AUDITING THE ECONOMIC ALIGNMENT OF LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Language models are increasingly used to make purchases and to negotiate on people's behalf. An agent trusted with such decisions should at least hold a coherent preference: whether it accepts a routine trade at a given price should be well defined, should not change when the same offer is reworded, and should not depend on details that are irrelevant to the trade, such as the names of the two parties. This paper tests that requirement directly. Eight open-weight instruction-tuned models were each asked, one price at a time, whether to buy a fixed quantity of water or of wheat, over dense price grids and under controlled variations of the prompt, yielding about 1.1 billion single-token yes/no decisions. Each price sweep is first tested for a clear switch from buying to declining inside the tested prices; where one exists, the switching price is located with a logistic fit.

The requirement is met only partially. Where a switching price exists, rewording the same offer moves it by more than a factor of five, and sampling the output at temperature one instead of decoding greedily makes six of the eight models flip between buying and declining at 20-50% of adjacent prices. Replacing the generic labels "Buyer" and "Seller" with names changes the share of prices at which a model buys by up to 95 percentage points for the typical name pair, and which names are used moves the switching price across a 2.5-fold range in the median model-commodity panel; most of this variation is attributable to individual names rather than to the region or gender labels the names carry. A coherence leaderboard summarizes the eight models on responsiveness to price and invariance to irrelevant cues; on it, an agent that never changes its decision outscores half the models. An economic delegate must therefore be evaluated on its whole decision curve and its invariances, not on a single reservation price. The 1.1 billion model responses and the analysis code will be released after the review period.

1 INTRODUCTION

Alignment is usually discussed in terms of values and intent. For an agent that transacts on a user's behalf, a narrower requirement comes first: its decisions must follow from the user's interest in the trade rather than from the wording of the offer, the sampling seed, or who happens to be the counterpart. Economics supplies the least demanding standard for this requirement. A buyer with a consistent valuation, offered the same good at rising prices, buys below a reservation price and declines above it (Becker et al., 1964). The step requires no optimization and nothing approaching full rationality, only a preference for paying less rather than more for an identical good. It is a floor on coherent behavior, and departures from it are informative in the way that human anchoring, framing, and charm pricing are (Tversky & Kahneman, 1974; 1981; Thomas & Morwitz, 2005).

This paper elicits the step directly, at scale, from eight open-weight instruction-tuned models: about 1.1 billion single-token yes/no purchase decisions over dense price grids for two goods, water and wheat. Each price sweep is read in two stages. The first asks whether the curve contains a decision boundary inside the tested prices at all; the second, when it does, locates the boundary and its width with a logistic fit and records how "turbulent" the surrounding curve is. Four properties of an aligned economic delegate are then tested against the curves. The boundary should exist and be stable under equivalent wording and under decoding noise; it should move in the expected direction when the quantity of the good changes; and it should not move when the trade is unchanged but the two parties are given names or professions.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

The results support four conclusions.

1. *Coherent purchasing is rare and does not follow model size.* Only qwen25-32b and gemma3-27b hold a clean step on water that survives rewording and sampling, 34 of 64 baseline curves contain no boundary, and a 1.5B model outranks the 8B Llama (Sections 5.1, 4.4).
2. *Models get the economics right but not robustly.* Every boundary rises with quantity, yet it moves more than fivefold under rewording and disappears for six of eight models once outputs are sampled (Sections 5.2, 5.3).
3. *The counterpart's name is the largest irrelevant factor, and it acts name by name.* Names move the median pair's demand by up to 95 points and the boundary by a median 2.5-fold across pairs, and only a fifth of that follows the region the names came from, so a group-level audit would miss most of it (Sections 5.4, 5.5).
4. *A single reservation price is the wrong evaluation.* An agent that never changes its decision outscores half the models; the whole curve, under the deployment's decoding, is the unit of evaluation (Sections 5.6, 6).

The names are used as invariance stimuli, not as proxies for groups. The trade is fully specified in the prompt, the good, its quantity, the price, and the buyer's need, so the counterpart's name carries no information about the value of the trade to the user; in an open-ended market a name might signal trust or reliability, but here nothing about delivery or quality varies with it. An aligned delegate should therefore produce the same curve for every counterpart, the variance across name pairs is the measurement, and sixty names suffice for it. The paper accordingly reports name-level effects in the main text, confines region- and gender-level summaries with their permutation nulls to the appendix, and makes no claim about the populations the names were drawn to represent. Three questions organize the results.

1. *Is the decision boundary well defined and stable?* Each curve is classified as constant, near-constant, or containing a boundary inside the tested range (Section 3.2); for the last group the boundary and its width are reported, and for every curve the share of prices at which the model buys (*average demand*), the adjacent-price flip rate (*turbulence*), and the share of the grid on which neither decision dominates (*turbulent fraction*), all defined in Section 3. Stability is measured across four phrasings and, in an ablation, across sampling temperatures.
2. *Does the boundary respond to what matters and ignore what does not?* Quantity is the economically relevant change and serves as a positive control. The irrelevant changes are the presence of names, the identity of the named or professionally described parties in a full buyer$\times$seller cross, and the phrasing. Throughout, a *panel* is one model on one good, sixteen in all.
3. *Where does instability live?* Turbulence is measured per curve and per name pair, related to the demand level, and tested for dependence on the counterpart's name.

The contributions are (i) a two-stage price-sweep analysis that tests for a step inside the tested range and then reports a logistic boundary with its width, average demand, and two instability metrics (Section 3.4); (ii) a coherence leaderboard scoring responsiveness to price and invariance to irrelevant cues (Section 4.4); (iii) name-level permutation tests that separate name-specific variation from region, gender, and pairing structure; and (iv) the 1.10 billion decisions and the analysis code, released after the review period.

2 Related work

Alignment of language agents and delegated decisions Work on aligning language agents distinguishes the intent a model should serve from the behaviors through which it serves it, and analyzes accidental misspecification as a source of misalignment (Kenton et al., 2021; Gabriel, 2020; Gabriel et al., 2024); scenario-based evaluations probe misalignment in high-stakes agentic settings (Lynch et al., 2025). This paper adopts the narrower view that, for a delegated economic decision, reliability of the decision under variation of irrelevant input is itself part of alignment. Language models are now evaluated as negotiators (Davidson et al., 2024; Bianchi et al., 2024) and audited as purchasing

agents, where position bias, preferences that differ across models, and inconsistent price sensitivity have been reported (Allouah et al., 2026); the present study takes the single decision underlying all of these, whether to accept a trade at a price, and asks whether it is coherent before asking whether it is skillful.

Coherence of LLM preferences and economic rationality Revealed-preference analysis reconstructs preferences from choices (Samuelson, 1938; Afriat, 1967; Varian, 1982). Applied to language models, Horton (2023) recreates classic behavioral experiments with LLM subjects, Chen et al. (2023) score budget-allocation choices for consistency with utility maximization, benchmarks test elements of microeconomic rationality and behavioral bias (Raman et al., 2024; 2025; Ross et al., 2024), Mazeika et al. (2025) fit utility models to pairwise choices and report coherence that grows with scale, and Liu et al. (2025) test transitivity and negation invariance of preferences, of which monotonicity in price is the one-dimensional case. Willingness-to-pay elicitation from language models has been compared with human data (Brand et al., 2023), and price variation in prompts has been shown to be confounded with the prompt itself unless the design controls for it (Gui & Toubia, 2023). This study characterizes the raw shape of the elicited decision curve, in the spirit of classic reservation-price procedures (Becker et al., 1964) and of machine psychology, which runs behavioral experiments on models rather than scoring them on a task (Binz & Schulz, 2023; Hagendorff et al., 2024; Coda-Forno et al., 2024); the vocabulary of *multiple switching* for repeated reversals in a price list is borrowed from the human literature (Jack et al., 2022; Yu et al., 2021), and the question of whether price-specific reversals, of the kind charm pricing produces in humans (Thomas & Morwitz, 2005), recur across models with the same number tokenization (Singh & Strouse, 2024) is examined in Appendix J.

Sensitivity to wording and decoding Language-model outputs vary with semantically irrelevant features of the prompt by amounts large enough to reverse model rankings (Sclar et al., 2024) and with the order of in-context examples (Lu et al., 2022), consistency under paraphrase has been proposed as a property to measure in its own right (Elazar et al., 2021; Moore et al., 2024), and, by analogy with the option-token priors documented for multiple-choice answers (Zheng et al., 2024), yes/no elicitation is exposed to answer priors independent of content. Sampling temperatures between 0 and 1 have been reported to leave accuracy on multiple-choice problems statistically unchanged (Renze, 2024), yet single-sample greedy evaluation hides the non-determinism a deployed agent would show (Song et al., 2025). The four phrasings of Experiment 1 and the temperature ablation quantify both sources of variation for the purchase decision, and every identity effect is reported against the phrasing noise as a benchmark.

Names as an invariance probe Correspondence audits hold a task fixed and change only a name; the design was popularized by hiring studies that measured call-back rates for otherwise identical resumes (Bertrand & Mullainathan, 2004) and has been transferred to language-model decisions on advice, valuations, hiring, salary, and other consequential choices (Salinas et al., 2024; An et al., 2024; Nghiem et al., 2024; Tamkin et al., 2023). Names are heterogeneous and imperfect signals of any demographic category; this paper uses them for a weaker purpose, since a delegate whose decision changes with the counterpart's name has failed an invariance that alignment requires whatever the name signals, and assigns identities to both parties so that pairing effects can be separated from each party's own effect.

3 Measuring Economic Coherence with Price Sweeps

3.1 Decision task and the ideal demand curve

A price sweep holds a proposed trade fixed while varying only its price. For a buyer with reservation value v, the economically coherent response is the step

$$d^*(p; v) = \mathbf{1}\{p \leq v\},$$

so the buyer chooses to buy every lower-priced instance of the trade and declines every higher-priced one. This is a deliberately weak condition: it requires only that paying more for an identical trade cannot lead the model to buy after declining at a lower price, and that economically irrelevant changes to the prompt leave v unchanged.

Every prompt in the corpus follows the same shape. One example (a low-value everyday commodity, the base phrasing, and generic buyer and seller roles) is:

```
Buyer is really thirsty.  Seller is offering a 500ml bottle
of drinking water for $0.10.  Should Buyer buy the water?
Output only Yes or No.
```

Within a sweep only the price changes; the price-indexed decisions form a *decision curve*. Within experiments, the persona strings may be names or professions, and the quantity and phrasing may vary (Section 4).

Each prompt receives a single-token greedy completion ($T = 0$), classified as buy, decline, or other; the last category, reported as refusals below, covers every response that is neither buy nor decline. Sorting these decisions by price gives the decision curve; a centered 50-position rolling mean is used only for visualization and smoothed summaries. The exact classifier, missing-value treatment, and smoother appear in Appendix A.

3.2 Decision patterns across price ranges

The first stage asks whether the curve contains a decision boundary inside the tested prices. A curve is *constant* when every decision is the same. Otherwise, the best hard step $\mathbf{1}\{p \le t\}$ is fitted to the raw decisions, choosing t to minimize the number of disagreements, and its *gain* is computed: the share of the minority decisions that the step explains, relative to the constant curve that ignores them. A curve is *near-constant buy* or *near-constant decline* when the gain is below one half, meaning that its minority decisions are scattered rather than concentrated on one side of a price. The remaining curves contain a step inside the range, and Section 3.3 locates it. These descriptions are explicitly range-limited: they do not establish how a model would decide outside the sampled prices.

3.3 The decision boundary

For a curve with a step inside the range, the boundary is located by maximum likelihood with a logistic curve in log price,

$$\mathrm{P}(\text{buy} \mid p) = \sigma\left(-\frac{\log p - \log \hat{t}}{\hat{s}}\right),$$

where $\hat{t}$ is the price at which the fitted buy probability crosses one half and $\hat{s}$ is the width of the transition in log price: the fitted probability falls from 0.73 to 0.27 between $\hat{t}e^{-\hat{s}}$ and $\hat{t}e^{+\hat{s}}$. $\hat{t}$ is called the *decision boundary* and is reported only when it lies inside the tested price range; a curve whose fitted crossing falls above or below the range is labeled accordingly and receives no boundary. On a clean downward step, $\hat{t}$ equals the reservation price and $\hat{s}$ is near zero. On an oscillating curve, $\hat{t}$ marks the center of the turbulent zone and $\hat{s}$ its extent, so it is read together with turbulence and turbulent fraction. The best-step threshold of Section 3.2 is retained as a check and agrees with $\hat{t}$ within 10% on three quarters of the curves. Appendix H compares the logistic center with a buy-price quantile.

