively capture individual variations in decision strategies, as reflected in comparisons between computational models, and can detect strategy shifts due to instructions. Our framework offers a scalable, powerful tool for understanding latent cognitive processes underlying human choice behavior.

Keywords: LLMs; think-aloud; decision strategy; behavioral modeling; intertemporal choice

Introduction

Intertemporal decisions involve the trade-off between smaller-sooner (SS) and larger-later (LL) options and play a central role in everyday decision-making (Frederick, et al., 2002). These choices range from everyday behaviors, such as meal portion sizes, to major life events, including educational attainment, family planning, health management, and financial savings (Berns et al., 2007). Understanding how people make such decisions is crucial for promoting rational intertemporal tradeoffs through behavioral and cognitive interventions (Reeck et al., 2017).

Decades of behavioral decision research have taken a "revealed preference" approach to make inferences about decision strategies from the observed choice data. This line of work suggests that intertemporal decisions are mostly made in an attribute-based fashion, meaning that people tend to compare along the payoff and delay attributes respectively (Dai & Busemeyer, 2014; Ericson et al., 2015; Scholten et al., 2014; Scholten & Read, 2010; Suter et al., 2016). However, this tendency is not universal. While most people's data are better described by attribute-based models, a significant portion are better described by alternative-based models (He et al., 2023). Moreover, people may switch between different choice strategies across trials. Broadly speaking, decision makers may adopt decision strategies adaptively, contingent on the choice problem under consideration (Payne et al., 1988, 1993). Yet the "revealed preference" approach is inherently limited in unraveling the fluctuation of decision strategies across trials, because it requires multiple choice problems for model selection to infer the underlying decision strategies.

Think-aloud protocols present an excellent opportunity to identify decision strategies at the trial level (Ericsson & Simon, 1993; van Solomon et al., 1994; Weber et al., 2007; Wurgaft et al., 2025). A think-aloud protocol asks participants to verbalize their thoughts, feelings, and actions in real time while performing a choice task. It gives researchers a window into participants' cognitive processes beyond just the final choice outcome. Recent work suggests that human decision makers are aware of their mental processes underlying their choices (Morris et al., 2025) which supports the feasibility of deriving their decision strategies from concurrent think-aloud data.

The think-aloud method often requires substantial training and time investment, undermining its unique potential in understanding human decision making and cognition in general (Larkin et al., 1980; Newell & Simon, 1972). Fortunately, the advent of natural language processing (NLP) technologies (such as speech-to-text models) and large language models (LLMs) has opened up new avenues to addressing the limitations of traditional manual coding of think-aloud protocols (Brown et al., 2020; Radford et al., 2023). The previously tedious and laborious work can now be implemented using a two-step pipeline. First, NLP technologies allow us to transcribe verbal reports from think-aloud protocols into texts easily. Second, commercial LLMs can be used to understand the transcribed texts, decode the decision strategies conveyed in the verbal reports and have created unprecedented opportunities for analyzing language data on a massive scale (Bhatia et al., 2026; Demszky et al., 2023; Fulawka et al., 2025; Wurgaft et al., 2025; Xie et al., 2025).

We tested the viability of this approach in intertemporal choice, asking LLMs to decode participants' latent decision strategies along the spectrum between the alternative-based and attribute-based evaluation at the trial level. Recent work has begun integrating LLMs with think-aloud protocol analysis in human choice behavior (Fulawka et al., 2025; Xie et al., 2025). Our work differs from these prior approaches in two key respects. First,

[*] Equal contribution.
[#] Corresponding authors.

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

whereas both Fulawka et al. (2025) and Xie et al. (2025) focus on risky decision making, our work focuses on intertemporal choice. Second, and more importantly, while previous work directly benchmarks LLM outputs to choice predictions, we compared the LLM outputs to latent decision strategies derived from behavioral data. Our approach offers a more critical test of the alignment between the decision strategies derived from think-aloud protocols and those derived from behavioral choice modeling. Therefore, this work also represents a process-level validation of the traditional behavioral modeling approach to strategy identification.

Two experiments were run to validate our LLM-derived decision strategies from think-aloud protocols. In Experiment 1, participants made binary choices in standard SS-LL intertemporal decision tasks under the free choice condition. They were allowed to make decisions in whatever way they wished. In Experiment 2, participants were explicitly instructed to follow either an attribute-based rule (i.e., the attribute-based instruction condition) or an alternative-based rule (i.e., the alternative-based instruction condition). In both experiments, participants were asked to verbally report their thoughts when making choices. We further validated the LLM-derived decision strategies from think-aloud protocols with the decision strategies inferred from choice data, the latter of which was accomplished with Bayesian model selection on behavioral choice data (He et al., 2023). In so doing, we introduced a scalable framework to identify decision strategies from the process-level data.

