Do LLMs Choose Like Humans? Using Cognitive Theory to Evaluate LLM Decision-Making

Johnathan Sun[1,2] **Andrei Shleifer**[1] **Yonatan Belinkov**[2,3]
[1]Harvard University [2]Kempner Institute, Harvard University
[3]Technion - Israel Institute of Technology
jlsun@college.harvard.edu shleifer@fas.harvard.edu
belinkov@technion.ac.il

Abstract

Large language models (LLMs) exhibit a range of human-like decision-making behaviors, but whether these reflect similar underlying mechanisms or surface-level mimicry remains unclear. We evaluate whether LLM context sensitivity aligns with a cognitive economic theory that explains human behavior through problem categorization and attention allocation. Across 12 open-source and commercial LLMs on a novel 140,000-trial product choice benchmark, context induces human-like shifts in choice and problem categorization, but does not reliably reweight attention between features like price and quality. Neither scale nor chain-of-thought reasoning reliably attenuates context sensitivity or generates human-like behavior. These results suggest that LLM decision mechanisms are distinct from human ones.

1 Introduction

Large language models (LLMs) increasingly make consequential economic decisions, but their underlying decision-making mechanisms remain poorly understood.

Behavioral evaluations have shown that models exhibit human-like preferences and cognitive biases (Malberg et al., 2025; Lior et al., 2026; Cook et al., 2026). Researchers have fit canonical economic models, such as prospect theory, on LLM behavior, but these theories offer no independent tests of mechanism and have so far shown limited explanatory power (Becker, 1962; Manski, 1995; Ross et al., 2024; Wang et al., 2026).

Recent work in cognitive economics has made progress towards a unified explanation of human decision-making, including choice anomalies, cognitive biases, and preference instability (Bordalo et al., 2012, 2013, 2020). We evaluate whether LLM decisions align with the cognitive theory of choice from Bordalo et al. (2026b), and, by extension, the extent to which LLM decision-making is behaviorally and mechanistically consistent with humans.

We explore a central feature of decision-making: context sensitivity. Many behaviors, from canonical biases to phenomena traditionally modeled as belief biases or preference primitives, are instances of context sensitivity, so a sharp characterization of it in LLMs could help explain and predict behavior across many settings (Gabaix, 2019; Jones and Steinhardt, 2022; Enke and Graeber, 2023).

Is a jar of jam a luxury or a staple? Are stocks opportunities or gambles? Bordalo et al. (2026b) argue that humans make decisions by first characterizing the problem at hand: A decision-maker (DM) attends to context features like location and time to find the most similar category of choices made in the past. The selected category causes the DM to differentially attend to choice features, such as price and quality, which forms a mental representation of the current decision that is used to make a choice. Context influences how a person thinks about a problem, an explanation that unifies major biases in judgment and decision-making.

To evaluate whether LLMs exhibit behavior consistent with these mechanisms in a canonical economic setting, we introduce a 140,000-trial product choice dataset that experimentally varies context via cues that induce hedonic (quality-focused) or functional (price-focused) consumption in humans. We measure context effects on choices and feature sensitivities across 12 open-source and commercial LLMs, finding that:

1. **Context effects on choice align with humans, but effects on feature sensitivities largely do not.** Cues have substantial effects on choice rates across all models, but only a quarter of model-cue combinations exhibit predicted shifts in feature sensitivity.

2. **Context effects follow a problem categorization process.** Cue effects peak where catego-

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 46552–46569
October 24-29, 2026 ©2026 Association for Computational Linguistics

rization is ambiguous and shrink as baseline decisions grow more certain. Cues on prior in-context items predictably shift choices on a neutral test question.

3. **Scale and reasoning do not reliably attenuate context effects.** Context sensitivity varies with scale across model families, remains high in frontier models, and can increase with chain-of-thought reasoning, contrasting with bounded-cognition accounts of human bias.

2 Related Work

LLM decision-making and cognitive economics. LLMs exhibit a range of canonical biases and decision-making behaviors (Jones and Steinhardt, 2022; Echterhoff et al., 2024; Cheung et al., 2025; Malberg et al., 2025), and researchers have fit economic models on LLM behavior (Chen et al., 2023; Ross et al., 2024; Bini et al., 2026). We utilize a cognitive economic model that unifies context effects usually explained with domain-specific theories, which typically describe rather than mechanistically explain behavior, and often face documented experimental violations (Plott and Zeiler, 2005; Barr et al., 2009; Bernheim and Taubinsky, 2018; Dean and Neligh, 2023).

Cognitive economic models are potentially well-suited to LLMs. Such models are built from mechanisms of attention, memory, salience, and internal representations (Bordalo et al., 2012, 2016, 2020, 2026b), each with a natural analogue in transformers, which dynamically allocate attention across tokens (Vaswani et al., 2017), use feed-forward layers as associative memory (Geva et al., 2021), and form context-sensitive hidden representations that shape decisions (Sun and Zhang, 2026).

Structure of context sensitivity. Models are highly sensitive to prompt content, including implicit demographic cues (Gupta et al., 2024; Pawar et al., 2025), affective framing (Lior et al., 2026), and spurious syntax modifications (Sclar et al., 2024). Our dataset is the first to measure the effects of hedonic and functional contexts, a longstanding distinction in the consumer choice literature (Dhar and Wertenbroch, 2000).

Similarity between LLMs and humans. Researchers in cognitive and social science increasingly use LLMs as experimental proxies for humans (Anthis et al., 2025; Manning and Horton, 2026; Park et al., 2026). However, human and LLM behavior often diverge in economic settings, and heuristic approaches such as prompt engineering or fine-tuning often fail to recover human-like behavior (Gui and Toubia, 2025; Gao et al., 2025; Hullman et al., 2026). To the extent that Bordalo et al. (2026b) accurately describe humans, our results imply that underlying decision mechanisms in LLMs are distinct.

3 Behavioral Framework

Decision-makers weigh price and quality. We evaluate LLMs on a fundamental economic setting: choosing between products. The tradeoff between price and quality is a canonical testbed in economics, with extensive human data and problem features that are easy to vary (Rosen, 1974; Bordalo et al., 2013). Moreover, LLMs increasingly encounter similar price-quality tradeoffs as they become full-fledged economic agents (Schumacher et al., 2025; Stripe, 2025).

Our benchmark consists of product choice questions; we use the running example of purchasing jam to explain the mechanisms at play. Suppose a DM faces a problem D in which they choose between a standard jar of jam ($j = 0$) and a fancy one ($j = 1$). Jam j has price p_j and quality q_j, with $p_1 > p_0$, $q_1 > q_0$. Let the DM's valuation of jam j be

$$u_j(\alpha^*) = \alpha_Q q_j - \alpha_P p_j, \tag{1}$$

where $\alpha^* = (\alpha_Q, \alpha_P)$ are context-dependent attention weights on price and quality. The DM chooses the jam with the higher valuation, which is the fancy jam iff $\alpha_Q(q_1 - q_0) > \alpha_P(p_1 - p_0)$.

Context affects price-quality tradeoff. Bordalo et al. (2026b) explain context-sensitive behavior in humans by specifying how attention weights (α_Q, α_P) are pinned down in the mental representation of a decision.

When is jam a luxury or a staple? Consumer psychology models propose two ways of looking at the problem (Hirschman and Holbrook, 1982; Dhar and Wertenbroch, 2000):

1. **Hedonic consumption** implicates affective or sensory experiences associated with quality sensitivity: Jam as a luxury.

2. **Functional consumption** implicates accomplishing a practical task associated with price sensitivity: Jam as a staple.

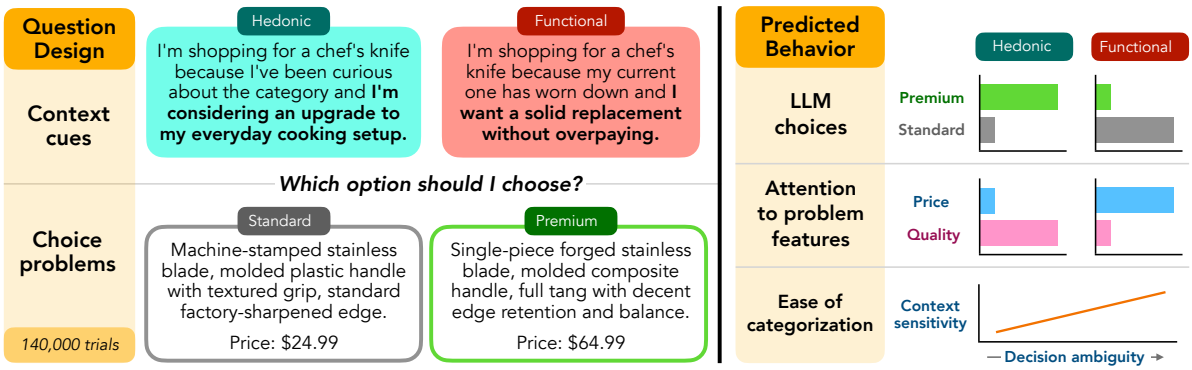


Figure 1: **Product choice questions evaluate LLM decisions against human behavior and cognitive theory.** (**Left**) Hedonic and functional context cues, drawn from the consumer choice literature, frame a binary choice between standard and premium products. (**Right**) Given human behavior explained by Bordalo et al. (2026b), we predict that hedonic framing shifts LLM choices towards the premium option, reweights attention from price to quality, and produces larger effects when baseline categorization is ambiguous.