3.4 Turbulence and other curve summaries

The decision boundary alone cannot distinguish a near-ideal step from a curve that repeatedly reverses direction; $\hat{s}$ measures the width of the transition, and both are paired with *turbulence*, the adjacent-price buy/decline flip rate (the analogue of multiple switching in a price list), and *turbulent fraction*, the share of the smoothed curve with $0.3 < m < 0.7$. Appendix A gives the exact estimators.

4 Experimental Design

4.1 Experiments

The study comprises four experiments; each varies one element of the prompt while holding the rest fixed, and asks whether the decision curve moves with it. Experiment 1 uses four semantically equivalent phrasings (Section 5.2); Experiment 2 varies each commodity across five quantities

(Section 5.3); Experiment 3 crosses 18 buyer professions with 18 seller professions (Section 5.5); and Experiment 4 crosses 60 buyer names with 60 seller names (Section 5.5). Comparing the base phrasing of Experiment 1, whose parties are the generic roles "Buyer" and "Seller", with the named pairs of Experiment 4 gives the naming test of Section 5.4. Experiments 1-2 also provide the curves used in Appendix J to test whether price extrema recur across models. One decision curve is the full price sweep for a fixed model, commodity, condition, buyer, and seller.

4.2 Models and inference

Eight instruction-tuned open-weight models are evaluated (Qwen Team, 2024; DeepSeek-AI, 2024b; Grattafiori et al., 2024; Meta AI, 2024a; Gemma Team, 2025). The selection spans two developer-origin regions and two price-tokenization regimes: the Qwen, DeepSeek, and Gemma tokenizers encode each digit of a price separately, as verified on the released tokenizers and documented in their reports (Bai et al., 2023; DeepSeek-AI, 2024a; Gemma Team, 2025), whereas the Llama 3 tokenizer groups digits into one- to three-digit tokens, a property of the pre-tokenizer pattern in its released code (Meta AI, 2024b), unlike earlier Llama releases (Singh & Strouse, 2024). Whether shared price-specific irregularities follow either grouping is examined in Appendix J.

In the text and the per-panel tables the models are referred to by the short names qwen25-0.5b, qwen25-1.5b, qwen25-32b, deepseek-v2-lite, gemma3-27b, llama32-1b, llama32-3b, and llama31-8b; Table 4 and the leaderboard tables use the release names. Every model receives the same materialized prompts under the decoding and classification procedure of Section 3.1; Appendix A lists the models by tokenization regime.

4.3 Stimuli, identities, and corpus scope

The retained commodities are drinking water and milling wheat, each tested over quantity-scaled price ranges from cents to hundreds of thousands of dollars. The piecewise grid uses one-cent steps below $1,000 and ten-dollar steps above it; displayed prices change from `$X.XX` to `$X.XXk` at the same boundary. The profession set contains three occupations from each of the six skill tiers (TEER categories) of Canada's National Occupational Classification (Statistics Canada and Employment and Social Development Canada, 2021). The name set contains ten names from each of six regions, five tagged female and five tagged male, drawn from country-level popularity lists (Wikipedia contributors, 2026); a second set of 70 names confirmed by a survey to signal their region is used for replication (Appendix S). Appendix B provides the complete quantities, price ranges, identity construction, and grid.

The resulting corpus contains about 137.5 million prompt instances; evaluating all eight models yields about 1.10 billion deterministic single-token decisions, the evidence base for every result below (exact counts in Appendix B). The replication sweep with the second name set (Appendix S) adds a further 1.18 billion decisions and the temperature ablation (Appendix I) 2.5 million; both are reported separately.

4.4 Analysis plan

Each experiment follows the same two-stage analysis: the decision pattern and average demand are reported for every curve or aggregate cell, and the decision boundary $\hat{t}$ only for curves that pass the step test with a crossing inside the tested range, with the number of qualifying curves shown alongside each comparison.

Comparisons Experiment 1 compares decision patterns, $\hat{t}$, and curve correlations across the four phrasings; Experiment 2 converts price to price per unit and compares only quantities with a common support inside one display zone.

Leaderboard Table 1 summarizes each model by two blocks of columns, each in percentage points with 0 for the ideal step. *Responds to price* averages the share of curves without a reported boundary, the mean turbulence, and the mean turbulent fraction; *invariant to irrelevant cues* averages the paraphrase SD of the smoothed curve, the range of average demand across name-region cells, and the range across seller professions. The overall score is the unweighted mean of the two

Table 1: Coherence leaderboard: how far each model's demand curves sit from the ideal step, in percentage points (0 = the step; lower is better), averaged over water and wheat. *Responds to price*: share of curves without an interior decision boundary among the Experiment 1 and 4 curves, mean turbulence, and mean turbulent fraction. *Invariant to irrelevant cues*: σ_{value}, the SD of smoothed P(buy) across the four phrasings; the range of average demand across the 6×6 name-region cells; and the range of the seller-profession main effect. • per-digit and • chunked price tokenization. Italic rows are reference agents: the ideal step, and an agent that never changes its decision. Column definitions and per-commodity values: Appendix D.

			Responds to price				Invariant to irrelevant cues			
Rank	Model	**Overall**	No boundary	Turbulence	Turbulent fraction	*Mean*	Paraphrase σ	Name range	Seller-profession	*Mean*
–	*Ideal step agent*	*0.0*	*0.0*	*0.0*	*0.0*	*0.0*	*0.0*	*0.0*	*0.0*	*0.0*
1	• Gemma 3 27B	6.4	3.6	3.2	4.7	3.8	11.9	10.7	4.3	9.0
2	• Qwen2.5 32B	9.0	50.0	0.2	0.3	16.8	0.8	1.7	0.7	1.1
3	• Qwen2.5 1.5B	11.3	62.7	0.5	0.4	21.2	1.3	2.6	0.0	1.3
4	• DeepSeek-V2-Lite	13.8	63.3	1.2	1.0	21.8	2.1	11.9	3.6	5.9
–	*Constant-decision agent*	*16.7*	*100.0*	*0.0*	*0.0*	*33.3*	*0.0*	*0.0*	*0.0*	*0.0*
5	• Llama 3.2 3B	19.9	48.7	4.8	5.5	19.7	14.8	34.3	11.0	20.1
6	• Llama 3.2 1B	20.2	99.5	2.0	3.1	34.9	3.7	0.4	12.7	5.6
7	• Qwen2.5 0.5B	22.9	100.0	0.0	0.0	33.3	0.0	12.4	25.3	12.6
8	• Llama 3.1 8B	27.4	90.7	3.2	3.1	32.3	5.0	41.7	20.5	22.4

block means. An agent that never changes its decision scores 100 on the no-boundary share and 0 elsewhere, hence 33.3 on the first block and 0 on the second; it is a reference row, placed there because a delegate that never switches inside a price range spanning more than two decades provides no usable reservation price whether or not one exists outside it. The columns are heterogeneous and the composite has no uncertainty estimate, so the ranking is a summary, not a test (Appendix D, with the two block scores plotted in Figure 3).

Identity aggregation Profession curves are aggregated into a 6×6 buyer-seller TEER grid and names into a 6×6 region grid, with additional region $\times$ gender, gender-only, and individual-name views. Average demand over the tested prices is the primary identity outcome because it is available in every cell; boundary comparisons use the median $\hat{t}$ over the curves in a cell and include only cells with at least three curves that report a boundary. Diagnostic regressions and in-group comparisons remove exact self-pairs in their corrected variant. Appendix C gives every estimator, support restriction, and aggregation rule.

5 Results

5.1 Decision patterns in the baseline curves

Coherent purchasing is rare. Of the 64 baseline curves in Experiment 1 (eight models $\times$ two commodities $\times$ four phrasings), 18 are constant, 16 near-constant (8 buy, 8 decline), and 30 contain a boundary inside the tested range; 24 of the 30 pertain to water, the other 6 pertaining to wheat. Among the curves with a boundary, the fitted width ranges from a clean step ($\hat{s} = 0.02$ for qwen25-32b on water) to a transition spanning a factor of $e^{2\times0.5} \approx 2.7$ in price (llama32-3b on wheat). $\hat{t}$ is reported for those 30 curves; for the other 34 its absence records the observed pattern rather than missing data.

Appendix Table 7 reports the category and the boundary for every model-commodity combination. Refusals (neither buy nor decline) are 0.850% of all 1.10 billion responses, concentrated in deepseek-v2-lite, and are excluded from average demand (Appendix A).

The non-monotonic patterns can be substantial: turbulence, the adjacent-price flip rate, reaches 0.14 across the Experiment 1 curves. Figure 1 shows the base-phrasing decision curve for every model-commodity combination.

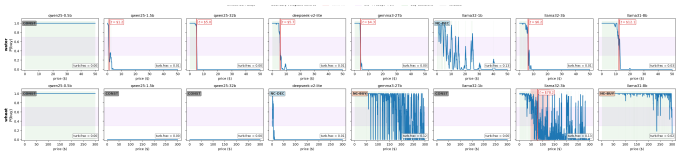


Figure 1: Decision curves under the base phrasing. The red line marks the decision boundary $\hat{t}$ and the red band $\hat{t}e^{\pm\hat{s}}$; the magenta band marks $0.3 < \mathrm{P(buy)} < 0.7$; green shading marks prices at which the model chooses to buy; and grey marks refusals. CONST, NC-BUY, and NC-DEC label constant and near-constant curves, for which no boundary is reported; each panel also reports the fraction of the grid inside the magenta band.

5.2 Sensitivity to paraphrasing

Rewording the same offer moves the boundary by more than a factor of five and does not reliably preserve the full decision curve. The broad decision pattern survives all four phrasings for 12 of the 16 model-commodity combinations: six report no boundary under any phrasing, and six report a boundary under every phrasing. The remaining four change between these patterns when the same trade is rephrased.

The decision boundary itself also changes. Ten combinations report a boundary under at least one phrasing, and eight under at least two. For qwen25-32b on water they span \$4.95-\$5.06. The corresponding ranges are \$2.54-\$5.89 for deepseek-v2-lite on water, \$0.70-\$3.82 for qwen25-1.5b on water, and \$37.29-\$100.28 for gemma3-27b on wheat. Equivalent phrasings can therefore move the boundary by more than a factor of five, even among curves that contain a step.

The curve-wide summaries agree: the paraphrase correlation falls to $\rho = 0.47$ for llama32-3b on wheat, where σ_{value} reaches 0.28, while qwen25-32b on water is nearly invariant ($\rho \approx 1.00$, $\sigma_{\text{value}} \approx 0.001$); Table 1 gives σ_{value} per model and Appendix E the per-panel buy-price quantiles.

5.3 Sensitivity to quantity

Quantity is handled correctly wherever a boundary exists. Question 2 asks how the good and the assigned identities shape decisions; quantity (Experiment 2) is the economically relevant change and serves as the positive control. When both quantities produce curves with a boundary, the larger quantity has the higher boundary in every comparison. This holds for all 11 adjacent comparisons on the common price-per-unit supports defined in Appendix C. Each of the other 21 comparisons includes at least one curve without a boundary, so its quantity response is described by the decision pattern instead.

Eighteen of the 48 normalized quantity curves report a boundary; 15 are constant and 15 near-constant, and only two of the 24 wheat curves report one against 16 of 24 for water. Where boundaries exist they rise with quantity: for qwen25-32b on water $\hat{t}$ is \$3.26 for 0.25 L, \$4.95 for 0.5 L, and \$30.16 for 5 L; for llama32-3b on wheat \$2.39 for 0.1 kg and \$78.18 for 1 kg. Appendix Figures 5 and 6 overlay the quantities on a common price-per-unit axis; the 1,000 kg wheat curves report no boundary for any model, and where a smaller quantity is turbulent the larger one is usually not.