Experiment 1

Methods

Participants We recruited 300 participants (mean age = 29.69 years; 54% female) from Credemo (https://www.credamo.com), an online crowdsource platform in China. Two-hundred and twelve participants submitted usable think-aloud recordings (pass rate = 70.7%; see below for detailed exclusion criteria). This pass rate was similar to that in Wurgaft et al.(2025), who collected think-aloud data on Prolific (www.prolific.com).

Stimuli To obtain a dataset that contains rich behavioral regularities, we designed the set of stimuli to elicit two of the most well-known phenomena in intertemporal choice: the common difference effect and the magnitude effect. The magnitude effect refers to the phenomenon that people tend to be more patient with large payouts than small ones (Baker et al., 2003; Benhabib et al., 2010; Benzion et al., 1989; Green et al., 1994; Holcomb & Nelson, 1992; Petry, 2001; Thaler, 1981). The common difference effect refers to the pattern that people tend to be less patient when both options are deferred further into the future (Loewenstein & Prelec, 1992).

The choice items were generated in the following way. The larger-later option, $LL = [x, d]$, offered a larger amount of x CNY, where $x \in \{52, 208, 1664\}$, after a delay of d weeks. The variation in x was designed to elicit the magnitude effect . The smaller-sooner option, $SS = [y, f]$, offered y CNY after a shorter delay of f weeks (also called front-end delay), where f took values of 0, 4, or 8 weeks. The variation in f was designed to elicit the common difference effect. The smaller amount y took three different values for each of the three x values in the following formulas: $y = x \times \rho$, where $\rho \in \{0.3, 0.6, 0.9\}$.

To prevent participants from easily detecting the relationship between x and y, we introduced variability to the y values by incorporating a random integer drawn from a discrete uniform distribution within the range of -4 to 4. Finally, we relied on the classic exponential model to set reasonable d values, and used the exponential discounting model to equalize the subjective utility between the two options, by setting $y \times \alpha^f = x \times \alpha^d$,where α took values of 0.9 or 0.95 (Samuelson, 1937). The full set of intertemporal choice items was identified by crossing three x values, three f values, three ρ values and two α values, resulting in 54 different items.

Procedures Upon receiving the two options in each problem, participants were asked to verbally express their thoughts concerning this choice. They were instructed to report any form of thought related to the choice problem. After reporting their thoughts, participants submitted the audio recording and indicated their choices between the two options. To ensure recording quality, participants were reminded to finish the experiment in a quiet environment to minimize external interference, and each recording was required to last at least 5 seconds in length . Two practice trials were also provided to familiarize participants with the procedures.

Audio Transcripts and Text Screening All participants' think-aloud audio recordings were transcribed into texts through the cutting-edge Chinese speech recognition API provided by iFlytek (Zhao, 2023). To further ensure transcription quality, we designed a three-stage text screening process. First, we conducted a quick screening, removing empty texts, texts containing only punctuation, vocal fillers or interjections (such as "um um um," "ah ah ah," etc.) and pure option repetitions (e.g., "I choose A"). We also identified mixed Chinese-English expressions that may arise from transcription errors and manually corrected the errors.

Second, we used DeepSeek for automated text screening. Specifically, to be included as a valid transcript, it has to (1) relate to the intertemporal choice task, (2) express choice reasons or subjective feelings, (3) use understandable Chinese expressions, and (4) be fluent with no obvious typos. If a text transcription failed to meet any of the four criteria, it would be considered invalid and removed from formal analysis.

Third, two of the authors manually reviewed each transcribed text considered valid and corrected typos if applicable. At the same time, they also checked the think-aloud items that were considered invalid to identify and bring back the items that were falsely excluded in the second stage.

After this three-stage processing, we obtained 9,981 valid think-aloud transcripts from 227 participants.

Subsequently, participants who had fewer than five trials with valid transcripts were excluded from formal data analysis. As a result, 212 participants with both choice data and at least five valid think-aloud trials were included in the formal data analysis.

> Intertemporal decision-making involves two types of decision strategies: alternative-based evaluation strategies and attribute-based comparison strategies.
> The **alternative-based evaluation** strategies refer to the decision-maker integrating and evaluating the magnitude of the amount and the length of the delay within each option separately, assessing the subjective value of each option on its own, without making cross-option comparisons within the same attribute.
> The **attribute-based comparison** strategies refer to the decision-maker comparing the difference in monetary amounts between the two options, or comparing the difference in delay lengths between the two options, and making a decision based on these within-attribute comparisons.
> In this decision task, one option is "receiving *x* Yuan in *d*weeks}," and the other option is "receiving *y* Yuan in *f*weeks}."
> **The decision maker's think-aloud verbalization is *transcription*.** Based on this, please assess the decision maker's strategy tendency and express it as a numerical value between 0 and 100 (0 represents a fully alternative-based evaluation strategy, 100 represents a fully attribute-based comparison strategy, and 50 represents neutral).
> **Please report only the number, with no other text.**

Figure 1: Prompt for LLM-based strategy decoding.