The core intuition is that context influences whether the DM views a decision as hedonic or functional consumption, which changes how the DM weighs price and quality.

For example, the shopping venue influences whether humans characterize their purchasing as functional or hedonic (Hamilton and Chernev, 2013). To capture this effect, let κ_D denote the scalar context of the current problem D. Suppose $\kappa_D = 1$ in an organic farm store (hedonic), and $\kappa_D = 0$ in a convenience store (functional). Let F_{hed} and F_{fun} denote the frequencies of hedonic and functional consumption in past problems, which represent experiences stored in memory. The DM chooses a jam to purchase in two steps:

1. **Problem categorization**: The DM selects a category from memory that is most contextually similar to the current problem. Here, the DM categorizes D as hedonic if $\kappa_D > \kappa(F_{\text{fun}}/F_{\text{hed}})$ and functional otherwise, where $\kappa(\cdot)$ is increasing. Intuitively, the DM is more likely to regard the jam decision as hedonic consumption in the farm store because they were hedonic there in the past.

2. **Attention reweighting**: The DM forms a mental representation with the selected category, which influences relative attention to decision features. If D is hedonic, $\alpha_Q > \alpha_P$, so the DM is more sensitive to quality than price. If D is functional, $\alpha_P > \alpha_Q$.

Model implications. Hedonic/functional categorization predicts attribute weighting, price elasticity, and purchase justification in humans (Wakefield and Inman, 2003; Okada, 2005; Khan and Dhar, 2006). We evaluate LLMs against three human phenomena explained by the cognitive model:

1. **Context affects choices**: Under Equation (1), hedonic context raises α_Q relative to α_P, which shifts choices near indifference towards the premium option. The DM is more likely to choose the fancy jam in the farm store, and vice versa in the convenience store.

2. **Context induces asymmetric feature sensitivity**: Hedonic context raises the marginal rate of substitution α_Q/α_P, the amount of additional price the DM accepts for a marginal gain in quality. Intuitively, the DM's decision is relatively more sensitive to jam qualities than jam prices in the farm store.

3. **Context effects depend on ease of categorization**: Holding context fixed, the DM is more likely to categorize D as hedonic if $F_{\text{fun}}/F_{\text{hed}}$ is smaller, which occurs when the hedonic category has been used relatively more often than the functional category. The DM is more likely to view jam as a luxury if they made relatively more recent hedonic purchases.

4 Experimental Design

Baseline product choice problems. To evaluate LLM behavior, we design a 140,000-trial product choice dataset. We use Claude Sonnet-4.6 to generate 100 products ("chef's knife") from 20 sectors ("kitchen & food"). For each product, Claude generates five options that gradually increase in price

Cue Type	Hedonic Cue	Functional Cue	References
Purchase trigger	Exploring the category out of curiosity or considering an upgrade	Routine replacement or replenishment without overpaying	Bell et al. (2011)
Retail architecture	Browsing a curated specialty page with staff picks and quality notes	Browsing a marketplace comparison page sorted by price with filters	Koschmann and Isaac (2018)
Payment source	Paying with a gift card or credit, keeping any unspent balance	Paying from a checking account, keeping any unspent balance	Thaler (1999)
Justification	Recently earned a small reward; this feels like a reasonable treat	Ordinary purchase with no special reason; just a practical choice	Okada (2005)
Reasoning frame	Focused on why the product would improve the user experience	Focused on how to complete the purchase and the concrete tradeoffs	Liberman and Trope (1998)

Table 1: **Contextual purchasing cues.** Each cue type includes a hedonic/functional pair drawn from the consumer choice literature. Claude Sonnet-4.6 instantiates these templates into concrete cues for each product, designed to shift consumption framing without changing product utility.

p_t and descriptive quality d_t. Descriptions match in length (± 5 words) and attributes described per product, and brand names are omitted to limit associations with unobservable features like reputation.

To map description to quality, we use GPT 5.4 to generate 0-100 quality scores q_t for each option. GPT is employed so Claude does not self-evaluate its quality manipulations, though GPT and Claude scores are highly correlated ($r = 0.96$). We manually audited 10% of GPT scores to confirm that better-described options consistently received higher scores, and that score gaps tracked the size of described quality differences.

As in the jam example, we write questions where an LLM with context κ outputs a binary choice A/B between standard ($j = 0$) and premium ($j = 1$) options for a given product, with $p_1 > p_0$, $q_1 > q_0$. For each product, we pair the five options by adjacent quality, and include both option orders to account for position effects. This process yields 800 total pairs of product options.

Context cues drawn from hedonic/functional consumption literature. Context κ varies under three conditions: The **neutral** condition presents the statement "I need a {product}" before the question. The **hedonic** and **functional** conditions prepend one of five types of paired cues to the neutral statement. Claude Sonnet-4.6 generates cues using templates that induce or are associated with hedonic/functional consumption in the consumer choice literature, such as whether a DM intends to replace (functional) or upgrade (hedonic) their existing product (Table 1). With five types of paired hedonic/functional cues and one neutral condition, the baseline dataset contains 8,800 problems. Figure 1 presents an example problem, and Appendix Section A includes more technical dataset details.

Datasets to estimate feature sensitivity. We estimate the sensitivity of choices to problem features using two additional datasets that slightly perturb the price and quality of baseline questions. The **price sweep** dataset fixes descriptions and scales premium option prices by 10 evenly log-spaced factors from one to two, generating 88,000 problems. The **quality sweep** dataset fixes prices and varies premium option descriptions by five small differences in quality, adding a further 44,000 problems.

To investigate how cues shift feature sensitivity, we measure decisions as a function of log feature ratios, a form that maps the marginal rate of substitution α_Q/α_P onto a regression coefficient. For each model m and sweep $s \in \{\text{price}, \text{quality}\}$, we estimate the regression

$$\ell_i = \beta_H \mathrm{Hed}_i + \beta_R \log r_i^{(s)} + \beta_{HR} \mathrm{Hed}_i \log r_i^{(s)} + \alpha_{p(i)} + \gamma \mathrm{PF}_i + \varepsilon_i,$$

where ℓ_i is the logit difference between premium and standard options; $\mathrm{Hed}_i \in \{0, 1\}$ indicates a hedonic cue (functional is the reference); $\log r_i^{(\text{price})}$ is the log-multiplier applied to the premium price in trial i, while $\log r_i^{(\text{quality})} = \log(q_i^{\text{prem}}/q_i^{\text{std}})$ is the log ratio of premium-to-standard quality scores; $\alpha_{p(i)}$ are option pair and cue type fixed effects; and PF_i controls for option order. The coefficient β_R captures the model's sensitivity to feature ratios under functional cues, and β_H captures the average level shift induced by hedonic cues. The key coefficient is the interaction β_{HR}: the marginal change in feature sensitivity between hedonic and functional cues. Human evidence and theory predict $\beta_{HR} > 0$ in both price and quality sweeps, which corresponds to hedonic cues attenuating negative

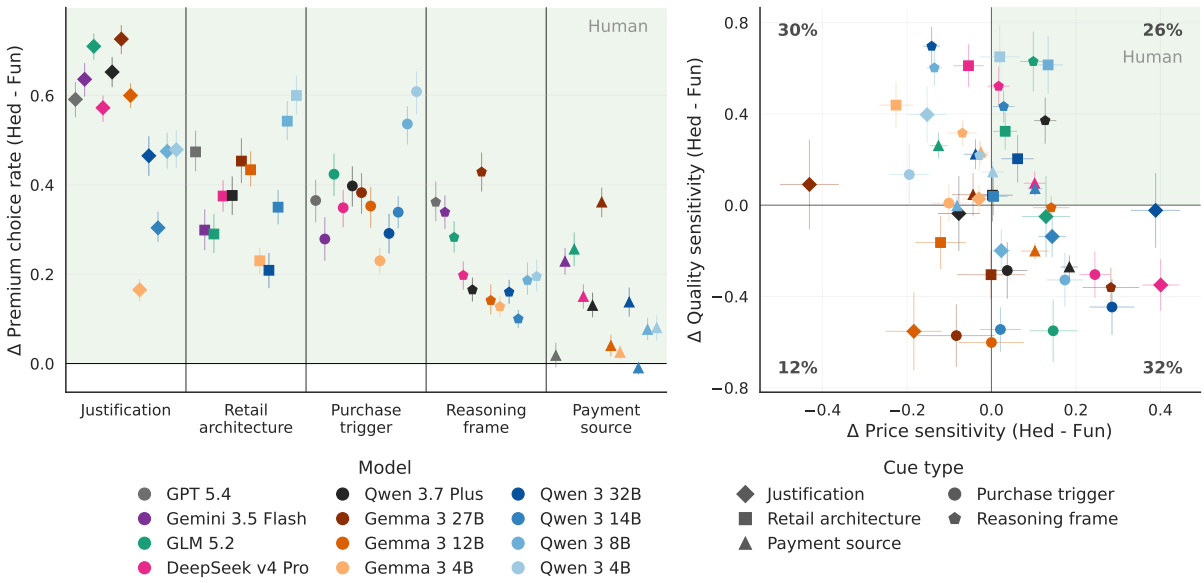


Figure 2: **Models match human context effects on choices, but diverge on feature sensitivity.** (**Left**) Difference in premium choice rates between hedonic and functional cues. Shaded region indicates human-like behavior for all panels—here, hedonic cues lead to more premium choices. See Appendix Figure 5 for the difference in premium choice rates relative to the neutral condition. (**Right**) Difference in standardized price (x-axis) and quality (y-axis) sensitivity between hedonic and functional cues, estimated by linear regression on choice logits (see Section 4). Shaded region denotes behavior where hedonic cues attenuate (negative) price sensitivity and amplify quality sensitivity. Percentages report the share of model-cue combinations in each quadrant. All bars are 95% CIs.

price sensitivity and amplifying positive quality sensitivity relative to functional cues.