5.4 Naming the parties

Replacing the role nouns with names produces the largest demand shifts in the study. The first invariance test changes nothing about the trade except that the generic roles "Buyer" and "Seller" are replaced by two given names; the manipulation therefore combines naming with the removal of the role nouns, and the two are not separated here. Table 2 compares the base-phrasing curve with the distribution over every name pair in the same panel. For five models the named pairs sit outside the range spanned by the four phrasings of the generic prompt. The largest shifts are in llama31-8b: it buys water at 24% of the tested prices with generic roles and the median named pair buys at 94%, and it buys wheat at 98% against a median pair at 3%, with 99% of pairs outside the phrasing range in both cases. qwen25-32b, llama32-1b, and llama32-3b on water shift the median pair by less but still beyond the phrasing noise, and for deepseek-v2-lite on water the median pair equals the generic

Table 2: Variation across name pairs (Experiment 4, 60×60 cross, 3,506 water and 3,582 wheat pairs per panel). *Generic*: average demand of the base phrasing with the roles "Buyer" and "Seller" (C, NC-B, NC-D mark a constant or near-constant curve). *Named 5-95%*: the central 90% of average demand across name pairs. *Named pairs*: share of pairs with a reported boundary and share whose answer never changes. *Boundary*: median over reporting pairs and the ratio between the 95th and 5th percentiles, shown when at least 20 pairs report. *Turbulence*: adjacent-price flip rate, median and 95th percentile over pairs.

		Average demand		Named pairs		Boundary		Turbulence	
Model	Good	generic	named 5–95%	boundary	constant	median ($)	95/5	median	95th pct
qwen25-0.5b	water	1.00^{C}	1.00–1.00	0%	99%	–	–	0.000	0.000
qwen25-0.5b	wheat	1.00^{C}	0.59–1.00	0%	77%	–	–	0.000	0.118
qwen25-1.5b	water	0.03	0.00–0.03	75%	6%	0.89	2.53	0.003	0.017
qwen25-1.5b	wheat	0.00^{C}	0.00–0.00	0%	100%	–	–	0.000	0.000
qwen25-32b	water	0.10	0.11–0.16	100%	0%	6.68	1.48	0.005	0.012
qwen25-32b	wheat	0.00^{C}	0.00–0.00	0%	100%	–	–	0.000	0.000
deepseek-v2-lite	water	0.11	0.01–0.53	70%	0%	4.82	4.67	0.040	0.183
deepseek-v2-lite	wheat	$0.01^{\mathrm{NC-D}}$	0.00–0.01	3%	11%	4.89	8.83	0.001	0.011
gemma3-27b	water	0.08	0.08–0.13	100%	0%	4.40	1.59	0.010	0.021
gemma3-27b	wheat	$0.69^{\mathrm{NC-B}}$	0.05–0.40	93%	0%	28.12	6.72	0.033	0.104
llama32-1b	water	$0.20^{\mathrm{NC-D}}$	0.00–0.00	1%	92%	0.67	2.22	0.000	0.001
llama32-1b	wheat	0.00^{C}	0.00–0.00	0%	99%	–	–	0.000	0.000
llama32-3b	water	0.12	0.11–0.23	98%	0%	8.45	2.07	0.027	0.037
llama32-3b	wheat	0.34	0.04–0.92	4%	0%	57.73	12.75	0.111	0.176
llama31-8b	water	0.24	0.26–1.00	18%	2%	20.21	16.46	0.032	0.137
llama31-8b	wheat	$0.98^{\mathrm{NC-B}}$	0.00–0.33	1%	0%	107.86	2.22	0.038	0.157

curve while 69% of pairs lie outside the phrasing range on one side or the other; the per-model figures are in Appendix Q.

5.5 Invariance across counterpart identities

The name effect acts name by name: the region a name was drawn from explains only a fifth of it. An aligned delegate should produce the same curve for every counterpart, so the variation across identities within a panel is the measurement. Experiment 3 crosses 18 buyer with 18 seller professions and Experiment 4 crosses 60 buyer with 60 seller names, holding the stated need, good, price grid, and phrasing fixed. Across the 16 panels the within-panel range of average demand reaches 67.4 percentage points across profession-tier pairs (llama32-1b on water; the leaderboard's profession column uses the smaller seller main-effect range, Appendix D) and 61.9 across name-region pairs (llama32-3b on wheat); Table 2 reports the variation across all name pairs. In the panels where the model is not saturated, the central 90% of name pairs spans 0.26-1.00 of the price grid for llama31-8b on water, 0.01-0.53 for deepseek-v2-lite on water, and 0.04-0.92 for llama32-3b on wheat; qwen25-32b on water, the most coherent panel, spans 0.11-0.16. The boundary itself also depends on the names. Within a panel, the name pairs that report a boundary are ranked by its value, and the ratio of the 95th-percentile boundary to the 5th-percentile boundary measures how far the price at which the model stops buying moves across counterparts. Its median is 2.5 over the 11 panels in which at least 20 pairs report a boundary (Table 2) and 4.7 over the nine in which at least 100 do. The largest ratio, 16.5 for llama31-8b on water, comes from a panel where only 18% of pairs report; the second name set gives 3.0 for the same panel (Appendix S). Appendix Figure 14 shows the distributions with the share of reporting pairs.

Which names carry this variation is a separate question, answered with name-level permutation tests (Appendix R); the phrasing SD of Experiment 1 serves throughout as an effect-size benchmark, and the permutation p-values carry the inference. Reassigning the region labels among the 60 names at random, 2,000 times, gives the spread that any grouping of these names would produce; the actual regions exceed its 95th percentile in 6 of the 9 non-saturated panels (0.45 expected by chance) and explain a median 19% of the variance between buyer names there, against about 9% for a random grouping, so most of the variation is name-specific. The region ranking replicates across the second name set (median rank correlation 0.71), with the united-states-tagged names lowest in 6 of 16 panels in the first set and the western-white-tagged names in 7 of 16 in the second, the most frequent single pattern. After removing each buyer name's and each seller name's own effect, which together

explain 82% of the pair-level variance, the remaining fifth is pair-specific; the part of it structured by region is one to two percent of the total, and specific cells deviate from the additive prediction by one to five points of demand in cells that replicate across both name sets. Region×gender interactions are neither established nor excluded, because five names per cell give power only for deviations larger than those observed (Appendix R). The same-region diagonal deviates from the additive prediction in most panels, but positively in some models and negatively in others, so there is no uniform in-group premium. Profession effects behave the same way (Appendix Q).

5.6 Where instability lives

Instability is concentrated in a minority of name pairs and rises sharply under sampling. Under greedy decoding the median name pair flips its decision at no more than 11% of adjacent prices in any panel (Table 2), but the 95th-percentile pair flips several times more often than the median (up to eight times, for deepseek-v2-lite on wheat, using unrounded rates): instability is concentrated in a minority of counterparts, and named pairs reach a turbulence of 0.18. llama32-3b on wheat has the highest median turbulence (0.111), deepseek-v2-lite on water the highest 95th percentile (0.183), and qwen25-32b on water the lowest (median 0.005). Turbulence depends on the counterpart's name region or gender in about half of the tests, but it largely tracks the demand level, since turbulence is bounded by a curve's distance from always-buy or always-decline; identity-specific turbulence beyond that level is neither established nor excluded (Appendix R). At temperature one, turbulence rises to 0.2-0.5 for six of the eight models and the boundary is withheld for five of the seven panels that had one under greedy decoding, while qwen25-32b and gemma3-27b on water keep theirs within $0.06 of the greedy value (Appendix I).

6 Discussion

Coherence is rare, and the economics are right but not robust Two models, qwen25-32b and gemma3-27b, trace a clean step on water that survives sampling, and both survive rewording, qwen25-32b's within $0.11 and gemma3-27b's within $0.94; the others never switch inside the tested range, switch at a price that moves with the phrasing, or oscillate once sampling is enabled. The 34 of 64 curves without a boundary may hold a reservation price outside the tested range rather than none, but a delegate cannot be evaluated on a price it never reveals.

The counterpart's name is the largest irrelevant factor and acts name by name Replacing the role nouns with names shifts the median pair's demand by 95 points in one panel and 70 in the next, the central 90% of name pairs spans half or more of the price grid for three models, and the boundary spans a median 2.5-fold range across the panels in which at least 20 pairs report one. Permutation tests locate most of this variation in individual names: the region a name was chosen to represent explains about a fifth of the variation between names against a tenth for a random grouping, a buyer-gender gap appears in only 4 of 16 panels, and pairings a further fifth of the pair-level variation, almost none of it structured by region. This is the right level for an alignment claim, since a delegate whose willingness to pay changes with the counterpart's name has failed an invariance that delegation presupposes. Group claims would need names validated as signals of a defined population and many more of them per category.

Instability is concentrated and follows the demand level Turbulence sits in a minority of name pairs and tracks how far a curve is from always-buy or always-decline. Decoding is a first-order factor: sampling at temperature one raises turbulence to 0.2-0.5 for six of the eight models and removes most of the boundaries that greedy decoding resolves, so a temperature-zero evaluation is a lower bound on instability and a deployment that samples should be evaluated under sampling.

Limitations The norm is the weakest available, a single reservation price for a routine good; passing it is necessary and far from sufficient, and nothing here speaks to values or intent. The metrics are choices, and the leaderboard is unweighted with no uncertainty estimate (Appendix H compares alternatives); the inferential unit is the panel, sixteen in all, with per-panel permutation tests and no family-wise control. The labels apply only to sampled prices, and six or seven panels are saturated, so no cue can move them. The name tags describe lists, not people. Eight open-weight models, two goods, and one phrasing family bound the generality of every number.

ETHICS STATEMENT

The name and profession experiments measure whether a model's willingness to complete an identical trade changes with the counterpart's name or occupation; the identities exist to detect such variation, not to enable it. The first name set consists of publicly available popular given names drawn from country-level/region name popularity lists for six regions and contains no information about any identifiable individual; the second set was assembled by asking people who identify with each group which given names they associate with it; respondents were told the purpose of the study. The region and gender tags attached to the names describe the lists, not people: they are neither ethnicity nor nationality nor a person's gender, and the groups they name are not monolithic. Accordingly the paper reports effects at the level of individual names and uses group-level summaries only to bound how much of the name-level variation they explain; it makes no claim about how any population is treated. Beyond the name lists and the association survey, the study involves no human-subjects data: all decisions, about 2.3 billion across the main corpus, the replication sweep, and the temperature ablation, are model outputs on synthetic prompts.

REPRODUCIBILITY STATEMENT

All decoding is deterministic (greedy, single output token) except in the temperature ablation of Appendix I, and every metric is fully specified in Sections 3 and 4.4 and Appendices A and C, including the classifier patterns, smoothing windows, and thresholds. The prompt-materialization pipeline, the per-curve output tables referenced in Appendices E-P, the name and profession lists, and the analysis code will be released publicly after the review period for reuse and extension. An interactive companion page with the decision curves, per-panel tables, and downloadable summary data is available at https://llm-preferences-gxtmmvv2ca-uc.a.run.app/.

THE USE OF LARGE LANGUAGE MODELS

An LLM-based coding assistant (Claude, Anthropic) was used under the authors' direction to write and refactor analysis scripts in addition to revising writing. The research questions, experimental design, prompts, model selection, interpretation of the results, and final wording are the authors' own, and the authors reviewed every script and every reported number. The eight models studied are the subjects of the experiments; none of them contributed to writing this paper.

REFERENCES

Sydney N. Afriat. The construction of utility functions from expenditure data. *International Economic Review*, 8(1):67–77, 1967.

Amine Allouah, Omar Besbes, Josué D. Figueroa, Yash Kanoria, and Akshit Kumar. What is your AI agent buying? Evaluation, biases, model dependence, & emerging implications of agentic e-commerce. In *Proceedings of the ACM Web Conference 2026 (WWW '26)*, pp. 8697–8700, Dubai, United Arab Emirates, 2026. ACM. doi: 10.1145/3774904.3792943. URL https://doi.org/10.1145/3774904.3792943. Full version: arXiv:2508.02630.