LLM Coding of Think-Aloud Transcripts We then employed LLMs to decode participants' decision strategies from their think-aloud data. Specifically, we designed a spectrum between alternative-based and attribute-based evaluation rules. The former refers to decision-makers integrating the amount and delay of each option internally, evaluating each option's subjective value before making choices; the latter emphasizes direct comparison of certain attributes (such as amount or delay) across options. LLMs were instructed to rate participant's tendency to make alternative-based or attribute-based choices from the think-aloud data in each trial on a 101-point scale, with 0 indicating purely alternative-based choices, 100 indicating purely attribute-based choices and 50 indicating neutral (i.e., a mixture of attribute-based and alternative-based evaluation with no obvious bias toward either side). The full prompt structure provided to each LLM is illustrated in Figure 1.

The same coding task was independently completed by three popular Chinese LLMs: DeepSeek (V3-0324, Wu et al., 2025), Qwen (Qwen-plus, Feng et al., 2024) and ERNIE (ernie-4.5-turbo, Sun et al., 2021). We retained all core generative parameters as their official defaults. To ensure robustness, we repeated each rating five times and took the average across repetitions as the final rating.

Behavioral Modeling Benchmark Following previous work (He et al., 2023), we conducted an extensive evaluation of the behavioral choice data by comparing 13 alternative-based discounting models against 5 attribute-based models of intertemporal decision making, to identify individual-level decision strategies from choice data. The detailed model functional forms and parameters can be found in He et al. (2023).

Bayesian model selection was used to evaluate participants' decision strategies from their behavioral choice data. Our main interest is in the relative performance between discounting models vs. time-as-attribute models, which tells us whether alternative-based or attribute-based decision strategies were adopted by the participants. Thus, for every participant's choice data, we calculated a log Bayes factor (logBF) between pairs of models (Kass & Raftery, 1995). This metric measures the relative performance of models i and j as: $\log \mathrm{BF}_{ij} = \log \frac{p(D_n|i)}{p(D_n|j)} = \log p(D_n|i) - \log p(D_n|j) = \log ML_i - \log ML_j$, where $\log ML_i$ stands for the log marginal likelihood for a discounting model i in accounting for choice data D_n and MLL_j stands for the log marginal likelihood for a time-as-attribute model j in accounting for the same choice data.

To obtain $\log ML_i$, we employed a simple Monte Carlo method to estimate the log marginal likelihood (logML) for each model for each participant's choice data (see He et al., 2019; 2023; Scholten et al., 2016 for similar applications): $\log ML_i = \log p(D_n|i) = \log \int p(D_n|i,\Theta)p(\Theta|i)d\Theta$, where D_n is participant n's choice data and Θ is the set of free parameters for model i. For each estimation, we first drew one million samples of parameter values from the prior distribution. With each sample of parameter values, we calculated the likelihood of the choice data, obtaining one million likelihood values. The marginal likelihood was estimated as the average of the one million likelihood values.

Since our behavioral modeling analysis involved 13 alternative-based discounting models and 5 time-as-attribute models, there were 13 × 5 = 65 pairs of an alternative-based model and an attribute-based model. To obtain a representative log bayes factor between time-as-attribute models and discounting models, we defined a new metric from Bayesian model comparison: $\log \mathrm{BF}_{AD} = \log ML_A - \log ML_D$, where $\log ML_D$ is the average log marginal likelihoods for the set of discounting models, and MLL_A is the average log marginal likelihoods for the set of time-as-attribute models . Note that the 65 pairs of models yielded highly consistent model comparison results with high correlations between different pairs (with median Pearson's R = .724).

Results

LLM-coded Decision Strategies From Think-Aloud Protocols LLMs' ratings from the think-aloud data suggest that participants displayed stronger tendencies of

attribute-based evaluation than alternative-based evaluation, with mean ratings of all LLMs above 50 (all *p*s <.001). On the trial level, DeepSeek coded 66.7% of the trials as adopting an attribute-based choice (i.e., with ratings above 50). Similarly, the remaining two language models also identified the majority of the trials as attribute-based, with Qwen identifying 73.1% of the trials to be attribute-based, ERNIE identifying 55.1% of the trials as attribute-based. The average of the three models identified 65.0% of the trials as attribute-based. This is consistent with prior work that suggests that most participants were better described by time-as-attribute models than by discounting models (He et al., 2023). As we shall see later, the analysis of behavioral data demonstrated the same pattern. The ratings by the three LLMs were also highly correlated (Figure 2), indicating consistency of strategy rating across LLMs.