Models. We evaluate 12 instruct-tuned open-source and commercial LLMs: GPT 5.4, Gemini 3.5 Flash, GLM 5.2, DeepSeek v4 Pro, Qwen 3.7 Plus, Gemma 3 (4B, 12B, and 27B), and Qwen 3 (4B, 8B, 14B, and 32B) (Google DeepMind, 2025; Gemma Team et al., 2025; Yang et al., 2025; OpenAI, 2026; GLM-5 Team et al., 2026; DeepSeek-AI et al., 2026). For the main experiments, we query each model once per question at temperature zero. GPT and Gemini, for which temperature controls and logits are unavailable, are run with medium (default) reasoning and excluded from the feature-sensitivity analysis in Section 5.2. In Section 5.4, we additionally query the four Qwen 3 models once per question in thinking mode with a 4096-token budget at temperature 0.6.

5 Comparing LLM and Human Behavior

5.1 Human-LLM similarity in choice rates

Every LLM is more likely to choose premium with hedonic cues than functional cues, consistent with context effects on human choice (Figure 2, left). This effect is significant ($p < 0.001$, one-sided paired t-test), robust across almost every model-cue combination, and substantial: The average LLM is more than twice as likely to choose premium with a hedonic cue as with a functional cue.

Effect sizes vary across cue types. The hedonic justification cue—buying a special treat instead of a normal purchase—raises premium choice rates the most relative to the corresponding functional cue (248% increase). The hedonic payment source cue—paying with a gift card instead of a checking account—has the smallest effect (36% increase), though similar interventions induce substantial changes in human behavior (Levav and McGraw, 2009; Helion and Gilovich, 2014). The moderate effect of hedonic reasoning frame cues is also unexpected, since these cues effectively instruct the LLM to adopt a consumption mode.

Finally, hedonic (functional) cues raise (lower) premium choice rates relative to the neutral condition for every LLM ($p < 0.05$, one-sided paired t-test), so both cue directions significantly impact choices. That said, the neutral premium choice rate is 2.8 times closer on average to the hedonic-cue rate than to the functional-cue rate. Hence, the hedonic-functional gap in premium choice is largely because functional cues alter premium-leaning "default" behavior.

5.2 Human-LLM contrast in feature sensitivity

Sensitivities diverge from human-like behavior. While LLMs demonstrate human-like context effects on choices, they do not reweight feature sensitivities similarly. Only 26% of model-cue combinations are less price sensitive and more quality sensitive in hedonic cues compared to functional, while 12% exhibit the opposite behavior in both (Figure 2, right). The remaining 62% of model-cue combinations exhibit human-like reweighting in one feature, but not the other. These distinct effects on feature sensitivity are robust to different regression specifications, such as estimating sensitivity to differences in feature values instead of ratios (Appendix Figure 6).

Context effects on feature sensitivity are highly heterogeneous. Effects vary significantly by cue type. For instance, Qwen 32B is more price and quality sensitive under hedonic reasoning frame cues, but less sensitive to both under hedonic purchase trigger cues. No LLM exhibits human-like sensitivity across all cue types.

Effects also vary by model: hedonic reasoning frame cues reduce both price and quality sensitivity in Gemma 27B, but raise both in Qwen 8B. No cue type induces human-like sensitivity effects in all LLMs, though the retail architecture cues are most successful (6/10 models) and the purchase trigger cues least (0/10 models). Effects on feature sensitivity span at least three quadrants of Figure 2 (right) for six out of ten LLMs.

Aggregating models does not recover human-like sensitivities. Context effects may appear stable across a human population but vary between individuals, domains, and problems (Liew et al., 2016). Similarly, while individual LLM behavior is noisy, population-level behavior may be more human-like (Hagendorff et al., 2023). However, aggregating LLMs for each cue type reveals that only reasoning frame cues induce significant human-like sensitivities, while only four out of ten models display this pattern when aggregating across cues (GLM 5.2; DeepSeek v4 Pro; Qwen 8B and 32B).

Despite diverging sensitivities, LLMs do choose coherently. Another concern is that sensitivity patterns reflect incoherent choices rather than meaningful differences from human behavior. However, feature sensitivities are tightly estimated, and LLM choices have meaningful structure. Every LLM exhibits negative price sensitivity and positive quality sensitivity under both hedonic and functional cues. Holding quality fixed, doubling the premium price ratio reduces premium choice by 15 pp. on average; holding price fixed, doubling the quality ratio raises it by 44 pp. While changes to feature sensitivities from context vary, LLMs do respond consistently to feature variation.

5.3 LLMs are sensitive to ease of categorization

Since LLMs do not exhibit human-like feature reweighting, we ask whether LLMs behave consistently with the preceding mechanism: problem categorization. Recall from Section 3 that a DM categorizes problem D as hedonic if $\kappa_D > \kappa(F_{\text{fun}}/F_{\text{hed}})$. If LLMs categorize problems, we expect three behaviors: (i) Directly manipulating $F_{\text{hed}}/F_{\text{fun}}$ in-context shifts choices. (ii) Context effects on choice attenuate in problems with obvious categorizations. (iii) Context effects on choice are largest when decisions are ambiguous, irrespective of category taxonomy. We find evidence of all three patterns across models.

In-context evidence shifts categorization. Raising $F_{\text{hed}}/F_{\text{fun}}$ lowers the threshold for categorizing a current problem as hedonic, which captures the intuition that humans are generally more likely to recall and act upon frequent and recent memories (Kahana, 2012; Bordalo et al., 2025; Jiang et al., 2025). To proxy memory, we manipulate $F_{\text{hed}}/F_{\text{fun}}$ in LLMs by prepending eight cued example problems before an un-cued test problem. The eight examples include $n_{\text{hed}} \in \{0, 1, \ldots, 8\}$ hedonic and $n_{\text{fun}} = 8 - n_{\text{hed}}$ functional cues of the same type, present different products from the test question, and omit choices to avoid pattern-matching decisions. Consistent with the theory, increasing n_{hed} from zero to eight raises the average premium choice rate from 49% to 74%, and a 10 pp. increase in the share of hedonic examples raises premium choice rates by 2.5 pp. on average (Figure 3, left).

Implicit associations moderate context effects. In humans, buying a fountain pen feels more hedonic than buying dish soap (Wakefield and Inman, 2003). Since LLMs train on vast amounts of human data, products themselves may be associated with types of consumption, generating a test of categorization (Caliskan et al., 2017; Li et al., 2024).

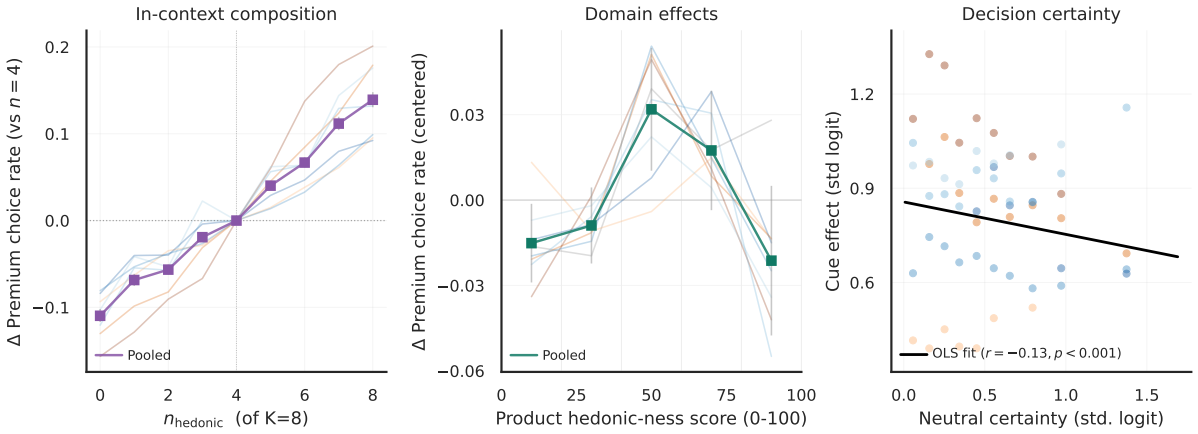


Figure 3: **Context effects follow a categorization process: in-context evidence shifts categorization (left), and ambiguity amplifies cue effects (center, right).** Panels pool across Gemma 3 (4B, 12B, 27B) and Qwen 3 (4B, 8B, 14B, 32B). (**Left**) Difference in premium choice rate relative to $n_{\text{hed}} = 4$ (midpoint) as a function of the number of hedonic cues among eight in-context examples. (**Center**) Difference in premium choice rate between hedonic and functional cues (de-meaned by model) binned by product hedonic-ness scores. (**Right**) Difference in (premium - standard logit) between hedonic and functional cues as a function of (premium - standard logit) in neutral choice problems, both standardized to neutral-SD units. All bars are 95% CIs.