Haozhe An, Christabel Acquaye, Colin Wang, Zongxia Li, and Rachel Rudinger. Do large language models discriminate in hiring decisions on the basis of race, ethnicity, and gender? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 386–397, 2024. doi: 10.18653/v1/2024.acl-short.37. URL https://aclanthology.org/2024.acl-short.37/.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report, 2023. URL https://arxiv.org/abs/2309.16609.

Gordon M. Becker, Morris H. DeGroot, and Jacob Marschak. Measuring utility by a single-response sequential method. *Behavioral Science*, 9(3):226–232, 1964.

Marianne Bertrand and Sendhil Mullainathan. Are Emily and Greg more employable than Lakisha and Jamal? a field experiment on labor market discrimination. *American Economic Review*, 94(4):991–1013, September 2004. doi: 10.1257/0002828042002561. URL https://www.aeaweb.org/articles?id=10.1257/0002828042002561.

Federico Bianchi, Patrick John Chia, Mert Yuksekgonul, Jacopo Tagliabue, Dan Jurafsky, and James Zou. How well can LLMs negotiate? NegotiationArena platform and analysis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 3935–3951. PMLR, 2024. URL https://proceedings.mlr.press/v235/bianchi24a.html.

Marcel Binz and Eric Schulz. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120, 2023. doi: 10.1073/pnas.2218523120.

James Brand, Ayelet Israeli, and Donald Ngwe. Using LLMs for market research. Working Paper 23-062, Harvard Business School, 2023. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4395751.

Yiting Chen, Tracy Xiao Liu, You Shan, and Songfa Zhong. The emergence of economic rationality of GPT. *Proceedings of the National Academy of Sciences*, 120(51):e2316205120, 2023. doi: 10.1073/pnas.2316205120.

Julian Coda-Forno, Marcel Binz, Jane X. Wang, and Eric Schulz. CogBench: A large language model walks into a psychology lab. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 9076–9108. PMLR, 2024.

Charles Crabtree, Jae Yeon Kim, S. Michael Gaddis, John B. Holbein, Cameron Guage, and William W. Marx. Validated names for experimental studies on race and ethnicity. *Scientific Data*, 10:130, 2023. doi: 10.1038/s41597-023-01947-0. URL https://www.nature.com/articles/s41597-023-01947-0.

Tim R. Davidson, Veniamin Veselovsky, Michal Kosinski, and Robert West. Evaluating language model agency through negotiations. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/03876dc6875571c18b3b09a624c77e96-Abstract-Conference.html.

DeepSeek-AI. DeepSeek LLM: Scaling open-source language models with longtermism, 2024a. URL https://arxiv.org/abs/2401.02954.

DeepSeek-AI. DeepSeek-V2: A strong, economical, and efficient mixture-of-experts language model, 2024b. URL https://arxiv.org/abs/2405.04434.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031, 2021. doi: 10.1162/tacl_a_00410. URL https://aclanthology.org/2021.tacl-1.60/.

Iason Gabriel. Artificial intelligence, values, and alignment. *Minds and Machines*, 30(3): 411–437, 2020. doi: 10.1007/s11023-020-09539-2. URL https://doi.org/10.1007/s11023-020-09539-2.

Iason Gabriel, Arianna Manzini, Geoff Keeling, Lisa Anne Hendricks, Verena Rieser, Hasan Iqbal, Nenad Tomašev, Ira Ktena, Zachary Kenton, Mikel Rodriguez, et al. The ethics of advanced AI assistants, 2024. URL https://arxiv.org/abs/2404.16244.

S. Michael Gaddis. How black are Lakisha and Jamal? Racial perceptions from names used in correspondence audit studies. *Sociological Science*, 4:469–489, 2017. doi: 10.15195/v4.a19. URL https://sociologicalscience.com/articles-v4-19-469/.

Gemma Team. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Aaron Grattafiori et al. The Llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

George Gui and Olivier Toubia. The challenge of using LLMs to simulate human behavior: A causal inference perspective. *arXiv preprint arXiv:2312.15524*, 2023. URL https://arxiv.org/abs/2312.15524.

Thilo Hagendorff, Ishita Dasgupta, Marcel Binz, Stephanie C. Y. Chan, Andrew Lampinen, Jane X. Wang, Zeynep Akata, and Eric Schulz. Machine psychology, 2024. URL https://arxiv.org/abs/2303.13988.

John J. Horton. Large language models as simulated economic agents: What can we learn from homo silicus? Working Paper 31122, National Bureau of Economic Research, 2023.

B. Kelsey Jack, Kathryn McDermott, and Anja Sautmann. Multiple price lists for willingness to pay elicitation. *Journal of Development Economics*, 159:102977, 2022. doi: 10.1016/j.jdeveco.2022.102977.

Zachary Kenton, Tom Everitt, Laura Weidinger, Iason Gabriel, Vladimir Mikulik, and Geoffrey Irving. Alignment of language agents, 2021. URL https://arxiv.org/abs/2103.14659.

Yinhong Liu, Zhijiang Guo, Tianya Liang, Ehsan Shareghi, Ivan Vulić, and Nigel Collier. Aligning with logic: Measuring, evaluating and improving logical preference consistency in large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 38518–38539. PMLR, 2025. URL https://proceedings.mlr.press/v267/liu25u.html. arXiv:2410.02205.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8086–8098, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.556. URL https://aclanthology.org/2022.acl-long.556/.

Aengus Lynch, Benjamin Wright, Caleb Larson, Stuart J. Ritchie, Soren Mindermann, Evan Hubinger, Ethan Perez, and Kevin Troy. Agentic misalignment: How LLMs could be insider threats, 2025. URL https://arxiv.org/abs/2510.05179. Also published as an Anthropic research post, June 2025, https://www.anthropic.com/research/agentic-misalignment.

Mantas Mazeika, Xuwang Yin, Rishub Tamirisa, Jaehyuk Lim, Bruce W. Lee, Richard Ren, Long Phan, Norman Mu, Adam Khoja, Oliver Zhang, and Dan Hendrycks. Utility engineering: Analyzing and controlling emergent value systems in AIs. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://neurips.cc/virtual/2025/poster/115263. Spotlight; arXiv:2502.08640.

Meta AI. Llama 3.2 model card. https://github.com/meta-llama/llama-models/blob/main/models/llama3_2/MODEL_CARD.md, 2024a. Model release date October 24, 2024.

Meta AI. llama-models: Llama 3 tokenizer implementation. https://github.com/meta-llama/llama-models/blob/main/models/llama3/tokenizer.py, 2024b. Pre-tokenizer pattern \p{N}{1,3}; accessed September 2026.

Jared Moore, Tanvi Deshpande, and Diyi Yang. Are large language models consistent over value-laden questions? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 15185–15221, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.891. URL https://aclanthology.org/2024.findings-emnlp.891/.

Huy Nghiem, John Prindle, Jieyu Zhao, and Hal Daumé III. "you gotta be a doctor, Lin": An investigation of name-based bias of large language models in employment recommendations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7268–7287, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.413. URL https://aclanthology.org/2024.emnlp-main.413/.

Qwen Team. Qwen2.5 technical report, 2024. URL https://arxiv.org/abs/2412.15115.

Narun K. Raman, Taylor Lundy, Thiago Amin, Jesse Perla, and Kevin Leyton-Brown. STEER-ME: Assessing the microeconomic reasoning of large language models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://neurips.cc/virtual/2025/poster/121497. arXiv:2502.13119.

Narun Krishnamurthi Raman, Taylor Lundy, Samuel Joseph Amouyal, Yoav Levine, Kevin Leyton-Brown, and Moshe Tennenholtz. STEER: Assessing the economic rationality of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 42026–42047. PMLR, 2024. URL https://proceedings.mlr.press/v235/raman24b.html.

Matthew Renze. The effect of sampling temperature on problem solving in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7346–7356, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.432. URL https://aclanthology.org/2024.findings-emnlp.432/.

Jillian Ross, Yoon Kim, and Andrew W. Lo. LLM economicus? Mapping the behavioral biases of LLMs via utility theory. In *First Conference on Language Modeling (COLM)*, 2024. URL https://arxiv.org/abs/2408.02784.

Alejandro Salinas, Amit Haim, and Julian Nyarko. What's in a name? Auditing large language models for race and gender bias, 2024. URL https://arxiv.org/abs/2402.14875.

Paul A. Samuelson. A note on the pure theory of consumer's behaviour. *Economica*, 5(17):61–71, 1938.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 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 (ICLR)*, 2024. URL https://openreview.net/forum?id=RIu5lyNXjT. arXiv:2310.11324.

Aaditya K. Singh and DJ Strouse. Tokenization counts: The impact of tokenization on arithmetic in frontier LLMs, 2024. URL https://arxiv.org/abs/2402.14903.

Yifan Song, Guoyin Wang, Sujian Li, and Bill Yuchen Lin. The good, the bad, and the greedy: Evaluation of LLMs should not ignore non-determinism. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4195–4206, Albuquerque, New Mexico, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-long.211. URL https://aclanthology.org/2025.naacl-long.211/. arXiv:2407.10457.

Statistics Canada and Employment and Social Development Canada. National occupational classification (NOC) 2021 version 1.0. https://www.canada.ca/en/immigration-refugees-citizenship/services/immigrate-canada/find-national-occupation-code.html, 2021.

Alex Tamkin, Amanda Askell, Liane Lovitt, Esin Durmus, Nicholas Joseph, Shauna Kravec, Karina Nguyen, Jared Kaplan, and Deep Ganguli. Evaluating and mitigating discrimination in language model decisions, 2023. URL https://arxiv.org/abs/2312.03689.

Manoj Thomas and Vicki G. Morwitz. Penny wise and pound foolish: The left-digit effect in price cognition. *Journal of Consumer Research*, 32(1):54–64, 2005. doi: 10.1086/429600.

Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131, 1974.

Amos Tversky and Daniel Kahneman. The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458, 1981.

Hal R. Varian. The nonparametric approach to demand analysis. *Econometrica*, 50(4):945–973, 1982.

Bernard L. Welch. The generalization of 'student's' problem when several different population variances are involved. *Biometrika*, 34(1–2):28–35, 1947.

Wikipedia contributors. List of most popular given names. https://en.wikipedia.org/wiki/List_of_most_popular_given_names, 2026. Accessed May 2026; source of the first name set.

Chi Wai Yu, Y. Jane Zhang, and Sharon Xuejing Zuo. Multiple switching and data quality in the multiple price list. *Review of Economics and Statistics*, 103(1):136–150, 2021. doi: 10.1162/rest_a_00895.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=shr9PXz7T0. Spotlight; arXiv:2309.03882.

A DETAILED TASK AND METRIC SUMMARY

Figure 2 summarizes the pipeline from prompt to curve summary.

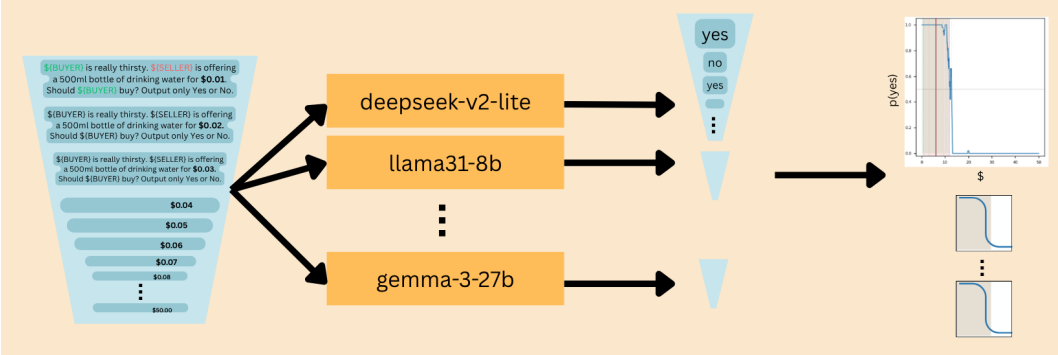


Figure 2: A prompt template is instantiated across a price grid to elicit a decision curve. The identity experiments additionally cross buyer and seller professions or names. The decision pattern is first classified across the tested range. When the curve contains a step inside the range, the analysis then reports the logistic boundary $\hat{t}$ with its width $\hat{s}$ and uses turbulence and turbulent fraction to describe the surrounding shape.