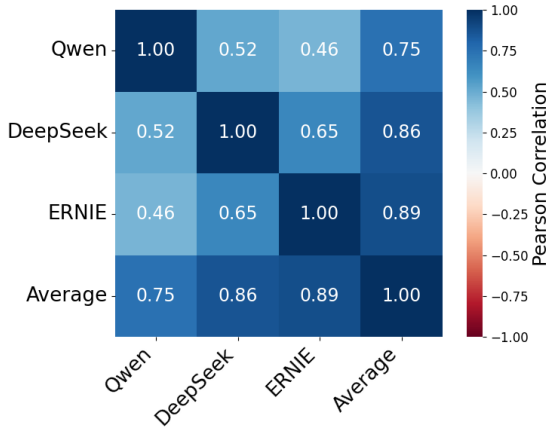

Figure 2: Correlations of LLM-coded decision strategies across different LLMs in Experiment 1.

Decision Strategies Identification Via Behavioral Modeling The traditional way to uncover decision strategies is through model selection using behavioral choice data. Participants are typically assumed to make decisions with the best-performing models. In this case, if a participant's choice data were best described by time-as-attribute models, they tended to align with attribute-based choice. In contrast, if a participant's choice data were best described by discounting models, they tended to align with alternative-based choice. Time-as-attribute models typically described individual participants' choice data better than discounting models,with over 85% of participants best fit by a time-as-attribute. This is not only consistent with previous work (Dai & Busemeyer, 2014; Ericson et al., 2015; He et al., 2023), but is also consistent with the LLM-coded decision strategies from the think-aloud data.

Individual-Level Agreement Between LLM-Coded Strategies and Behavioral Modeling Benchmark Bayesian model selection at the individual level allowed us to reveal each individual participant's decision strategies from their choice data. We used the average log Bayes factor between time-as-attribute models and discounting models, $\log \mathrm{BF}_{AD}$, as a behavioral benchmark for strategy identification. A high $\log \mathrm{BF}_{AD}$ indicates that the participant tended to adopt attribute-based strategies while a low $\log \mathrm{BF}_{AD}$ indicates that the participant tended to adopt alternative-based (discounting) strategies.

We then correlated the LLM-coded strategy scores (from the think-aloud data) with the behavioral benchmark. We found that the LLM-coded scores were moderately correlated with the behavioral signature identified from Bayesian model selection. This agreement between think-aloud data and behavioral data holds regardless of the LLM used for strategy scoring (Figure 3). Note that the correlational patterns remained robust even if we excluded outliers (i.e. participants beyond ±2 SD from the group mean for either variable) from the analysis.

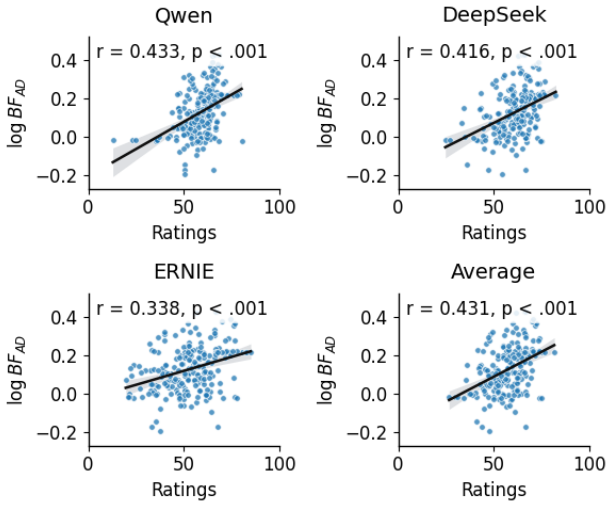

Figure 3: Correlations between LLM-derived decision strategies and behavioral modeling benchmarks in Experiment 1.

Discussion

Experiment 1 tested the validity of the LLM-based strategy scoring approach under a free-choice condition. We found that LLM-based strategy scoring successfully predicts two behavioral patterns of decision strategies in intertemporal choice. First, it predicted that a majority of participants tend to align more with attribute-based evaluation than with alternative-based evaluation in intertemporal decision making. Second, at the individual level, LLM-based strategy scoring correlated with the behavioral signature of alternative-based vs. attribute-based strategy from Bayesian model selection. These results present strong evidence that think-aloud data effectively reveal the latent decision strategies people use in their choice behavior in the free-choice condition.

Experiment 2

Experiment 2 was designed to test whether the LLM-coded decision strategies from the think-aloud data can detect the switch of strategies due to external instructions. Specifically, we explicitly asked participants to adopt

either an alternative-based or attribute-based decision regardless of their own preferences.

Methods

Participants In Experiment 2, two hundred participants were recruited on Credemo using a between-subjects design (61.5% female; mean age = 28.58, SD = 6.30). Participants were randomly assigned to one of two conditions: the attribute-based instruction condition and the alternative-based instruction condition, with 100 participants in each group. A total of 192 participants submitted usable think-aloud recordings, yielding a pass rate of 94%. After the same screening procedure as in Experiment 1, we were left with 188 participants for formal data analysis.