We can decompose context as $\kappa_D(p, c) = x_p + \gamma_c$, where x_p is context conveyed by product p and γ_c is context from an explicit cue c. The cue can only change categorization if κ_D is sufficiently close to the threshold $\tau \in [x_p + \gamma_{\text{fun}}, x_p + \gamma_{\text{hed}}]$. For strongly hedonic or functional products, both cues produce the same categorization, attenuating their effect on choice.

To test this behavioral signature, we use a GPT 5.4-based judge to rate each product's name from 0 (fully functional) to 100 (fully hedonic), averaging over three trials (Asirvatham et al., 2026). We manually audited all scores, and Appendix Figure 7 includes the score distribution. Measuring tape and dish soap are the most functional products, while watercolor paints and greeting cards are the most hedonic. Behavior is consistent with our prediction: cue effects on choices across LLMs peak for products with midrange hedonic-ness ratings and shrink at both extremes (Figure 3, center).

Decision ambiguity amplifies context effects. The hedonic-functional analysis assumes that these are the two relevant problem categories. A weaker test of categorization is that context should matter most when an un-cued problem does not induce a stable representation and choice. We proxy categorization ambiguity with the decision certainty of LLMs under neutral cues, measured as the absolute logit difference between premium and standard options (larger = more certain). Across models, cue effects on choice rates are negatively associated with baseline certainty (Figure 3, right).

While our interpretation assumes that competing categories imply different choices (e.g., hedonic and premium, functional and standard), this behavioral pattern is hard to reconcile with cues generating fixed shifts in the price-quality tradeoff. We propose that when an LLM is uncertain about how to categorize a problem, a cue provides a greater update to its posterior beliefs. Similar logic motivates Bayesian and rational inattention accounts of context dependence, and is consistent with evidence in humans that comparison complexity and cognitive uncertainty drive context dependence and preference instability (Enke and Graeber, 2023; Shubatt and Yang, 2026).

5.4 Scale and reasoning do not consistently reduce context sensitivity

Bordalo et al. (2026b) explain human context sensitivity in part through limited attention, and a broader literature links such behaviors to bounded cognition and computational constraints (Stanovich and West, 2000; Gabaix, 2019; Maćkowiak et al., 2023). If the same constraint shapes LLM decisions, more capable models should be less context-sensitive. However, two margins of capability—scale and inference-time compute—do not consistently attenuate context effects on choices. Reasoning can recover human-like context effects on feature sensitivities, but only when cues explicitly

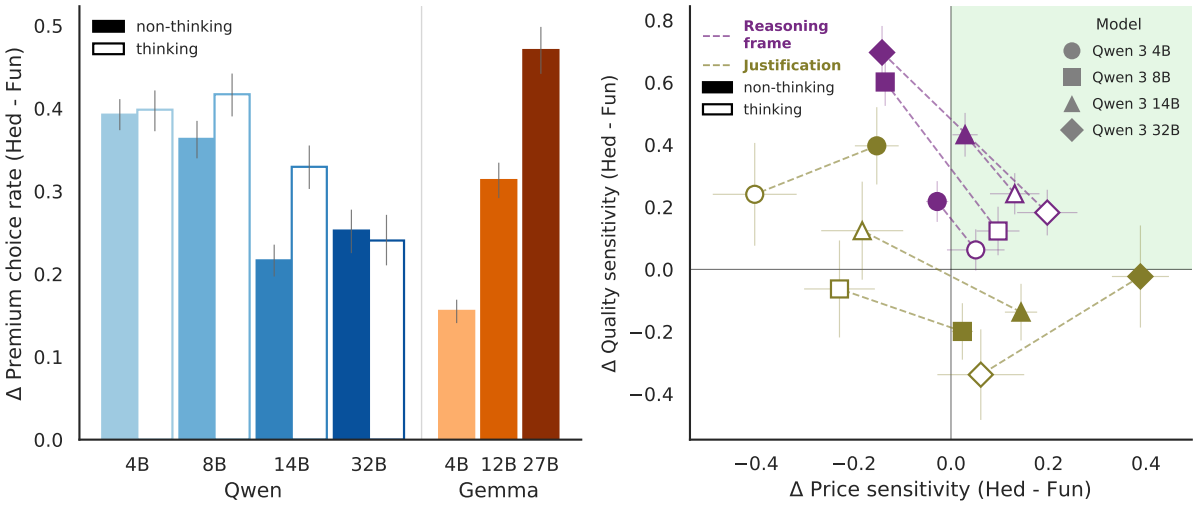


Figure 4: **Scale and reasoning do not reliably reduce context sensitivity.** (**Left**) Difference in premium choice rate between hedonic and functional cues for Qwen 3 (paired non-thinking/thinking) and Gemma 3 (non-thinking) model families. (**Right**) Difference in standardized price (x-axis) and quality (y-axis) sensitivity between hedonic and functional cues for Qwen models in thinking/non-thinking modes. Shaded region denotes human-consistent behavior. All bars are 95% CIs.

invoke the hedonic/functional dimension.

Context effects on choice vary with scale across LLM families. We evaluate context effects on choice for four Qwen 3 models and three Gemma 3 models. The two largest Qwen models are less sensitive than the two smaller ones, but Gemma 27B is roughly three times more sensitive than Gemma 4B (Figure 4, left). Context effects for GPT and Gemini, two commercial frontier LLMs, are also larger than those of Qwen 32B on average. One potential explanation is that post-training pressure—instruction tuning, RLHF, or training to infer user attributes—may work against any rationality gained with scale (Itzhak et al., 2024).

Reasoning recovers human-like behavior only when cues are explicit. We evaluate the four Qwen 3 models in thinking mode. Chain-of-thought reasoning does not reduce context effects on choice and significantly increases them for Qwen 8B and 14B (Figure 4, left).

The effect of reasoning on feature sensitivity varies by cue. The reasoning frame cue, which effectively instructs an LLM to engage in hedonic or functional consumption, induces human-like sensitivity effects in all four Qwen models only in thinking mode (Figure 4, right). The effect of reasoning is very inconsistent for other cues; we provide the justification cue as an example in Figure 4 and include all five cue types in Appendix Figure 8. Human-like sensitivity effects may be achievable, but they appear to be fragile, consistent with previous research on human simulation with LLMs (Hullman et al., 2026).

6 Discussion

Decision mechanisms in LLMs are distinct from those in humans. While context effects on choice and problem categorization are consistent with the theory of Bordalo et al. (2026b) and human behavior, feature sensitivities are not. Notably, LLMs do not exhibit human-like instability in price elasticities (Wakefield and Inman, 2003).

Our results contribute to a growing literature on the potential hazards of treating LLMs as human proxies (Gui and Toubia, 2025; Gao et al., 2025). In particular, researchers should be cautious when using comparative statics from LLM choice data to evaluate claims about humans. Agreement between human and LLM choices, when used alone, does not identify mechanistic correspondence.

That said, LLMs have made progress in accurately predicting human behavior, whether by default or with deliberate scaffolding, across social and cognitive science (Binz et al., 2025; Ashokkumar et al., 2026). An important avenue of research is understanding why LLMs seem to behave like humans in some settings, while sharply diverging from humans in others.

For instance, Bordalo et al. (2026b) argue that categories arise because humans selectively recall

memories from past experiences—including context, reasoning, and decisions—when encountering a new problem. We lack a rigorous understanding of whether LLMs encode analogous "experiences" about choice from training, and whether they use similar recall mechanisms in decision-making. To that end, training LLMs with targeted human data or structuring LLM representations to mimic those of humans could be promising steps towards faithful simulation (Binz et al., 2025).

Understanding LLM decision-making on its own terms. As LLMs become full-fledged economic agents, their decision representations, preferences, and mechanisms are becoming highly relevant objects of study. Our work is an example of using economic reasoning to understand LLM behavior (Imas et al., 2025; Cook et al., 2026; Buchanan and Foster, 2026). While LLMs may not behave fully like humans, we show that their decisions are consistent with utility maximization and problem categorization and that context predictably shifts their choices.

Our understanding of LLM behavior is much less developed than for humans, which opens several avenues for future work. First, the magnitude and direction of context effects, especially sensitivity to problem features, vary substantially across tested cues and models. We hope to study this heterogeneity in future work, including in more naturalistic settings and other types of decision problems.

Second, unlike human studies, we can directly measure and intervene on the internal representations of choice problems across contexts. Interpretability methods could complement behavioral studies by causally identifying circuits responsible for context sensitivity and bias, which may help disambiguate between competing theories and develop de-biasing interventions (Vig et al., 2020; Geiger et al., 2024; Karvonen and Marks, 2025).

Finally, our experiments are grounded in the bias literature. While existing human studies provide a search heuristic for LLM failures, biases and behaviors *unique* to LLMs remain under-explored (Jones and Steinhardt, 2022; Stella et al., 2023; McCoy et al., 2024; Cheung et al., 2025). One potential explanation for our results is implicit personalization: LLMs make decisions given user beliefs and preferences inferred from context, which may explain why the frontier models in our sample are the most sensitive to context (Xie et al., 2022; Marks et al., 2026).

Limitations

Behavioral similarity does not imply mechanistic correspondence. We evaluate LLM behavior against empirical human regularities and the behavioral predictions of a cognitive theory. While divergences between LLM and human behavior do imply distinct underlying mechanisms, similarities in behavior, such as context effects on choices or evidence on problem categorization in Section 5.3, do not imply shared mechanisms. Other theories may generate similar predictions about behavior; we do not disambiguate between competing explanations in this work. Mechanistic evidence from probing, steering, or causal interventions on internal representations is likely necessary to identify the processes that drive the behaviors we observe.