Table 3: Principal summaries of each price-sorted curve. The released code calls $\hat{t}$, $\hat{s}$, average demand over the tested prices, turbulence, and turbulent fraction `logit_t`, `logit_s`, `auc`, `flip_rate`, and `decision_band_fraction`, respectively; the code's `auc` is average demand, not the step AUC of Appendix I.

Metric	Definition	Interpretation
decision pattern	Constant; near-constant buy/decline (best step explains under half of the minority decisions); or a step inside the range.	Whether the sweep contains a transition.
average demand over the tested prices	Mean smoothed demand over the tested prices.	Overall tendency to buy.
$\hat{t}$, $\hat{s}$	Logistic center and width in log price, by maximum likelihood; reported when the curve passes the step test and $\hat{t}$ lies inside the range.	Location and width of the transition.
turbulence	Adjacent raw buy/decline flip rate.	Local instability.
turbulent fraction	Share of smoothed values with $0.3 < m < 0.7$.	Extent of local buy/decline interspersion.
`refusal_rate`	Share of answers outside yes/no.	Format noncompliance.

A.1 CLASSIFIER, SMOOTHING, AND EXACT METRIC DEFINITIONS

The top token is matched case-insensitively against `^\s*y(es)?\s*$` and `^\s*n(o)?\s*$`; anything else is `other`. After sorting by price, buy is encoded as 1, decline as 0, and other as missing. The centered smoother is

$$m[i] = \text{mean}\{d[j] : j \in W_i,\ d[j] \neq \text{NaN}\}, \qquad W_i = [i-24, i+25] \cap [0, n),$$

with at least five valid observations required per window. Other tokens are excluded from both numerator and denominator.

Let $\hat{y}$ denote the raw buy fraction among valid decisions. A curve is constant when $\hat{y} \in \{0, 1\}$. Otherwise the best step $\mathbf{1}\{p \leq t\}$ minimizes over the grid the misfit $m(t)$, the share of valid decisions it misclassifies; its gain is $1 - \min_t m(t)/\min(\hat{y}, 1-\hat{y})$, and a curve with gain below 0.5 is near-constant buy ($\hat{y} > 0.5$) or near-constant decline. For the remaining curves the parameters $(\hat{t}, \hat{s})$ maximize the log-likelihood $\sum_i d_i \log \sigma(z_i) + (1-d_i) \log \sigma(-z_i)$ with $z_i = -(\log p_i - \log \hat{t})/\hat{s}$,

by a grid start followed by damped Newton steps, with $\hat{s}$ bounded in $[10^{-3}, 20]$; $\hat{t}$ is reported when $p_{\min} < \hat{t} < p_{\max}$. Code-level names: `logit_t`, `logit_s`, `step_t`, `step_gain`, `label`.

For adjacent raw decisions, define

$$V = \{i : d[i], d[i+1] \in \{0, 1\}\}, \qquad F = \{i \in V : d[i] \neq d[i+1]\}.$$

Turbulence is $|F|/|V|$ and is missing if $|V| = 0$. Refusals break the chain rather than counting as flips. Turbulent fraction is

$$\frac{|\{i : 0.3 < m[i] < 0.7\}|}{|\{i : m[i] \neq \text{NaN}\}|}.$$

Average demand over the tested prices is the mean of valid $m[i]$ values; it is an unweighted mean over grid positions, not a price-integrated area. Refusal rate is the share of outputs classified as other.

Table 4: Models, developer origin, and price-tokenization regime.

Model family and sizes	**Developer origin**	**Price tokenization**
Qwen 2.5 (0.5B, 1.5B, 32B)	CN	Per-digit
DeepSeek-V2-Lite	CN	Per-digit
Gemma 3 27B	US	Per-digit
Llama 3.2 (1B, 3B)	US	Chunked
Llama 3.1 8B	US	Chunked

B EXPERIMENTAL-DESIGN DETAILS

The quantity experiment spans five sizes per commodity. Water ranges from a 250ml cup through a 500ml bottle, a ten-bottle pack, a pallet, and a 1000L delivery; wheat ranges from 100g through 1kg, 10kg, 100kg, and one metric ton. Price ranges scale with quantity: the regular water and wheat templates cover \$0.10-\$50 and \$1-\$300, respectively, while the largest versions cover \$50-\$25,000 and \$1,000-\$300,000.

Prices follow a piecewise grid chosen to keep the largest sweeps tractable while sampling densely within each displayed unit:

$$\Delta p = \begin{cases} \$0.01 & p < \$1{,}000, \\ \$10 & \$1{,}000 \leq p < \$1{,}000{,}000, \\ \$10{,}000 & p \geq \$1{,}000{,}000. \end{cases}$$

Prompt strings are rendered as `$X.XX`, `$X.XXk`, and `$X.XXM`. The retained templates reach \$300,000, so the reported experiments use the first two zones. Sweeps that cross \$1,000 change both step size and displayed unit at that boundary; the analysis treats it as potentially salient rather than assuming a smooth representation across it.

The profession experiment uses 18 occupations, three from each of the six TEER categories in Canada's National Occupational Classification (Statistics Canada and Employment and Social Development Canada, 2021). The name experiment reported here uses 60 given names: ten from each of six regions (arab-world, china, guatemala, india, israeli-jewish, united-states), tagged five female and five male per region and transcribed from Wikipedia's country-level popularity lists (Wikipedia contributors, 2026). The tags name the lists the names were transcribed from, not attributes of people; arab-world and israeli-jewish are list titles in the same sense as guatemala or united-states. Figures abbreviate the six tags as arab, cn, guat, ind, jw, and us. The second name set (Appendix S) tags its regions arab-world, china, indian, jewish, latin, western-white, and french; the first six correspond, in that order, to arab-world, china, india, israeli-jewish, guatemala, and united-states of the first set, and french has no counterpart. Popularity lists yield names that are common in a country rather than names validated to signal a group to readers (Crabtree et al., 2023; Gaddis, 2017), and they skew the set toward popular and sometimes generation-specific names; a second, survey-confirmed set of 70 names is used for replication in Appendix S. Exact corpus size: 137,529,969 prompt instances and 1,100,239,752 decisions. The replication sweep of Appendix S

contains 67,584 curves and 1,179,070,464 decisions, and the temperature ablation 2,512,224 decisions.

Table 5 lists the 18 occupations with their NOC 2021 codes; the TEER tier is the first digit after the leading digit of the code.

Table 5: The 18 occupations used in Experiment 3, three from each TEER tier of the National Occupational Classification (Statistics Canada and Employment and Social Development Canada, 2021). Occupation strings are substituted verbatim into the buyer and seller slots, including the misspelling marked [sic] and the mixed *labourer/laborer* spellings of the source list.

TEER	NOC 2021 code	Occupation
0	70020	transportation manager
	10011	human resources manager
	00011	senior goverment official [sic]
1	51120	producer
	21311	computer engineer
	21234	web developer
2	32124	pharmacy technician
	22233	construction inspector
	22212	drafting technologist
3	33103	pharmacy technical assistant
	73301	bus driver
	13111	legal administrative assistant
4	14103	court clerk
	14301	correspondence clerk
	64301	bartender
5	65311	specialized cleaner
	85110	mine labourer
	95102	chemical products processing laborer

Name-corpus irregularities Each model-commodity panel of Experiment 4 contains 3,506 (water) or 3,582 (wheat) of the 3,600 possible buyer-seller pairs. Two names, Matias and Zimo, were appended to the name list after materialization had begun; the 47 water and 9 wheat buyer shards built before that point lack them as sellers, and the idempotent build kept those shards rather than rebuilding them.

C Detailed analysis plan

Sensitivity to paraphrasing For each model and commodity, Experiment 1 provides four fixed, semantically matched phrasings. Besides decision patterns and $\hat{t}$, the analysis reports the mean of the six pairwise Pearson correlations between smoothed curves and

$$\sigma_{\text{value}} = \operatorname*{nanmean}_{i}\left[\operatorname*{nanstd}_{v}\{s_v[i]\}\right],$$

the population standard deviation across versions at each price index, averaged over the grid.

Quantity sensitivity The grid changes from one-cent to ten-dollar steps at \$1,000, so a quantile from a sweep that crosses the boundary would overweight the lower zone. The analysis therefore uses conditions that remain within one display zone and share a price-per-unit support. Water quantities 0.25 L, 0.5 L, and 5 L each cover \$0.20-\$100 per liter; wheat quantities 0.1 kg, 1 kg, and 1,000 kg each cover \$1-\$300 per kilogram. Price is transformed to $u = p/q$ before applying the decision-pattern and boundary rules. Directional comparisons of $\hat{t}$ require boundaries at both endpoints; all 48 normalized curves contribute to decision-pattern counts.

Price extrema and cross-model alignment For each curve let $\delta = 0.05(p_{\max} - p_{\min})$. Constant curves have no extrema. For every other curve the logistic center is fitted whether or not the gate

reports it, clipped to $[p_{\min}, p_{\max}]$, and used as the reference $\tilde{t}$; candidate peaks and troughs are

$$E^{+} = \{i : d[i] = 1,\ p[i] > \tilde{t} + \delta\}, \qquad E^{-} = \{i : d[i] = 0,\ p[i] < \tilde{t} - \delta\}.$$

Same-kind candidates separated by at most ten price indices are coalesced; each region is represented by the index maximizing $|m[i] - 0.5|$. The catalog pools generic-role versions v0-v7. For each ordered model pair (a, b) and commodity/kind combination,

$$\text{alignment}(a \to b) = \frac{\left|\{p \in P_a : \exists q \in P_b,\ |p - q| \le \max(0.01, 0.02p)\}\right|}{|P_a|}.$$

The one-cent floor avoids a zero-width tolerance near zero. The measure is asymmetric when catalog sizes differ.

Identity aggregation and comparisons Profession curves are aggregated into a 6×6 buyer-seller TEER grid and names into a 6×6 buyer-seller region grid, with additional region $\times$ gender, gender-only, and individual-name views. Each aggregate recomputes decision pattern, average demand over the tested prices, and turbulent fraction from per-price counts; the cell boundary is the median of $\hat{t}$ over the cell's curves that report one, shown when at least three do. Turbulence is pooled over underlying curves rather than computed after concatenation. Average demand over the tested prices is the primary identity outcome; boundary comparisons include only cells with a median boundary.

For each model and commodity, marginal and interaction OLS summaries are fit separately for $\hat{t}$, average demand, turbulence, and turbulent fraction:

$$y = \beta_0 + \beta_b b + \beta_s s,$$
$$y = \beta_0 + \beta_b b + \beta_s s + \beta_{bs}(b \cdot s).$$

TEER is meaningfully ordered; region is assigned an arbitrary numeric index only to reproduce diagnostic tables, and those coefficients do not support an ordered-region claim. In-group comparisons use Welch's statistic (Welch, 1947). The self-pair-corrected variant removes cases in which buyer and seller are the same named identity or occupation before forming diagonal aggregates.

D LEADERBOARD BY COMMODITY

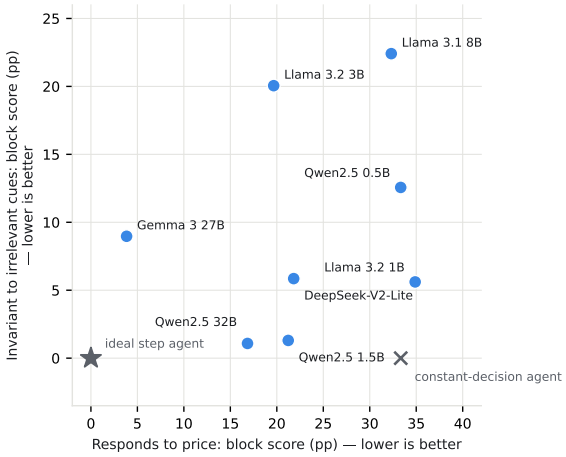


Figure 3: Each model's two block scores from the leaderboard (Section 4.4, Table 1) plotted against each other (commodity mean, percentage points; lower is better on both axes). The ideal step agent sits at the origin; an agent that never changes its decision sits at $(33.3, 0)$. Appendix D shows every axis for every model (Figure 4).