Stimuli and Procedures The choice stimuli for Experiment 2 were the same as in Experiment 1. The procedures were similar to those in Experiment 1, except for the strategy prompts. Specifically, the alternative-based instruction condition required participants to evaluate the overall value of each option separately and avoid directly weighing the relative strength of the two attributes. By contrast, the attribute-based instruction condition required participants to weigh the relative strength of the two attributes and to avoid evaluating the overall value of each option separately. These requirements were highlighted in all trials throughout the experiment.

Data Processing Transcription of the think-aloud recordings, text screening, LLM scoring, and Bayesian model selection on behavioral data were identical to those used in Experiment 1. After the data screening, a total of 9,852 valid think-aloud texts from 192 participants were retained, including 4,989 from the alternative-based instruction condition and 4,863 from the attribute-based instruction condition. Four participants with less than five valid think-aloud texts were further removed, leaving 188 participants for formal data analysis. The analysis of think-aloud and choice data followed the same procedures as in Experiment 1.

Results

LLM-Coded Decision Strategies From Think-Aloud Protocols We began by testing the differences in the LLM-coded strategies from the think-aloud data. As shown in Figure 3, all three LLMs suggested that participants in the attribute-based instruction condition were more aligned with the attribute-based evaluation rule (all ps < .001). This suggests that LLMs can capture the switch of decision strategies due to the instructions.

Decision Strategies Identification Via Behavioral Modeling Like in Experiment 1, we again used the comparison between discounting models and time-as-attribute models to reveal participants' decision strategies from choice data. Time-as-attribute models again described individual participants' choice data better than discounting models in both conditions in Experiment 2. A total of 82.5% of the participants were best fit by a time-as-attribute model than a discounting model, despite the alternative-based instructions in one of the experimental conditions. A breakdown according to the instruction conditions shows that 74 of the 93 participants (79.6%) in the alternative-based instruction condition were best fit by a time-as-attribute model and 81 of the 95 participants (85.3%) in the attribute-based instruction condition were best fit by a time-as-attribute model.

With the average log Bayes factor between time-as-attribute models and discounting models, $\log \mathrm{BF}_{AD}$, as a behavioral benchmark for strategy identification, we tested the differences in behavioral strategies between the attribute-based and alternative-based instruction conditions. However, we found that participants under attribute-based instructions were slightly more likely to adopt attribute-based strategies (vs. alternative-based discounting strategies), as measured by $\log \mathrm{BF}_{AD}$, although the difference between the two conditions did not reach the conventional statistical significance threshold (t = 0.98, p = .327, see Figure 4, bottom-right panel).

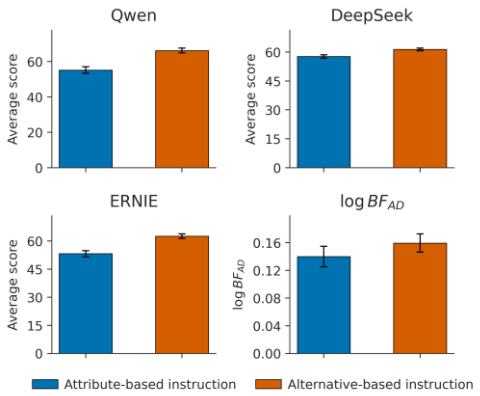

Figure 4: Between-condition comparison of decision strategies from LLM-coded think-aloud data and behavioral modeling in Experiment 2.

Individual-Level Agreement Between LLM-Coded Strategies and Behavioral Modeling Benchmarks We then correlated the LLM-coded strategies (from the think-aloud data) with the behavioral benchmark. We found that the LLM-coded scores were moderately correlated with the behavioral signature identified from Bayesian model selection. This agreement between think-aloud data and behavioral data holds regardless of the LLMs used for strategy scoring (Figure 5). All correlation coefficients were statistically significant, indicating stable convergence between the LLM-derived decision strategies and the behavioral modeling benchmark.

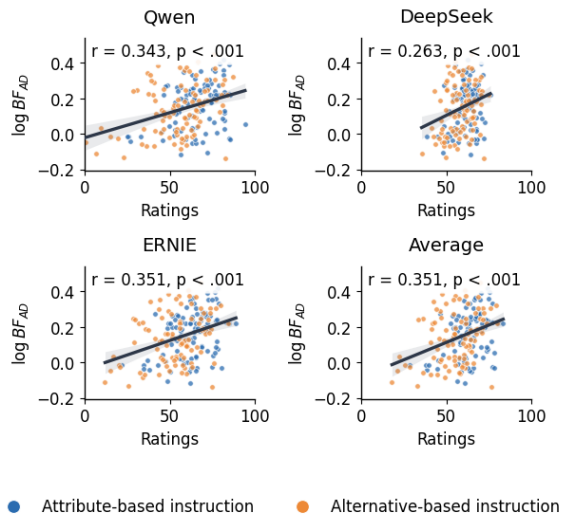

Figure 5: Correlations between LLM-derived decision strategies and behavioral modeling benchmarks in Experiment 2.