Experiments are imperfect analogies between humans and LLMs. Attention in humans has been measured through physical methods such as eye tracking or through self-reported scores (Reutskaja et al., 2011; Bordalo et al., 2026a). The human-like interpretation of token-level attention is tenuous, so we estimate LLM attention via regression as implied weights on problem features, a standard technique in economics (Bibal et al., 2022). In our experiments, higher attention means that an LLM behaves as if a feature is more important for choice. We leave the proper measurement and interpretation of LLM attention for future work. Likewise, our experiments in Section 5.3 treat in-context examples as loosely analogous to human experiences in memory. In-context examples are not the same as training data, and both may fundamentally differ from human memories.

A related limitation is that our predictions are taken from cognitive theory and existing human studies rather than a matched LLM-human experiment with identical questions. Although we use cues closely borrowed from the consumer choice literature, our cues may differ in systematic ways from those in prior work.

Limited scope of comparison. We test one canonical economic setting: binary choice between standard and premium consumer products. The theory of Bordalo et al. (2026b) applies to a much broader class of decision problems, including risk, intertemporal choice, and social preferences. We encourage future work in these domains, and with more LLMs and theories, to investigate the underlying mechanisms of LLM decision-making.

Supplemental Materials

This work's supplemental materials, including questions, model responses, and analysis code, are available at github.com/johnathansun/llm-cognitive-choice.

Acknowledgments

We thank Joshua Schwartzstein, Pedro Bordalo, Suproteem Sarkar, and Victoria Li for helpful comments and discussions. This work has been made possible in part by a gift from the Chan Zuckerberg Initiative Foundation to establish the Kempner Institute for the Study of Natural and Artificial Intelligence at Harvard University. In addition, it was supported by the Israel Science Foundation (grant No. 2942/25) and the European Union (ERC, Control-LM, 101165402). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

References

Jacy Reese Anthis, Ryan Liu, Sean M. Richardson, Austin C. Kozlowski, Bernard Koch, Erik Brynjolfsson, James Evans, and Michael S. Bernstein. 2025. Position: LLM social simulations are a promising research method. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 81005–81034. PMLR.

Ashwini Ashokkumar, Luke Hewitt, Isaias Ghezae, and Robb Willer. 2026. Large language models can predict the results of social science experiments. *Nature*, 656(8126):115–122.

Hemanth Asirvatham, Elliott Mokski, and Andrei Shleifer. 2026. GPT as a measurement tool. Working Paper 34834, National Bureau of Economic Research.

Michael S. Barr, Sendhil Mullainathan, and Eldar Shafir. 2009. The case for behaviorally informed regulation. In David Moss and John Cisternino, editors, *New Perspectives on Regulation*, chapter 2, pages 25–61. The Tobin Project, Cambridge, MA.

Gary S. Becker. 1962. Irrational behavior and economic theory. *Journal of Political Economy*, 70(1):1–13.

David R. Bell, Daniel Corsten, and George Knox. 2011. From point of purchase to path to purchase: How preshopping factors drive unplanned buying. *Journal of Marketing*, 75(1):31–45.

B. Douglas Bernheim and Dmitry Taubinsky. 2018. Behavioral public economics. In B. Douglas Bernheim, Stefano DellaVigna, and David Laibson, editors, *Handbook of Behavioral Economics: Applications and Foundations 1*, volume 1, pages 381–516. North-Holland.

Adrien Bibal, Rémi Cardon, David Alfter, Rodrigo Wilkens, Xiaoou Wang, Thomas François, and Patrick Watrin. 2022. Is attention explanation? An introduction to the debate. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3889–3900, Dublin, Ireland. Association for Computational Linguistics.

Pietro Bini, Lin William Cong, Xing Huang, and Lawrence J. Jin. 2026. Behavioral economics of AI: LLM biases and corrections. Working Paper 34745, National Bureau of Economic Research.

Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K. Eckstein, Noémi Éltető, Thomas L. Griffiths, Susanne Haridi, Akshay K. Jagadish, Li Ji-An, Alexander Kipnis, Sreejan Kumar, Tobias Ludwig, Marvin Mathony, Marcelo Mattar, and 21 others. 2025. A foundation model to predict and capture human cognition. *Nature*, 644(8078):1002–1009.

Pedro Bordalo, Giovanni Burro, Nicola Gennaioli, Gad Nacamulli, and Andrei Shleifer. 2025. Ads as cues. Working Paper 34387, National Bureau of Economic Research.

Pedro Bordalo, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. 2016. Stereotypes. *The Quarterly Journal of Economics*, 131(4):1753–1794.

Pedro Bordalo, John Conlon, Nicola Gennaioli, Spencer Kwon, and Andrei Shleifer. 2026a. How people use statistics. *The Review of Economic Studies*, 93(1):250–285.

Pedro Bordalo, Nicola Gennaioli, Giacomo Lanzani, and Andrei Shleifer. 2026b. A cognitive theory of reasoning and choice. *The Quarterly Journal of Economics*, 141(3):1921–1963.

Pedro Bordalo, Nicola Gennaioli, and Andrei Shleifer. 2012. Salience theory of choice under risk. *The Quarterly Journal of Economics*, 127(3):1243–1285.

Pedro Bordalo, Nicola Gennaioli, and Andrei Shleifer. 2013. Salience and consumer choice. *Journal of Political Economy*, 121(5):803–843.

Pedro Bordalo, Nicola Gennaioli, and Andrei Shleifer. 2020. Memory, attention, and choice. *The Quarterly Journal of Economics*, 135(3):1399–1442.

Joy Buchanan and Joshua Foster. 2026. The innate economic preferences of language models. *arXiv preprint*. ArXiv:2607.26288 [econ].

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Yiting Chen, Tracy Xiao Liu, You Shan, and Songfa Zhong. 2023. The emergence of economic rationality of GPT. *Proceedings of the National Academy of Sciences*, 120(51):e2316205120.

Vanessa Cheung, Maximilian Maier, and Falk Lieder. 2025. Large language models show amplified cognitive biases in moral decision-making. *Proceedings of the National Academy of Sciences*, 122(25):e2412015122.

Thomas R. Cook, Sophia Kazinnik, Zach Modig, and Nathan M. Palmer. 2026. What do LLMs want? Finance and Economics Discussion Series 2026-006, Board of Governors of the Federal Reserve System.

Mark Dean and Nathaniel Neligh. 2023. Experimental tests of rational inattention. *Journal of Political Economy*, 131(12):3415–3461.

DeepSeek-AI, Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chengyu Hou, Chenhao Xu, Chenze Shao, Chong Ruan, Conner Sun, and 300 others. 2026. DeepSeek-V4: Towards highly efficient million-token context intelligence. *arXiv preprint*. ArXiv:2606.19348 [cs].

Ravi Dhar and Klaus Wertenbroch. 2000. Consumer choice between hedonic and utilitarian goods. *Journal of Marketing Research*, 37(1):60–71.

Jessica Maria Echterhoff, Yao Liu, Abeer Alessa, Julian McAuley, and Zexue He. 2024. Cognitive bias in decision-making with LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12640–12653, Miami, Florida, USA. Association for Computational Linguistics.

Benjamin Enke and Thomas Graeber. 2023. Cognitive uncertainty. *The Quarterly Journal of Economics*, 138(4):2021–2067.

Xavier Gabaix. 2019. Behavioral inattention. In B. Douglas Bernheim, Stefano DellaVigna, and David Laibson, editors, *Handbook of Behavioral Economics: Applications and Foundations 1*, volume 2, pages 261–343. North-Holland.

Yuan Gao, Dokyun Lee, Gordon Burtch, and Sina Fazelpour. 2025. Take caution in using LLMs as human surrogates. *Proceedings of the National Academy of Sciences*, 122(24):e2501660122.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. 2024. Finding alignments between interpretable causal variables and distributed neural representations. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pages 160–187. PMLR.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 196 others. 2025. Gemma 3 technical report. *arXiv preprint*. ArXiv:2503.19786 [cs].

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

GLM-5 Team, Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, Chenzheng Zhu, Congfeng Yin, Cunxiang Wang, Gengzheng Pan, Hao Zeng, Haoke Zhang, Haoran Wang, Huilong Chen, and 167 others. 2026. GLM-5: From vibe coding to agentic engineering. *arXiv preprint*. ArXiv:2602.15763 [cs].

Google DeepMind. 2025. Gemini 3 Flash model card. Model card.

George Gui and Olivier Toubia. 2025. The challenge of using LLMs to simulate human behavior: A causal inference perspective. *arXiv preprint*. ArXiv:2312.15524 [cs].

Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2024. Bias runs deep: Implicit reasoning biases in persona-assigned LLMs. In *The Twelfth International Conference on Learning Representations*.

Thilo Hagendorff, Sarah Fabi, and Michal Kosinski. 2023. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. *Nature Computational Science*, 3(10):833–838.

Ryan Hamilton and Alexander Chernev. 2013. Low prices are just the beginning: Price image in retail management. *Journal of Marketing*, 77(6):1–20.

Chelsea Helion and Thomas Gilovich. 2014. Gift cards and mental accounting: Green-lighting hedonic spending. *Journal of Behavioral Decision Making*, 27(4):386–393.