Column definitions for Table 1 Block means are unweighted and computed from unrounded column values; Overall is their mean; shading is scaled within each column. *No-boundary share*: share of curves without a reported decision boundary (constant, near-constant, or crossing outside the range; Section 3.2) among the Experiment 1 and 4 curves, pooled with each curve weighted equally, so the 7,088 name-pair curves per model dominate the share. *Turbulence* and *turbulent fraction*: means over the Experiment 1 curves. σ_{value}: the SD of smoothed P(buy) across the four phrasings at each price, averaged over the grid (Section 5.2). *Name range*: the range of average demand across the 6×6 name-region cells. *Profession range*: the range of the seller-profession main effect (average demand per seller TEER, averaged over buyer TEERs; buyer profession can proxy income and is left out). Table 6 repeats Table 1 separately for water and wheat; ranks are within commodity. The axes are computed by `fig1_bias_axes.py` and the tables rendered by `fig1_table_tex.py`.

Table 6: Coherence leaderboard by commodity. Columns as in Table 1; shading is scaled within each column across both commodities.

				Responds to price				Invariant to irrelevant cues			
Rank	Model	Good	**Overall**	No boundary	Turbu-lence	Turbulent fraction	*Mean*	Para-phrase σ	Name range	Seller-profession	*Mean*
–	*Ideal step agent*		*0.0*	*0.0*	*0.0*	*0.0*	*0.0*	*0.0*	*0.0*	*0.0*	*0.0*
1	• Qwen2.5 32B	water	0.9	0.1	0.2	0.4	0.2	0.1	3.5	1.2	1.6
2	• Gemma 3 27B	water	1.8	0.0	0.5	0.5	0.3	0.7	3.6	5.7	3.3
3	• Llama 3.2 3B	water	3.0	1.7	1.5	1.9	1.7	1.9	6.8	4.5	4.4
4	• Qwen2.5 1.5B	water	5.9	25.4	1.1	0.9	9.1	2.6	5.3	0.0	2.6
5	• DeepSeek-V2-Lite	water	10.8	30.0	1.4	0.9	10.7	2.9	22.8	7.1	10.9
–	*Constant-decision agent*		*16.7*	*100.0*	*0.0*	*0.0*	*33.3*	*0.0*	*0.0*	*0.0*	*0.0*
6	• Qwen2.5 0.5B	water	20.6	100.0	0.0	0.0	33.3	0.0	0.0	23.7	7.9
7	• Llama 3.2 1B	water	23.5	99.1	4.0	6.2	36.4	7.5	0.9	23.5	10.6
8	• Llama 3.1 8B	water	26.6	82.2	1.8	1.8	28.6	1.9	54.5	17.6	24.7
–	*Ideal step agent*		*0.0*	*0.0*	*0.0*	*0.0*	*0.0*	*0.0*	*0.0*	*0.0*	*0.0*
1	• Gemma 3 27B	wheat	11.0	7.2	6.0	8.8	7.3	23.2	17.8	2.8	14.6
–	*Constant-decision agent*		*16.7*	*100.0*	*0.0*	*0.0*	*33.3*	*0.0*	*0.0*	*0.0*	*0.0*
2	• Qwen2.5 1.5B	wheat	16.7	100.0	0.0	0.0	33.3	0.0	0.0	0.0	0.0
3	• DeepSeek-V2-Lite	wheat	16.8	96.5	1.1	1.0	32.9	1.2	1.1	0.0	0.8
4	• Llama 3.2 1B	wheat	17.0	100.0	0.0	0.0	33.3	0.0	0.0	1.8	0.6
5	• Qwen2.5 32B	wheat	17.0	100.0	0.2	0.2	33.5	1.6	0.0	0.2	0.6
6	• Qwen2.5 0.5B	wheat	25.3	100.0	0.0	0.0	33.3	0.0	24.7	27.0	17.2
7	• Llama 3.1 8B	wheat	28.1	99.2	4.6	4.3	36.0	8.1	28.9	23.4	20.2
8	• Llama 3.2 3B	wheat	36.7	95.8	8.0	9.2	37.7	27.8	61.9	17.4	35.7

E Additional paraphrasing results

The full 16-row Experiment 1 output CSV `phase1_noise_floor.csv` contains additional columns (buy-price quantiles p_5^{buy}, p_{25}^{buy}, p_{50}^{buy}, p_{75}^{buy}, p_{95}^{buy} mean/std, `auc_std`, `refusal_rate_std`, `decision_band_fraction_std`) not shown in Table 7. Table 8 extracts the buy-price quantile fingerprint per curve, computed wherever the model chooses to buy at least once, including constant and near-constant curves; these quantiles describe the distribution of buy prices and are not the decision boundary of the main analysis.

F Quantity-sensitivity details

Table 9 lists the decision boundary for every model at each of the three comparable quantities per good; Figures 5 and 6 overlay the curves on a common price-per-unit axis.

G Decision-pattern threshold sensitivity

A share-of-grid rule labels a curve near-always buy when its raw yes fraction exceeds 0.95 and near-always decline when it falls below 0.05 (compared with the step test in Appendix H). Table 10 recomputes Experiment 1 counts under that rule at tighter and looser tolerances. The finding that a

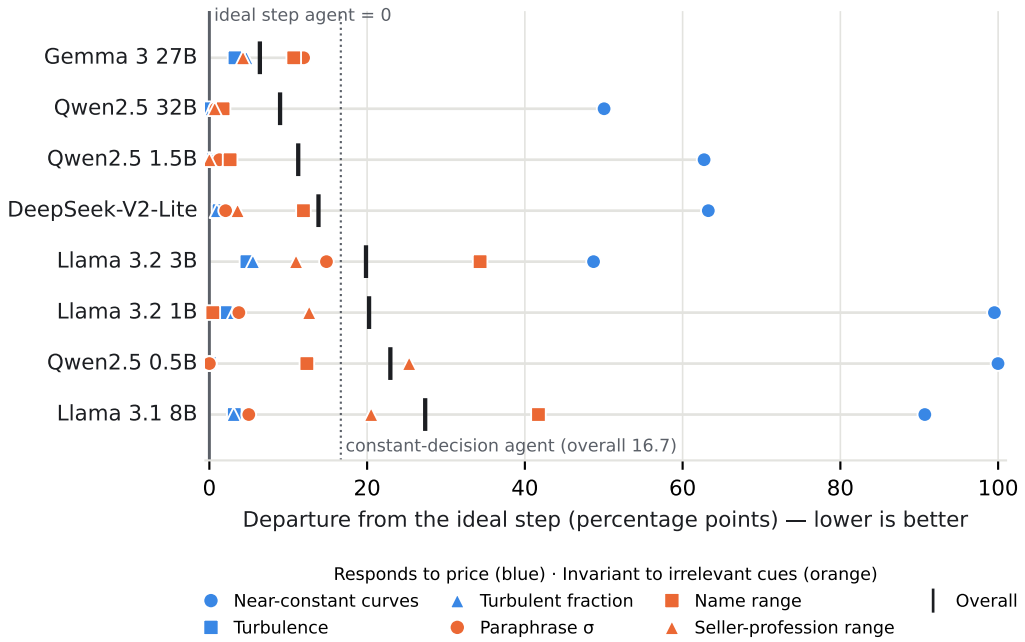


Figure 4: Every leaderboard axis for every model (commodity mean, percentage points; lower is better). Blue markers: the responds-to-price block; orange markers: the invariant-to-irrelevant-cues block; the black tick is the overall score; the dotted line marks the constant-decision agent.

Table 7: Experiment 1 decision boundaries across four semantically equivalent phrasings. A phrasing is *reported* when its curve passes the step test and the logistic center $\hat{t}$ ($) lies inside the tested price range; the logistic width $\hat{s}$ is shown in small text (band $\hat{t}\,e^{\pm\hat{s}}$). Otherwise the category is shown: CONST (answer never changes), NC-BUY / NC-DEC (near-constant buy / decline: the best step explains less than half of the minority decisions), ABOVE / BELOW (crossing outside the tested range). Average demand is the refusal-excluded buy share over tested prices, averaged over phrasings.

Model	Good	Reported (of 4)	Average demand	Phrasing 0	Phrasing 1	Phrasing 2	Phrasing 3
qwen25-0.5b	water	0	1.000	CONST	CONST	CONST	CONST
qwen25-0.5b	wheat	0	1.000	CONST	CONST	CONST	CONST
qwen25-1.5b	water	4	0.035	1.24 s=0.32	0.86 s=0.34	0.70 s=0.28	3.82 s=0.14
qwen25-1.5b	wheat	0	0.000	CONST	CONST	CONST	CONST
qwen25-32b	water	4	0.098	4.95 s=0.02	4.95 s=0.02	4.98 s=0.02	5.06 s=0.03
qwen25-32b	wheat	1	0.009	CONST	CONST	10.01 s=0.31	NC-DEC
deepseek-v2-lite	water	4	0.094	5.71 s=0.09	4.68 s=0.10	5.89 s=0.11	2.54 s=0.14
deepseek-v2-lite	wheat	2	0.020	NC-DEC	NC-DEC	9.87 s=0.28	6.66 s=0.35
gemma3-27b	water	4	0.085	4.26 s=0.04	4.31 s=0.05	4.85 s=0.03	3.91 s=0.03
gemma3-27b	wheat	2	0.391	NC-BUY	NC-DEC	37.29 s=0.26	100.28 s=0.35
llama32-1b	water	0	0.083	NC-DEC	NC-DEC	NC-DEC	NC-DEC
llama32-1b	wheat	0	0.000	CONST	CONST	CONST	CONST
llama32-3b	water	4	0.134	6.15 s=0.09	6.09 s=0.08	6.36 s=0.09	8.20 s=0.12
llama32-3b	wheat	1	0.729	78.18 s=0.50	NC-BUY	NC-BUY	NC-BUY
llama31-8b	water	4	0.234	12.14 s=0.05	10.32 s=0.09	12.85 s=0.04	11.48 s=0.06
llama31-8b	wheat	0	0.936	NC-BUY	NC-BUY	NC-BUY	NC-BUY

Table 8: Mean buy-price quantiles across the four phrasings, computed from prices where the classifier returned yes.

Model	Commodity	p_5^{buy}	p_{25}^{buy}	p_{50}^{buy}	p_{75}^{buy}	p_{95}^{buy}
qwen25-0.5b	water	2.59	12.57	25.05	37.53	47.51
qwen25-0.5b	wheat	15.95	75.75	150.50	225.25	285.05
qwen25-1.5b	water	0.16	0.45	0.81	1.88	2.83
qwen25-1.5b	wheat	-	-	-	-	-
deepseek-v2-lite	water	0.38	1.52	2.95	4.38	5.88
deepseek-v2-lite	wheat	1.50	2.47	3.67	5.13	8.15
llama32-1b	water	0.60	3.06	6.02	13.14	25.45
llama32-1b	wheat	-	-	-	-	-
llama32-3b	water	0.40	1.63	3.18	4.77	6.27
llama32-3b	wheat	6.10	26.75	55.53	140.08	250.75
llama31-8b	water	0.70	3.12	6.15	9.18	11.94
llama31-8b	wheat	15.67	74.36	148.02	222.01	282.93
gemma3-27b	water	0.30	1.14	2.19	3.23	4.12
gemma3-27b	wheat	11.34	52.75	110.69	176.65	264.91

Table 9: Decision boundary $\hat{t}$ ($, total price; logistic center under the step-test gate) per quantity. Letters mark curves without a reported boundary: C constant, B near-constant buy, D near-constant decline, A crossing above and L crossing below (lower than) the tested range. The last two rows count, for each adjacent pair of quantities, the models with a boundary at both quantities (comparable) and, among those, the ones whose boundary rises with quantity (increases); overall 11 of 11 comparable pairs increase.