General Discussion

Human decision strategies are central to behavioral and social sciences. It has spurred decades of transdisciplinary work from cognitive science, psychology, economics and management. Previous work primarily relies on behavioral choice data to infer decision strategies undertaken by participants. The advent of commercial LLMs allows us to decode the strategies from think-aloud data during decision making. We evaluate the feasibility of this approach in intertemporal choice regarding alternative-based evaluation vs. attribute-based evaluation. Over two experiments, we show that LLM-derived strategies are moderately correlated with the behavioral strategies identified from choice data (via Bayesian model selection). We also found that the LLM-derived decision strategies from the think-aloud data vary with task instructions in the predicted direction. These results collectively show that the proposed pipeline can derive latent decision strategies from think-aloud data.

Integrating Think-Aloud Protocols With LLMs

Think-aloud analysis has a long history in cognitive science (Ericsson & Simon, 1980; Newell & Simon, 1972; van Solomon et al., 1994). It provides direct access to participants' ongoing thought processes during task performance, rather than relying solely on retrospective accounts or behavioral outcomes (Ericsson & Simon, 1993). This capability to capture real-time cognitive activity has proven invaluable across diverse research domains, from investigating decision-making strategies and problem-solving heuristics to studying expertise development and reasoning processes (Eccles & Arsal, 2017; Fonteyn et al., 1993).

Yet, the advantages of think-aloud data were compromised by its laborious data processing. The information coding procedure has significantly limited its usage in modern cognitive science. Large language models' excellent humanlike capability for understanding, and responding to, textual inputs makes them ideal for scalable coding of the rich information from think-aloud protocols.

Decision Strategies in Intertemporal Choice

In this work, we focus on latent decision strategies underlying human intertemporal choice. Specifically, one major dispute in intertemporal choice research concerns whether intertemporal choice is made in an alternative-based or attribute-based manner (He et al., 2022). The economic tradition mostly adopts an alternative-based view, assuming that each option is assigned a utility value and that the option with the highest value is chosen (Loewenstein & Prelec, 1992; Samuelson, 1937). However, recent experimental work increasingly favors an attribute-based evaluation rule, which suggests that people treat time delays as a decision attribute like monetary amounts and the decisions are made based on the trade-off between the advantages and disadvantages of the options along both attributes (Dai & Busemeyer, 2014; Ericson et al., 2015; Read et al., 2013; Scholten & Read, 2010).

The discrimination between alternative-based and attribute-based models typically relies on behavioral choice data and different individuals may be disposed differently along the spectrum between alternative-based and attribute-based choice (He et al., 2023). Our LLM-derived strategies validate the strategies identified from behavioral data, showing convergent validity across multiple modalities of data.

Limitations and Future Directions

Our work has limitations. First, the behavioral benchmark via Bayesian model selection did not differ significantly between the two conditions (though the overall tendency is consistent with our theoretical prediction). One possibility is that the set of stimuli used in the experiments was not sensitive enough to elicit the behavioral differences due to instructions. In other words, changes in decision strategy due to instructions do not necessarily translate into observable differences in choice outcomes. Future work should curate stimuli with behavioral responses being more sensitive to task instructions or design new manipulations that more cleanly separate alternative- vs. attribute-based strategies.

Second, human decision strategies involve much more nuances than the alternative-based vs. attribute-based distinction. Although this work focuses on this distinction, the same approach can be applied to other dimensions of decision strategies and other decision tasks (Fulawka et al., 2025). The rich information in think-aloud data holds the promise to expand the scope of modeling human decision strategies that has been missing in behavioral decision research. As LLMs continue to advance, their integration with think-aloud data will become an increasingly valuable tool for bridging the gap between verbal cognition and formal models of decision making.

Acknowledgments

We thank Yuntian Chai for assistance in implementing LLM scoring. The authors acknowledge financial support from the National Natural Science Foundation of China (No. 72571166) and appreciate the High Performance Computing Center of Shanghai University, and Shanghai Engineering Research Center of Intelligent Computing System (No. 19DZ2252600) for providing the computing resources and technical support.

References

Baker, F., Johnson, M. W., & Bickel, W. K. (2003). Delay discounting in current and never-before cigarette smokers: Similarities and differences across commodity, sign, and magnitude. *Journal of Abnormal Psychology*, *112*(3), 382–392. https://doi.org/10.1037/0021-843X.112.3.382

Benhabib, J., Bisin, A., & Schotter, A. (2010). Present-bias, quasi-hyperbolic discounting, and fixed costs. *Games and Economic Behavior*, *69*(2), 205–223. https://doi.org/10.1016/j.geb.2009.11.003

Benzion, U., Rapoport, A., & Yagil, J. (1989). Discount Rates Inferred from Decisions: An Experimental Study. *Management Science*, *35*(3), 270–284. https://doi.org/10.1287/mnsc.35.3.270