Elizabeth C. Hirschman and Morris B. Holbrook. 1982. Hedonic consumption: Emerging concepts, methods and propositions. *Journal of Marketing*, 46(3):92–101.

Jessica Hullman, David Broska, Huaman Sun, and Aaron Shaw. 2026. This human study did not involve human subjects: Validating LLM simulations as behavioral evidence. *arXiv preprint*. ArXiv:2602.15785 [cs].

Alex Imas, Kevin Lee, and Sanjog Misra. 2025. Agentic interactions. SSRN Working Paper 5875162.

Itay Itzhak, Gabriel Stanovsky, Nir Rosenfeld, and Yonatan Belinkov. 2024. Instructed to bias: Instruction-tuned language models exhibit emergent cognitive bias. *Transactions of the Association for Computational Linguistics*, 12:771–785.

Zhengyang Jiang, Hongqi Liu, Cameron Peng, and Hongjun Yan. 2025. Investor memory and biased beliefs: Evidence from the field. *The Quarterly Journal of Economics*, 140(4):2749–2804.

Erik Jones and Jacob Steinhardt. 2022. Capturing failures of large language models via human cognitive biases. In *Advances in Neural Information Processing Systems*, volume 35, pages 11785–11799. Curran Associates, Inc.

Michael J. Kahana. 2012. *Foundations of Human Memory*. Oxford University Press, New York.

Adam Karvonen and Samuel Marks. 2025. Robustly improving LLM fairness in realistic settings via interpretability. *arXiv preprint*. ArXiv:2506.10922 [cs].

Uzma Khan and Ravi Dhar. 2006. Licensing effect in consumer choice. *Journal of Marketing Research*, 43(2):259–266.

Anthony Koschmann and Mathew S. Isaac. 2018. Retailer categorization: How store-format price image influences expected prices and consumer choices. *Journal of Retailing*, 94(4):364–379.

Jonathan Levav and A. Peter McGraw. 2009. Emotional accounting: How feelings about money influence consumer choice. *Journal of Marketing Research*, 46(1):66–80.

Peiyao Li, Noah Castelo, Zsolt Katona, and Miklos Sarvary. 2024. Frontiers: Determining the validity of large language models for automated perceptual analysis. *Marketing Science*, 43(2):254–266.

Nira Liberman and Yaacov Trope. 1998. The role of feasibility and desirability considerations in near and distant future decisions: A test of temporal construal theory. *Journal of Personality and Social Psychology*, 75(1):5–18.

Shi Xian Liew, Piers D. L. Howe, and Daniel R. Little. 2016. The appropriacy of averaging in the study of context effects. *Psychonomic Bulletin & Review*, 23(5):1639–1646.

Gili Lior, Liron Naccache, and Gabriel Stanovsky. 2026. Comparing the framing effect in humans and LLMs on naturally occurring texts. *arXiv preprint*. ArXiv:2502.17091 [cs].

Simon Malberg, Roman Poletukhin, Carolin M. Schuster, and Georg Groh. 2025. A comprehensive evaluation of cognitive biases in LLMs. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 578–613, Albuquerque, USA. Association for Computational Linguistics.

Benjamin S. Manning and John J. Horton. 2026. General social agents. Working Paper 34937, National Bureau of Economic Research.

Charles F. Manski. 1995. *Identification Problems in the Social Sciences*. Harvard University Press, Cambridge, MA.

Sam Marks, Jack Lindsey, and Christopher Olah. 2026. The persona selection model: Why AI assistants might behave like humans. Anthropic Alignment Science Blog.

Bartosz Maćkowiak, Filip Matějka, and Mirko Wiederholt. 2023. Rational inattention: A review. *Journal of Economic Literature*, 61(1):226–273.

R. Thomas McCoy, Shunyu Yao, Dan Friedman, Mathew D. Hardy, and Thomas L. Griffiths. 2024. Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, 121(41):e2322420121.

Erica Mina Okada. 2005. Justification effects on consumer choice of hedonic and utilitarian goods. *Journal of Marketing Research*, 42(1):43–53.

OpenAI. 2026. GPT-5.4 Thinking system card. System card.

Joon Sung Park, Carolyn Q. Zou, Jonne Kamphorst, Niles Egan, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Percy Liang, Robb Willer, and Michael S. Bernstein. 2026. LLM agents grounded in self-reports enable general-purpose simulation of individuals. *arXiv preprint*. ArXiv:2411.10109 [cs].

Siddhesh Milind Pawar, Arnav Arora, Lucie-Aimée Kaffee, and Isabelle Augenstein. 2025. Presumed cultural identity: How names shape LLM responses. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22147–22172, Suzhou, China. Association for Computational Linguistics.

Charles R. Plott and Kathryn Zeiler. 2005. The willingness to pay–willingness to accept gap, the "endowment effect," subject misconceptions, and experimental procedures for eliciting valuations. *American Economic Review*, 95(3):530–545.

Elena Reutskaja, Rosemarie Nagel, Colin F. Camerer, and Antonio Rangel. 2011. Search dynamics in consumer choice under time pressure: An eye-tracking study. *American Economic Review*, 101(2):900–926.

Sherwin Rosen. 1974. Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy*, 82(1):34–55.

Jillian Ross, Yoon Kim, and Andrew W. Lo. 2024. LLM economicus? Mapping the behavioral biases of LLMs via utility theory. In *First Conference on Language Modeling*.

Katharina Schumacher, Roger Roberts, and Katharina Giebel. 2025. The agentic commerce opportunity: How AI agents are ushering in a new era for consumers and merchants. Report, McKinsey & Company.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*.

Cassidy Shubatt and Jeffrey Yang. 2026. Tradeoffs and comparison complexity. *arXiv preprint*. ArXiv:2401.17578 [econ].

Keith E. Stanovich and Richard F. West. 2000. Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, 23(5):645–665.

Massimo Stella, Thomas T. Hills, and Yoed N. Kenett. 2023. Using cognitive psychology to understand GPT-like models needs to extend beyond human biases. *Proceedings of the National Academy of Sciences*, 120(43):e2312911120.

Stripe. 2025. Agentic Commerce Protocol. Web page. Open standard codeveloped by Stripe and OpenAI, announced 29 September 2025.

Johnathan Sun and Andrew Zhang. 2026. Persona vectors in games: Measuring and steering strategies via activation vectors. *arXiv preprint*. ArXiv:2603.21398 [cs].

Richard H. Thaler. 1999. Mental accounting matters. *Journal of Behavioral Decision Making*, 12(3):183–206.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems 33*, pages 12388–12401.

Kirk L. Wakefield and J. Jeffrey Inman. 2003. Situational price sensitivity: the role of consumption occasion, social context and income. *Journal of Retailing*, 79(4):199–212.

Rui Wang, Qihan Lin, Jiayu Liu, Qing Zong, Tianshi Zheng, Dadi Guo, Haochen Shi, Peixuan Han, Weiqi Wang, and Yangqiu Song. 2026. Rethinking prospect theory for LLMs: Revealing the instability of decision-making under epistemic uncertainty. *arXiv preprint*. ArXiv:2508.08992 [cs].

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2022. An explanation of in-context learning as implicit Bayesian inference. In *The Tenth International Conference on Learning Representations*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *arXiv preprint*. ArXiv:2505.09388 [cs].

A Experimental Details

A.1 Product pairs

Each product's five options form a labeled quality ladder (budget, economy, midrange, premium, luxury), with attributes drawn from a fixed per-product schema (e.g., blade material, handle, and edge for chef's knives) so that adjacent tiers differ in attribute values rather than description style. Binary pairs use adjacent tiers (1–2 through 4–5). Prices are taken as generated, and realized premium/standard price ratios span 1.4–3.5× (median 2.2×). Table 2 shows two example pairs.

A.2 Cue texts

Each of the five cue dimensions (Table 1) has one paired hedonic/functional cue per product, so a framed trial differs from its counterpart only in the cue sentence. The payment source, justification, and reasoning frame cues use a fixed template with a product-specific stem, while the purchase trigger and retail architecture cues are instantiated per product by Claude Sonnet 4.6 so that they read naturally for the specific product (e.g., staff notes on "blade feel" for the chef's knife). Cues never mention price, quality, or either option's attributes. Table 3 shows all five dimensions instantiated for one product.

A.3 Prompt template and trial structure

Cues are written from the shopper's perspective. The full template is

```
{cue_text} Here are two options.

**Option A: {a_name}**
{a_description}
```

```
Price: ${a_price}

**Option B: {b_name}**
{b_description}
Price: ${b_price}

Which option should I choose? Answer
with only "A" or "B".
```

Each of the 800 (product, tier pair, order) combinations appears under an 11-cell grid — 5 cue dimensions $\times$ 2 poles (functional/hedonic), plus one neutral cell — which the sweeps then cross with price multipliers and description variants.

A.4 Models and inference

Local runs apply each model card's recommended chat template. We measure the first-token log-probability of the "A"/"B" answer tokens. GPT 5.4 and Gemini 3.5 Flash expose no token logits, so they enter only choice rate analyses.

A.5 LLM-judge scores

Quality. We use GPT 5.4 as a numeric judge to produce quality scores. Each option is rated five times on a $[0, 100]$ scale and the ratings are averaged. The judge is instructed to rate the product description, and prices are not provided. Table 5 (Panel A) confirms that generation produces a smoothly increasing quality ladder; Panel B shows the mean rating across the five quality-sweep perturbations ($y_{\text{index}} \in \{-2, \ldots, +2\}$), each of which re-generates the option description with different attributes while holding price fixed.