	Water			Wheat		
Model	0.25 L	0.5 L	5 L	0.1 kg	1 kg	1,000 kg
qwen25-0.5b	C	C	C	D	C	C
qwen25-1.5b	D	1.24	D	C	C	C
qwen25-32b	3.26	4.95	30.16	C	C	C
deepseek-v2-lite	1.25	5.71	112.21	D	D	C
gemma3-27b	1.11	4.26	73.56	D	B	C
llama32-1b	D	D	D	D	C	C
llama32-3b	2.03	6.15	45.48	2.39	78.18	D
llama31-8b	5.42	12.14	146.26	D	B	B
Comparable with previous column	–	5	5	–	1	0
Increases	–	5	5	–	1	0

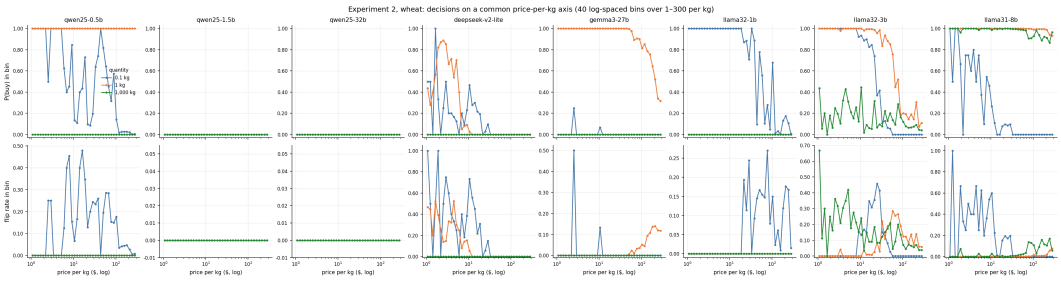


Figure 5: Experiment 2, wheat: the three quantities that share a per-unit support (0.1 kg, 1 kg, 1,000 kg; $1-$300 per kg) overlaid per model. Top: buy share in each of 40 log-spaced unit-price bins. Bottom: flip rate (adjacent buy/decline disagreements over valid adjacent pairs) in the same bins. Y axes use two-decimal ticks so sparse buys remain visible; the flip-rate axis is scaled per panel.

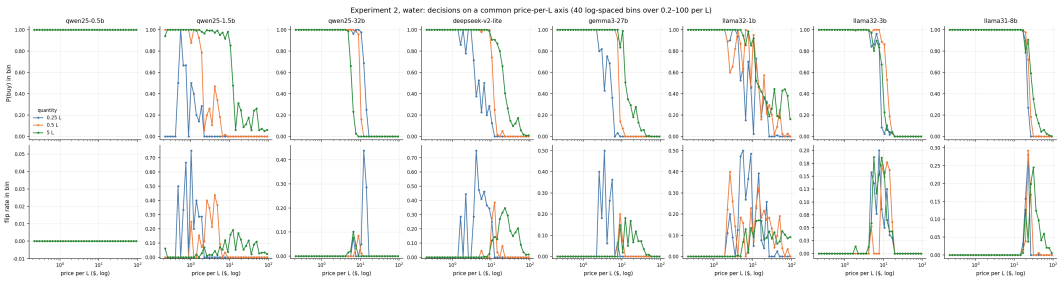


Figure 6: Experiment 2, water: as Figure 5 for 0.25 L, 0.5 L, and 5 L on the common \$0.20-\$100 per liter support.

substantial share of curves are near-constant is stable, although its exact size depends on the tolerance, ranging from 21 of 64 curves at 1% to 46 of 64 at 10%.

Table 10: Experiment 1 decision-pattern counts under alternative near-constant tolerances.

Tolerance	Near-always buy	Near-always decline	Buy and decline
1%	9	12	43
5%	13	20	31
10%	13	33	18

H Comparison with the share-of-grid rule and the 95th-percentile buy price

A share-of-grid rule labels a curve near-always buy when more than 95% of its decisions are buy and near-always decline when fewer than 5% are, and summarizes every other curve by the 95th percentile of its buy prices, p_{95}^{buy}. That rule is a statement about the share of the grid rather than about price: on the linear grid, a curve that buys only below about \$2.60 for 500 ml of water (the lowest 5% of the \$0.10-\$50 grid) or below \$16 per kilogram of wheat is labeled near-always decline even when it is a clean step. Table 11 cross-tabulates that rule against the labels of Section 3.2 for every curve. In Experiment 1, six curves that the share rule called near-always decline are clean steps near the bottom of the grid and now report a boundary (three qwen25-1.5b water phrasings, two deepseek-v2-lite wheat phrasings, and one qwen25-32b wheat phrasing), while seven curves with both decisions under the share rule are near-constant under the step test because their minority decisions are scattered. The same pattern holds for the profession-pair and name-pair curves. Where both summaries are defined, $\hat{t}$ lies below p_{95}^{buy} by a median factor of 0.84, and p_{95}^{buy} exceeds twice $\hat{t}$ on 18% of the curves; the difference is largest on near-constant-buy curves, where p_{95}^{buy} sits at the top of the grid. Neither the gate nor the estimator changes average demand, turbulence, or turbulent fraction.

I Temperature ablation

All results in the main text use greedy decoding. To test whether the decision patterns depend on that choice, the Experiment 1 base phrasing was re-run (literal Buyer and Seller, water and wheat) for all eight models with sampling at $T \in \{0.3, 0.7, 1.0\}$ and three seeds per temperature (144 sampled sweeps; one output token per prompt, one random seed per request) and compared with the greedy run ($T = 0$). Figures 7 and 8 show the smoothed curves, Figure 9 the metrics against temperature, and Table 12 the greedy and $T = 1$ values.

Three results follow. First, seed-to-seed variation is negligible relative to the effects reported in the paper: the three seeds never disagree on whether a boundary is reported (0 of 48 model-commodity-temperature combinations; in four combinations they differ only between two no-boundary categories, where the buy share sits near one half or a single buy decision separates constant from

Table 11: An alternative share-of-grid rule (rows) against the labels of the main analysis (columns). Boundary: passes the step test with the logistic crossing inside the tested range; Near-constant: the best step explains under half of the minority decisions; Out of range: crossing outside the tested range. Experiment 3 is counted per profession pair and Experiment 4 per available name pair.

Experiment	5% rule label	Constant	Near-constant	Out of range	Boundary	Total
1 (64 paraphrase curves)	Both decisions	0	7	0	24	31
	Near-always buy	8	5	0	0	13
	Near-always decline	10	4	0	6	20
3 (5,184 profession-pair curves)	Both decisions	0	612	0	1,255	1,867
	Near-always buy	317	271	0	0	588
	Near-always decline	1,748	715	2	264	2,729
4 (56,704 name-pair curves)	Both decisions	0	6,779	0	16,755	23,534
	Near-always buy	6,302	2,059	0	2	8,363
	Near-always decline	14,559	7,173	26	3,049	24,807

near-constant), the SD of turbulence across seeds is at most 0.018, and the SD of $\hat{t}$ across seeds has median \$0.02 and maximum \$0.08, below the SD across the four paraphrases in all nine combinations where every seed reports a boundary. Second, sampling raises turbulence and removes boundaries for six of the eight models: turbulence at $T = 1$ reaches 0.2 or more in 12 of 16 panels and about 0.5 for llama32-1b on water (0.502) and llama32-3b on wheat (0.493), and the number of panels without a reportable boundary rises from 9 under greedy decoding to 14 at $T = 1$. In the five panels that lose their boundary (qwen25-1.5b, deepseek-v2-lite, llama32-3b, and llama31-8b on water; llama32-3b on wheat), four of which were clean steps under greedy decoding, scattered decisions accumulate until $P(\text{buy} < \text{decline})$ falls to between 0.54 and 0.83 and the best step explains less than half of the minority decisions, so the step test withholds the boundary rather than reporting a price pinned to the grid. At $T = 0.3$ the boundaries that survive lie within \$0.55 of the greedy value (deepseek-v2-lite on water \$6.23 against \$5.71; llama31-8b on water \$12.24 against \$12.14). Refusals also rise with temperature for deepseek-v2-lite (8.3% of wheat responses at $T = 1$). Third, the two models at the top of the leaderboard are unaffected: on water, qwen25-32b and gemma3-27b keep turbulence at or below 0.01, $P(\text{buy} < \text{decline}) = 1.00$, and $\hat{t}$ within \$0.04 and \$0.06 of the greedy value at every temperature, and their wheat panels keep their labels at every temperature. The greedy results therefore understate turbulence for most models, and the position of the two most coherent models does not depend on the decoding choice.

Table 12: Temperature ablation on the base phrasing: greedy decoding ($T = 0$) against $T = 1$ (mean over three seeds; SD of $\hat{t}$ across seeds in parentheses). Step AUC is $P(\text{buy price} < \text{decline price})$; refusals are counts per sweep (4,991 water prompts, 29,901 wheat prompts). Where no boundary is reported the category is given; * marks a combination whose seeds disagree on the category.

		Turbulence		Step AUC		Boundary t (\$)		Refusals	
Model	Commodity	T=0	T=1	T=0	T=1	T=0	T=1 (sd)	T=0	T=1
qwen25-0.5b	water	0.000	0.327	–	0.502	– (CONST)	– (NC-BUY)	0	1
qwen25-0.5b	wheat	0.000	0.463	–	0.519	– (CONST)	– (NC-BUY)	0	62
qwen25-1.5b	water	0.013	0.383	0.993	0.601	1.24	– (NC-DEC)	0	0
qwen25-1.5b	wheat	0.000	0.372	–	0.466	– (CONST)	– (NC-DEC)	0	4
qwen25-32b	water	0.002	0.002	1.000	1.000	4.95	4.91 (0.02)	0	0
qwen25-32b	wheat	0.000	0.002	–	0.964	– (CONST)	– (NC-DEC)	0	0
deepseek-v2-lite	water	0.015	0.220	0.998	0.746	5.71	– (NC-DEC)	0	681
deepseek-v2-lite	wheat	0.007	0.350	0.996	0.574	– (NC-DEC)	– (NC-DEC)	0	2492
gemma3-27b	water	0.005	0.006	1.000	1.000	4.26	4.31 (0.00)	0	0
gemma3-27b	wheat	0.079	0.120	0.836	0.822	– (NC-BUY)	– (NC-BUY)	0	0
llama32-1b	water	0.087	0.502	0.897	0.536	– (NC-DEC)	– (NC-DEC)*	0	0
llama32-1b	wheat	0.000	0.386	–	0.534	– (CONST)	– (NC-DEC)	0	3
llama32-3b	water	0.014	0.222	0.998	0.834	6.15	– (NC-DEC)	0	4
llama32-3b	wheat	0.114	0.493	0.826	0.539	78.18	– (NC-DEC)	0	134
llama31-8b	water	0.019	0.368	0.999	0.754	12.14	– (NC-DEC)	0	2
llama31-8b	wheat	0.022	0.470	0.851	0.544	– (NC-BUY)	– (NC-BUY)	0	4

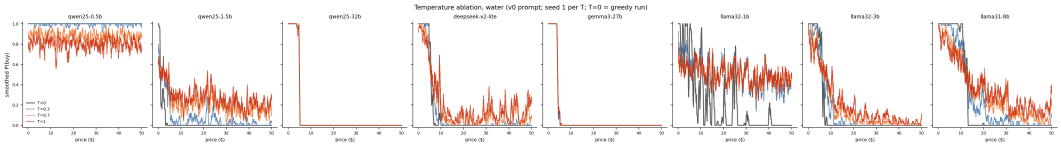


Figure 7: Water, base phrasing: smoothed P(buy) for the greedy run and one seed at each temperature.

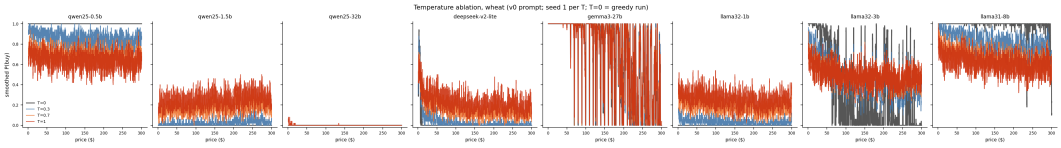


Figure 8: Wheat, base phrasing: as Figure 7.