Berns, G. S., Laibson, D., & Loewenstein, G. (2007). Intertemporal choice – toward an integrative framework. *Trends in Cognitive Sciences*, *11*(11), 482–488. https://doi.org/10.1016/j.tics.2007.08.011

Bhatia, S., Yang, A., & Cohen, T. R. (2026). The Structure of Social Situations: Insights From the Large-Scale Automated Coding of Text. *Psychological Science*, *37*(3), 180–205. https://doi.org/10.1177/09567976261418946

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language Models are Few-Shot Learners* (No. arXiv:2005.14165). arXiv. https://doi.org/10.48550/arXiv.2005.14165

Dai, J., & Busemeyer, J. R. (2014). A probabilistic, dynamic, and attribute-wise model of intertemporal choice. *Journal of Experimental Psychology: General*, *143*(4), 1489–1514. https://doi.org/10.1037/a0035976

Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., Eichstaedt, J. C., Hecht, C., Jamieson, J., Johnson, M., Jones, M., Krettek-Cobb, D., Lai, L., JonesMitchell, N., Ong, D. C., Dweck, C. S., Gross, J. J., & Pennebaker, J. W. (2023). Using large language models in psychology. *Nature Reviews. Psychology*, *2*(11), 688–701. https://doi.org/10.1038/s44159-023-00241-5

Eccles, D. W., & Arsal, G. (2017). The think aloud method: What is it and how do I use it? *Qualitative Research in Sport, Exercise and Health*, *9*(4), 514–531. https://doi.org/10.1080/2159676X.2017.1331501

Ericson, K. M., White, J. M., Laibson, D., & Cohen, J. D. (2015). Money Earlier or Later? Simple Heuristics Explain Intertemporal Choices Better Than Delay Discounting Does. *Psychological Science*, *26*(6), 826–833. https://doi.org/10.1177/0956797615572232

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, *87*(3), 215–251. https://doi.org/10.1037/0033-295X.87.3.215

Ericsson, K. A., & Simon, H. A. (1993). *Protocol Analysis: Verbal Reports as Data*. The MIT Press. https://doi.org/10.7551/mitpress/5657.001.0001

Feng, H., Du, S., Zhu, G., Zou, Y., Phua, P. B., Feng, Y., Zhong, H., Shen, Z., & Liu, S. (2024). Leveraging Large Language Models for Automated Chinese Essay Scoring. In A. M. Olney, I.-A. Chounta, Z. Liu, O. C. Santos, & I. I. Bittencourt (Eds.), *Artificial Intelligence in Education* (pp. 454–467). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-64302-6_32

Fonteyn, M. E., Kuipers, B., & Grobe, S. J. (1993). A Description of Think Aloud Method and Protocol Analysis. *Qualitative Health Research*, *3*(4), 430–441. https://doi.org/10.1177/104973239300300403

Frederick, S., Loewenstein, G., & O'Donoghue, T. (2002). Time discounting and time preference: A critical review. *Journal of Economic Literature*, *40*(2), 351–401.

Fulawka, K., Hertwig, R., & Wulff, D. U. (2025). *Large language models accurately identify decision reasons in verbal reports*.

Green, L., Fristoe, N., & Myerson, J. (1994). Temporal discounting and preference reversals in choice between delayed outcomes. *Psychonomic Bulletin & Review*, *1*(3), 383–389. https://doi.org/10.3758/BF03213979

He, L., Golman, R., & Bhatia, S. (2019). Variable time preference. *Cognitive Psychology, 111*, 53–79. https://doi.org/10.1016/j.cogpsych.2019.03.003

He, L., Wall, D., Reeck, C., & Bhatia, S. (2023). Information acquisition and decision strategies in intertemporal choice. *Cognitive Psychology*, *142*, 101562. https://doi.org/10.1016/j.cogpsych.2023.101562

He, L., Zhao, W. J., & Bhatia, S. (2022). An ontology of decision models. *Psychological Review*, *129*(1), 49–72. https://doi.org/10.1037/rev0000231

Holcomb, J. H., & Nelson, P. S. (1992). Another Experimental Look at Individual Time Preference. *Rationality and Society*, *4*(2), 199–220. https://doi.org/10.1177/1043463192004002006

Kass, R., & Raftery, A. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

Larkin, J., McDermott, J., Simon, D. P., & Simon, H. A. (1980). Expert and Novice Performance in Solving Physics Problems. *Science*, *208*(4450), 1335–1342. https://doi.org/10.1126/science.208.4450.1335

Loewenstein, G., & Prelec, D. (1992). Anomalies in Intertemporal Choice: Evidence and an Interpretation. *The Quarterly Journal of Economics*, *107*(2), 573–597. https://doi.org/10.2307/2118482

Morris, A., Carlson, R. W., Kober, H., & Crockett, M. J. (2025). Introspective access to value-based multi-attribute choice processes. *Nature Communications*,

16(1), 3733. https://doi.org/10.1038/s41467-025-59080-y

Newell, A., & Simon, H. A. (with Simon, H. A.). (1972). *Human problem solving*. Prentice-Hall.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1988). Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(3), 534–552. https://doi.org/10.1037/0278-7393.14.3.534

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The Adaptive Decision Maker*. Cambridge University Press.