Hedonic-ness. The product-level hedonic-ness scores of Section 5.3 also come from a GPT 5.4-based judge (Asirvatham et al., 2026). Scores span the full range but are right-skewed (median 33) (Figure 7).

B Within-Model Standardization

Models differ by an order of magnitude in the variance of $\ell_i = \text{logit}\, P(\text{premium})$ on neutral trials. For instance, Gemma 27B's neutral standard deviation (SD) is roughly 5$\times$ Qwen 4B's. To compare shapes of cue effects across models, we rescale every model's ℓ_i by its own neutral-trial SD,

$$\tilde{\ell}_{i,m} = \frac{\ell_{i,m}}{\widehat{\sigma}_m^{(0)}},\ \widehat{\sigma}_m^{(0)} = \text{SD}\big(\{\ell_i : i \in \text{neutral}_m\}\big)$$

so $\tilde{\ell}$ is in neutral-SD units within model m. All "Δ std. logit" axes in the paper (Figure 2 right, Figure 3 right, Figure 4 right) use this standardization, and pooled lines average $\tilde{\ell}$ across models with equal weight.

C Reproducibility Statement

Compute. Local inference runs on NVIDIA H100 and H200 GPUs. The benchmark requires approximately 1000 H100-hours for the Gemma 3 and Qwen 3 families. API costs for GPT 5.4, Gemini 3.5 Flash, GLM 5.2, DeepSeek v4 Pro, and Qwen 3.7 Plus totaled roughly $500.

Software. Local inference uses `torch` 2.9.1, `transformers` 4.57.1, and `sglang` 0.5.9; analysis uses `numpy`, `pandas`, `statsmodels` 0.14, `scipy`, `matplotlib`, and `seaborn`. The full pinned environment is released with the code.

Licenses and intended use. Gemma 3 weights are used under the Gemma Terms of Use and Qwen 3 weights under Apache 2.0; GLM 5.2, DeepSeek v4 Pro, and Qwen 3.7 Plus are accessed through the Fireworks API, and GPT 5.4 and Gemini 3.5 Flash through their providers' APIs, each under the respective terms of service. All use is non-commercial research. The questions and model responses will be released under CC BY 4.0. The dataset is intended for research on LLM behavior, not for training of commercial systems.

AI assistance. Claude Sonnet 4.6 generated the product descriptions, prices, and per-product cue instantiations (Sections A.1 and A.2); the generation prompts and parameters are released with the dataset. GPT 5.4 served as a numeric judge for quality and hedonic-ness scores (Section A.5). The authors used Claude and ChatGPT for coding assistance and for editing portions of the manuscript. All scientific decisions were made by the authors.

Risks and privacy. One risk is that the cue effects we document could be read as a recipe for steering LLM recommendations by adding framing effects that are invisible to users. That said, cues are borrowed from existing literature, and releasing the dataset adds little novel capability: all stimuli are synthetic, brand-free, and contain no demographic, geographic, or identity-related framing, and the manipulations are drawn from a longstanding consumer choice literature. We manually audited all 100 products and a random sample of cue texts; no PII or offensive material is present, and model responses contain no free-form text beyond A/B choices and log-probabilities.

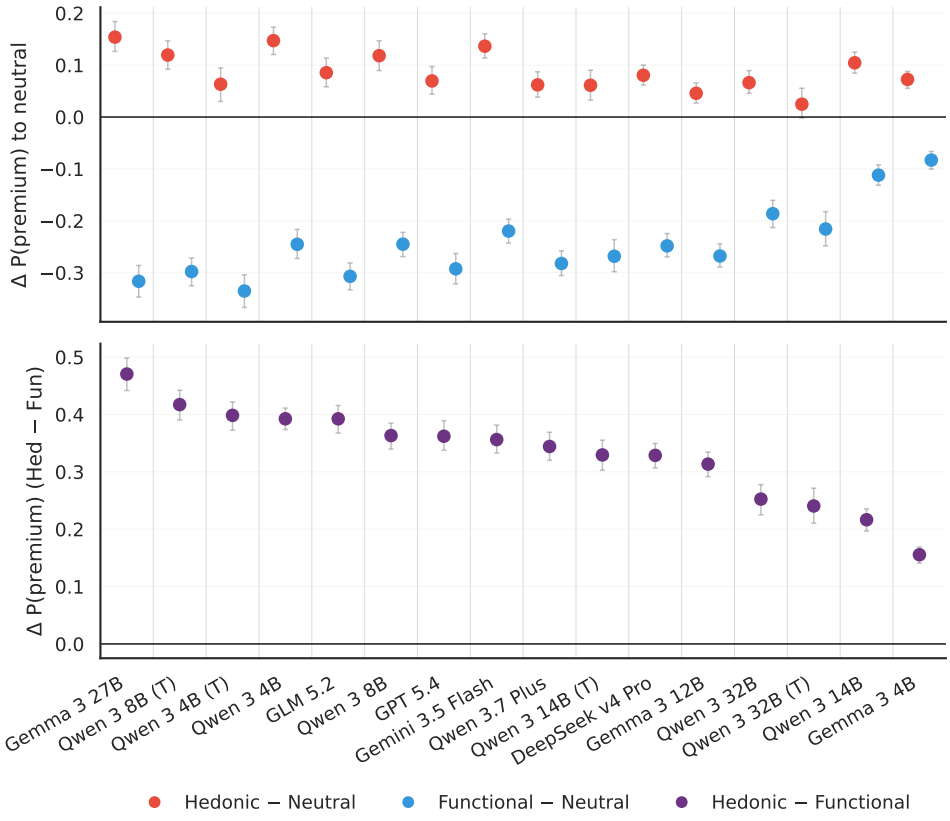


Figure 5: **Premium choice rates by model and cue condition on the baseline dataset**, for all sixteen configurations (twelve models plus the four thinking modes; (T) denotes thinking mode). (**Top**) Difference in premium choice rate between hedonic (functional) cues and the neutral condition. (**Bottom**) Difference in premium choice rate between hedonic and functional cues, averaged across cue types. All bars are 95% CIs.

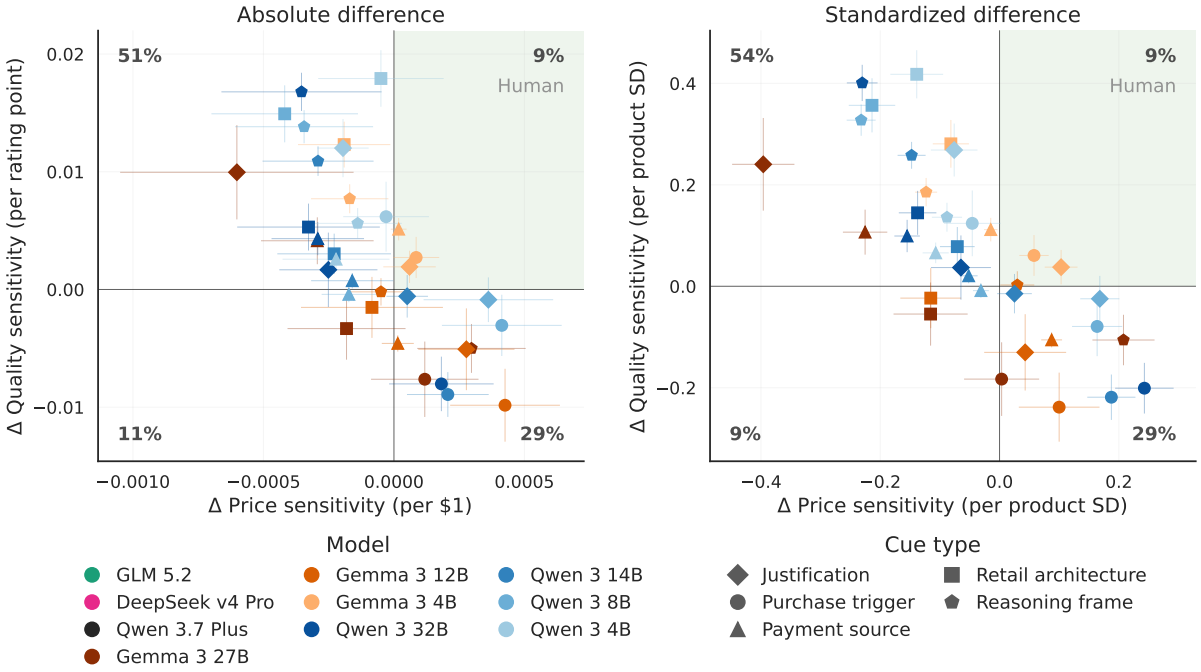


Figure 6: **Cue effects on feature sensitivity under alternate regression specifications.** The same quantity as Figure 2 (right), re-estimated with differences in feature values instead of log ratios: absolute differences (left; price sensitivity per dollar, quality sensitivity per rating point) and per-product-SD standardized differences (right). Percentages report the share of model-cue combinations in each quadrant.