Petry, N. M. (2001). Delay discounting of money and alcohol in actively using alcoholics, currently abstinent alcoholics, and controls. *Psychopharmacology*, *154*(3), 243–250. https://doi.org/10.1007/s002130000638

Radford, A., Kim, J. W., Xu, T., Brockman, G., Mcleavey, C., & Sutskever, I. (2023). Robust Speech Recognition via Large-Scale Weak Supervision. *Proceedings of the 40th International Conference on Machine Learning*, 28492–28518. https://proceedings.mlr.press/v202/radford23a.html

Read, D., Frederick, S., & Scholten, M. (2013). DRIFT: An analysis of outcome framing in intertemporal choice. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(2), 573–588. https://doi.org/10.1037/a0029177

Reeck, C., Wall, D., & Johnson, E. J. (2017). Search predicts and changes patience in intertemporal choice. *Proceedings of the National Academy of Sciences*, 114(45), 11890–11895. https://doi.org/10.1073/pnas.1707040114

Samuelson, P. A. (1937). A Note on Measurement of Utility. *The Review of Economic Studies*, *4*(2), 155–161. https://doi.org/10.2307/2967612

Scholten, M., & Read, D. (2010). The psychology of intertemporal tradeoffs. *Psychological Review*, *117*(3), 925–944. https://doi.org/10.1037/a0019619

Scholten, M., Read, D., & Sanborn, A. (2014). Weighing Outcomes by Time or Against Time? Evaluation Rules in Intertemporal Choice. *Cognitive Science*, *38*(3), 399–438. https://doi.org/10.1111/cogs.12104

Scholten, M., Read, D., & Sanborn, A. (2016). Cumulative weighing of time in intertemporal tradeoffs. Journal of Experimental Psychology: General, 145(9), 1177–1205. https://doi.org/10.1037/xge0000198

Sun, Y., Wang, S., Feng, S., Ding, S., Pang, C., Shang, J., Liu, J., Chen, X., Zhao, Y., Lu, Y., Liu, W., Wu, Z., Gong, W., Liang, J., Shang, Z., Sun, P., Liu, W., Ouyang, X., Yu, D., … Wang, H. (2021). *ERNIE 3.0: Large-scale Knowledge Enhanced Pre-training for Language Understanding and Generation* (No. arXiv:2107.02137). arXiv. https://doi.org/10.48550/arXiv.2107.02137

Suter, R. S., Pachur, T., & Hertwig, R. (2016). How Affect Shapes Risky Choice: Distorted Probability Weighting Versus Probability Neglect. *Journal of Behavioral Decision Making*, *29*(4), 437–449. https://doi.org/10.1002/bdm.1888

Thaler, R. (1981). Some empirical evidence on dynamic inconsistency. *Economics Letters*, *8*(3), 201–207. https://doi.org/10.1016/0165-1765(81)90067-7

van Solomon, M. W., Barnard, Y. F., & Sandberg, J. A. C. (1994). The think aloud method: A practical guide to modelling cognitive processes. *Information Processing & Management*, *31*(6), 906–907. https://doi.org/10.1016/0306-4573(95)90031-4

Weber, E. U., Johnson, E. J., Milch, K. F., Chang, H., Brodscholl, J. C., & Goldstein, D. G. (2007). Asymmetric Discounting in Intertemporal Choice: A Query-Theory Account. *Psychological Science (0956-7976)*, *18*(6), 516–523. https://doi.org/10.1111/j.1467-9280.2007.01932.x

Wu, J., Wang, Z., & Qin, Y. (2025). Performance of DeepSeek-R1 and ChatGPT-4o on the Chinese National Medical Licensing Examination: A Comparative Study. *Journal of Medical Systems*, *49*(1), 74. https://doi.org/10.1007/s10916-025-02213-z

Wurgaft, D., Prystawski, B., Gandhi, K., Zhang, C. E., Tenenbaum, J. B., & Goodman, N. (2025). Scaling up the think-aloud method. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *47*(0). https://escholarship.org/uc/item/8jc6r7h5

Xie, H., Xiong, H.-D., & Wilson, R. C. (2025). *Rethinking Think-Aloud in the Age of Language Models: Verbally reported thoughts are predictive of behavior*.

Zhao, M. (2023). The Application of iFLYTEK FiF Speaking Training System in Language Learning. *Proceedings of the 2022 5th International Conference on Machine Learning and Natural Language Processing*, 175–178. https://doi.org/10.1145/3578741.3578777