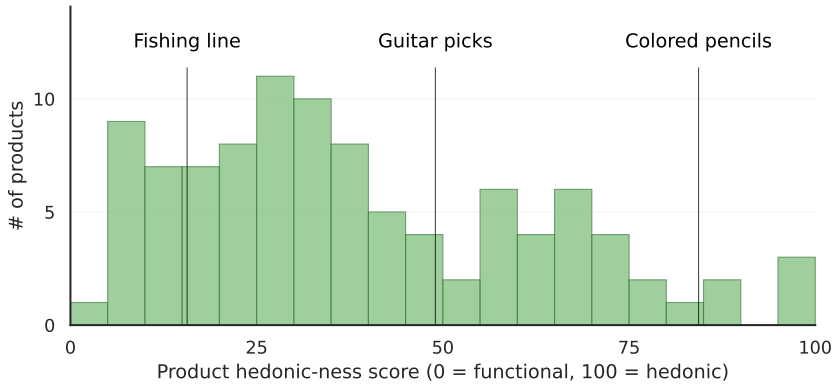


Figure 7: **Distribution of hedonic-ness scores across the 100 products.** Scores from the GPT 5.4 judge (Section A.5); 0 = purely functional, 100 = purely hedonic. The distribution is right-skewed (median 33), with labeled products marking representative points.

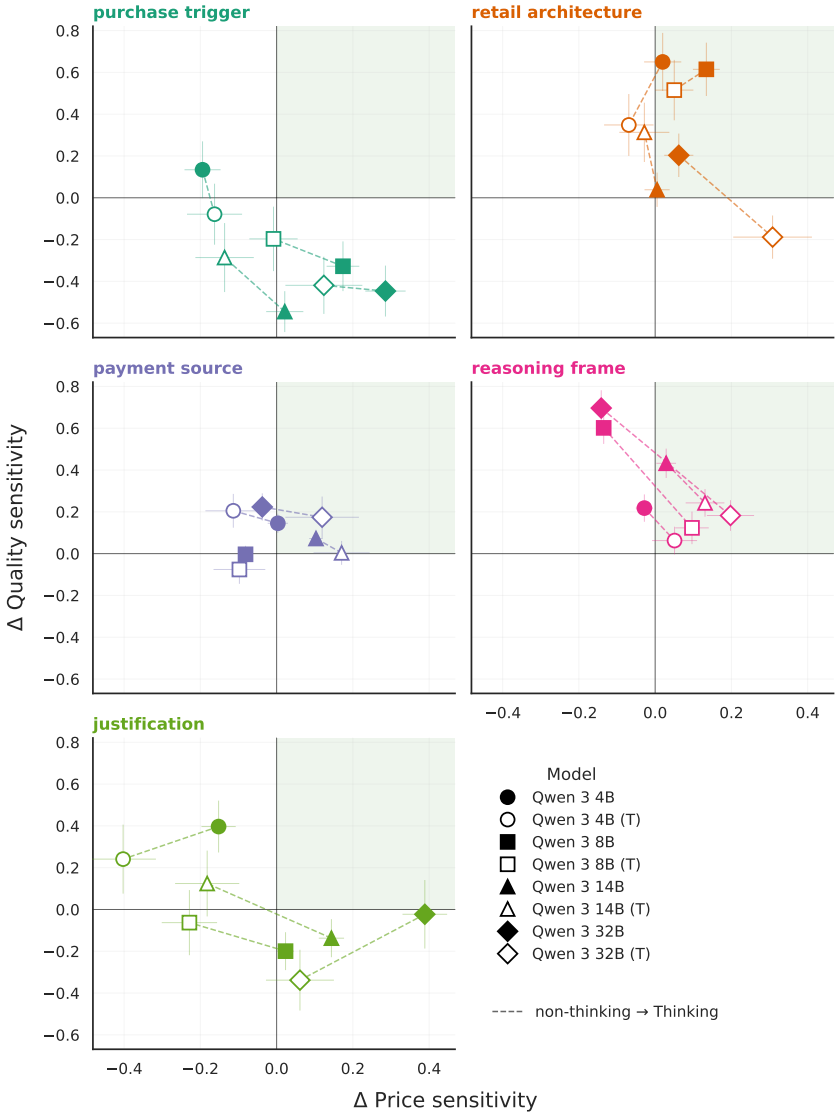


Figure 8: **Effect of thinking mode on price and quality sensitivity, by cue type.** Difference in standardized price (x-axis) and quality (y-axis) sensitivity between hedonic and functional cues for Qwen models in thinking/non-thinking modes. Shaded region denotes human-consistent behavior. Figure 4 (right) shows the justification and reasoning frame dimensions. All bars are 95% CIs.

Fishing reel	*outdoor & sporting goods* | price ratio 2.33	
Premium	**Aluminum spinning reel with sealed drag.** Aluminum alloy body and rotor, 6 ball bearings, machined aluminum spool with sealed multi-disc drag, suited for light saltwater fishing.	$69.99
Standard	**Graphite spinning reel, entry level.** Graphite composite body and rotor, 3 ball bearings, aluminum spool with basic anti-reverse mechanism, suited for freshwater fishing.	$29.99
Electric toothbrush	*personal care & grooming* | price ratio 3.13	
Premium	**Rechargeable oscillating toothbrush.** Single oscillating speed, USB charging cable, plastic handle, one replacement brush head and two-minute timer.	$24.99
Standard	**Basic battery-powered toothbrush.** Single-speed vibrating motor, AA-battery powered, fixed plastic handle, one nylon bristle brush head included.	$7.99

Table 2: **Two example product pairs.** Each row shows the option name (bolded) and description generated by Claude Sonnet 4.6, along with its price. Descriptions are length-matched and brand-free, and option order (A=premium vs. A=standard) is counterbalanced across trials.

Dimension	Pole	Cue
Purchase trigger	*Hed.*	I'm shopping for a chef's knife because I've been curious about the category and I'm considering an upgrade to my everyday cooking setup.
	Func.	I'm shopping for a chef's knife because my current one has worn down and I want a solid replacement without overpaying.
Retail architecture	*Hed.*	I'm browsing for a chef's knife on a curated kitchenware shop page where staff picks include quality notes on blade feel and durability.
	Func.	I'm browsing for a chef's knife on a marketplace comparison page sorted by price, with filters, unit costs, and price-match language displayed.
Payment source	*Hed.*	I'm looking for a chef's knife and will be paying with a gift card I received as a birthday present. Any unspent amount remains mine.
	Func.	I'm looking for a chef's knife and will be paying with money from my regular checking account. Any unspent amount remains mine.
Justification	*Hed.*	I'm looking for a chef's knife. I recently completed a demanding project and earned a small reward, so this feels like a reasonable treat.
	Func.	I'm looking for a chef's knife. This is an ordinary purchase with no special reason to treat myself—I'm just making a practical choice.
Reasoning frame	*Hed.*	I'm looking for a chef's knife. I'm focused on why this product would improve my use experience and the qualities I'd notice while using it.
	Func.	I'm looking for a chef's knife. I'm focused on how I'll complete the purchase and the concrete tradeoffs to compare.

Table 3: **Cue templates instantiated for one product** (chef's knife). Each of the five dimensions has a paired hedonic/functional cue. Cues for the remaining 99 products are generated analogously via Claude Sonnet 4.6. Appendix A.3 provides the full prompt template for choice questions.

Family	Trials	H.	F.	N.
Baseline	8,800	4,000	4,000	800
Price sweep	88,000	40,000	40,000	8,000
Quality sweep	44,000	20,000	20,000	4,000
ICL mix	36,000	0	0	36,000
Total	176,800			

Table 4: **Trial counts per model.** "H."/"F."/"N." are hedonic, functional, and neutral conditions. The same 800 option pairs are reused across all four experiment families, but not every model is evaluated in each experiment.

Panel A: baseline tier ratings ($y_{\text{index}} = 0$)

Tier	Label	Mean	SD	n
1	budget	30.4	9.2	100
2	economy	54.6	11.8	100
3	midrange	72.9	8.7	100
4	premium	84.8	4.3	100
5	luxury	89.8	3.7	100

Panel B: quality-sweep mean rating

Tier	-2	-1	0	$+1$	$+2$
1 (budget)	19.4	28.2	30.4	41.4	53.1
2 (economy)	31.7	46.6	54.6	61.7	67.5
3 (midrange)	49.3	65.9	72.9	76.6	81.0
4 (premium)	67.5	80.3	84.8	86.5	88.6
5 (luxury)	81.7	87.5	89.8	90.8	92.4

Table 5: **GPT 5.4 quality ratings.** Each option is rated five times independently from 0 to 100 with a rubric that ignores price and brand and scores only descriptive attributes (Section A.5). Panel A: mean rating per baseline tier across the 100 products ($y_{\text{index}} = 0$). Panel B: mean rating across the five quality-sweep perturbations.

Model	Access	Backend	Decoding	Choice DV
Gemma 3 4B/12B/27B-IT	Open	Local	Greedy	First-token A/B logprob
Qwen 3 4B/8B/14B/32B	Open	Local	Greedy	First-token A/B logprob
Qwen 3 4B–32B (Thinking)	Open	Local	T=0.6, top-p 0.95	A/B logprob after CoT
GLM 5.2	API	Fireworks	T=0	A/B logprob (quantized)
DeepSeek v4 Pro	API	Fireworks	T=0	A/B logprob (quantized)
Qwen 3.7 Plus	API	Fireworks	T=0	A/B logprob (quantized)
Gemini 3.5 Flash	API	Google AI Studio	Medium reasoning	Choice only
GPT 5.4	API	OpenAI Batch	Medium reasoning	Choice only

Table 6: **Models evaluated.**