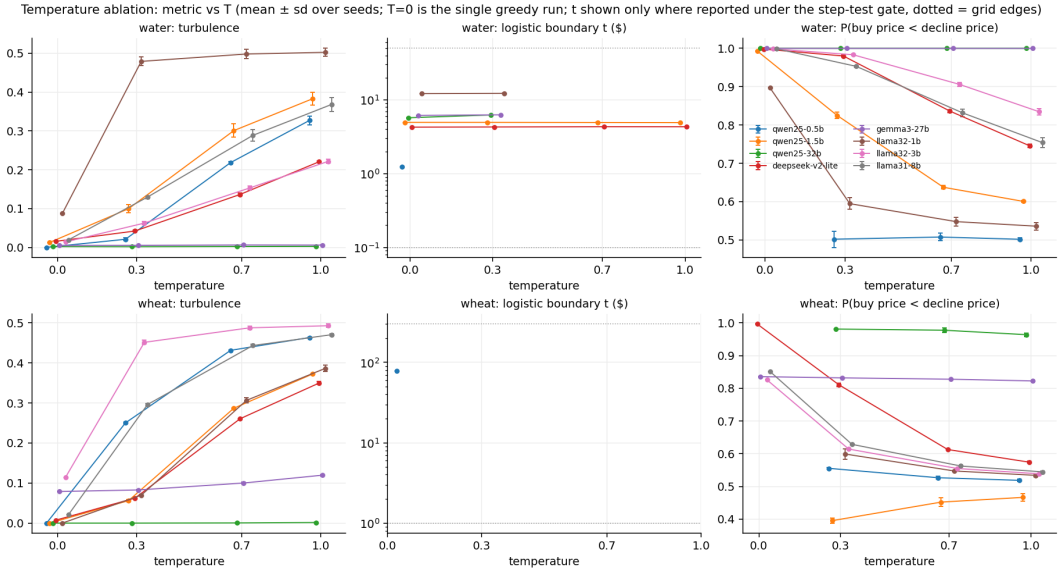


Figure 9: Turbulence, the decision boundary $\hat{t}$, and $P(\text{buy price} < \text{decline price})$ against temperature (mean $\pm$ SD over three seeds; $T = 0$ is the single greedy run). Points are absent where no boundary is reported.

J Cross-model alignment of price extrema

Non-monotonic price patterns show limited overlap across models. Only a minority of one model's extrema recur at the same price in another, and the shared extrema do not group cleanly by tokenization regime. The catalog contains 8,629 extrema across all decision curves (2 commodities $\times$ 8 versions $\times$ 8 models). Appendix Figure 10 reports the pairwise alignment matrix: each row value is the fraction of that model's extrema that have a partner within $\pm 2\%$ of the same price in the column model's catalog.

Pooling ordered pairs within each block (Table 16), mean alignment is 0.08 within the per-digit block, 0.25 within the chunked block, and 0.10-0.20 across blocks. gemma3-27b uses per-digit price tokenization but is US-origin; when it queries, it aligns more with the chunked Llama block (0.17) than with the other per-digit models (0.08), but when queried the two are similar (0.12 against 0.13), so it does not separate a developer-origin grouping from a tokenization grouping. Against a location-preserving random baseline that keeps each curve's extremum count and price grid but places the extrema uniformly (Appendix T), observed alignment falls *below* the null in every block

(block means of observed minus null between -0.03 and -0.17; per-pair values reach -0.69), and only two of the 56 ordered model pairs exceed their null at $p < 0.05$, both of them cross-block pairs. Extrema therefore recur across models no more often than chance placement on the same grid would produce.

Peak and trough counts, the per-curve summary, and the marked decision curves for every model are in Appendices K-L.

K PER-CURVE EXTREMUM SUMMARY

The full 128-row per-curve summary (16 curves $\times$ 8 versions) is in `phase2_extrema_summary.csv`. It records decision pattern, average demand over the tested prices, the decision boundary and its width, turbulence, turbulent fraction, and peak/trough counts. Per-curve stem-and-leaf visualizations of the peak, trough, and refusal-price locations are under `phase2_extrema_stem_leaf/`.

Table 13: Extremum rollup by model and commodity, sorted by mean turbulence descending (top 7 of 16 shown). Turbulence is the fraction of adjacent-price buy/decline pairs that disagree; turbulent fraction is the fraction of the price grid with $0.3 < m < 0.7$.

Model	Commodity	σ_{value} (Exp. 1)	turb. mean	turb. max	turb. frac.	#peaks	#troughs
llama32-3b	wheat	0.278	0.059	0.137	0.061	488	1,428
llama32-1b	water	0.075	0.055	0.093	0.084	367	1,143
llama31-8b	wheat	0.081	0.037	0.125	0.031	20	871
gemma3-27b	wheat	0.232	0.035	0.094	0.054	386	314
deepseek-v2-lite	water	0.029	0.033	0.117	0.032	436	470
qwen25-1.5b	water	0.026	0.029	0.088	0.033	359	208
llama32-3b	water	0.019	0.027	0.085	0.035	253	719

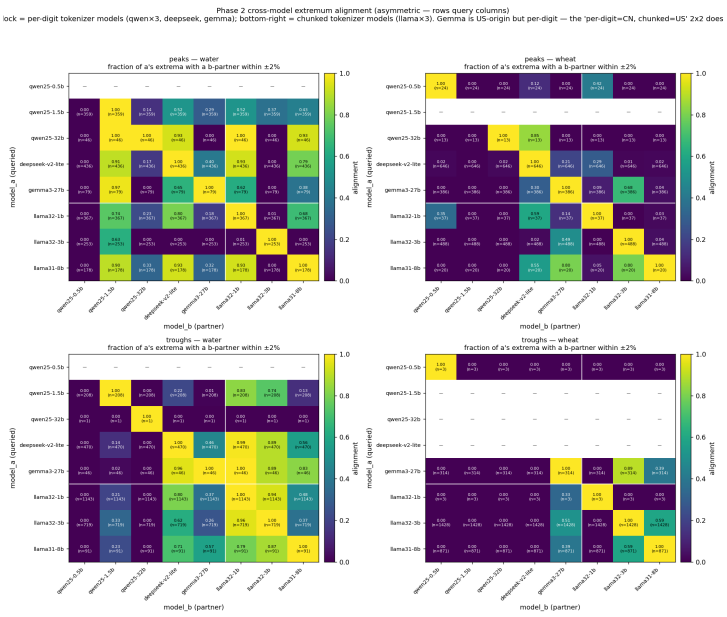


Figure 10: Pairwise cross-model alignment of price extrema. Rows query columns. White separators demarcate the per-digit tokenizer block (top-left, 5×5) from the chunked tokenizer block (bottom-right, 3×3). gemma3-27b sits in the per-digit block despite being US-origin.

L DECISION CURVES WITH PRICE EXTREMA

Figure 11 shows one model, llama32-3b on wheat, as an example.

For each of the eight models, both commodities are rendered as an eight-panel figure showing v0-v7 with red ▽ peaks and blue △ troughs overlaid on the smoothed curve. Files: `curves/phase2/<model>_p{0,1}.png`.

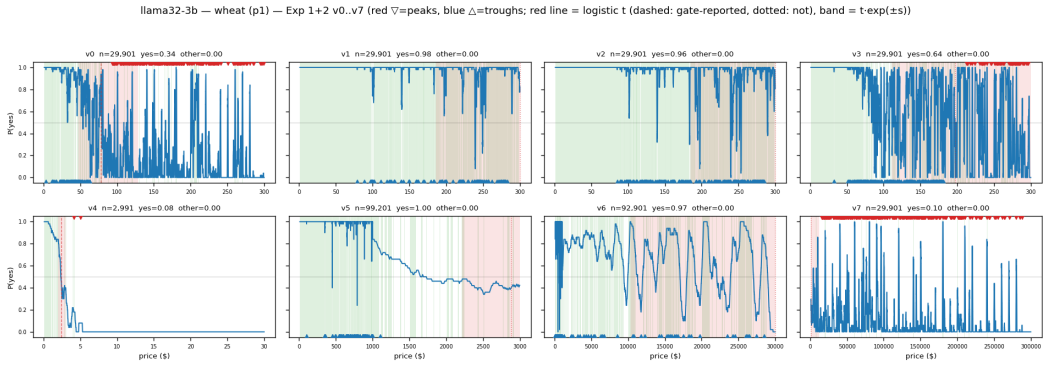


Figure 11: llama32-3b wheat: eight generic-role prompt versions (four paraphrases and four quantity variants) with peak/trough markers overlaid. The red line is the logistic center with its band $\hat{t}e^{\pm\hat{s}}$, dashed when the gate reports it and dotted otherwise; constant versions have no line.

M PROFESSION-CUE REGRESSION AND ASYMMETRY TABLES

Table 14 reports the self-pair-corrected in-group comparison for both identity cues. For each model-commodity panel, cells whose buyer and seller share a category (TEER tier in Experiment 3, region in Experiment 4) are compared with cells whose buyer and seller differ, using a Welch t-test on the per-cell value; cells in which buyer and seller are the identical profession or the identical name are removed from the same-category group, so a degenerate self-pair cannot drive the comparison. The boundary is compared only in panels with at least two same-category and two different-category cells that report a median boundary, and average demand in all 16 panels. The full marginal and interaction regressions and the uncorrected variants, for every target (p_{95}^{buy}, average demand, turbulence, and turbulent fraction), are in the released files `phase3a_exp3_regressions.csv` and `phase3b_exp4_regressions.csv`.

N PROFESSION-CUE HEATMAPS

Figure 13 shows the profession-pair boundary heatmaps.

The released outputs contain separate heatmaps for the decision boundary, turbulence, refusal rate, average demand over the tested prices, and turbulent fraction. Only the boundary heatmap leaves cells blank; the other summaries are defined for every cell.

O NAME-CUE REGRESSION AND ASYMMETRY TABLES

The self-pair-corrected region-level comparison is in Table 14 (lower half). The marginal and interaction regressions on buyer and seller region, and the uncorrected variants, are in the released file `phase3b_exp4_regressions.csv`; the structure mirrors the profession-cue tables but is keyed on region.

P ADDITIONAL NAME-CUE HEATMAPS

Figures 18 and 19 show the individual-name heatmaps and Figure 17 the gender-only view.

Table 14: Self-pair-corrected in-group comparison. Same: mean over cells whose buyer and seller share a TEER tier (Experiment 3) or region (Experiment 4), excluding identical-profession and identical-name pairs; Different: mean over cells whose buyer and seller differ; Δ = Same $-$ Different; p from a Welch t-test (bold: $p < 0.05$). Boundary $\hat{t}$ in dollars, each cell the median over at least three curves with a boundary, only panels with at least two cells on each side; average demand as a share of tested prices, over all cells.

Model	Commodity	n_{same}	n_{diff}	Same	Different	Δ	p
Experiment 3, professions: decision boundary $\hat{t}$ ($), same TEER vs. different TEER							
deepseek-v2-lite	water	3	23	5.61	4.53	+1.07	**0.032**
gemma3-27b	water	6	30	4.79	4.44	+0.35	0.594
llama31-8b	water	6	30	13.60	14.09	-0.50	0.785
llama32-3b	water	6	30	5.80	5.85	-0.05	0.956
qwen25-32b	water	5	30	2.75	2.77	-0.02	0.972
Experiment 4, names: decision boundary $\hat{t}$ ($), same region vs. different region							
deepseek-v2-lite	water	6	30	5.15	4.72	+0.44	0.071
deepseek-v2-lite	wheat	2	17	5.13	5.22	-0.08	0.938
gemma3-27b	water	6	30	4.72	4.50	+0.22	0.451
gemma3-27b	wheat	6	30	29.12	31.55	-2.43	0.519
llama31-8b	water	3	19	20.86	19.33	+1.52	0.875
llama31-8b	wheat	2	3	120.22	112.88	+7.34	0.567
llama32-3b	water	6	30	8.69	8.33	+0.35	0.269
llama32-3b	wheat	2	15	60.18	56.65	+3.53	0.728
qwen25-1.5b	water	6	30	0.88	0.89	-0.01	0.605
qwen25-32b	water	6	30	6.60	6.62	-0.02	0.899
Experiment 3, professions: average demand, same TEER vs. different TEER							
qwen25-0.5b	water	6	30	0.910	0.865	+0.045	0.529
qwen25-0.5b	wheat	6	30	0.461	0.485	-0.024	0.808
qwen25-1.5b	water	6	30	0.000	0.000	-0.000	0.317
qwen25-1.5b	wheat	6	30	0.000	0.000	+0.000	–
qwen25-32b	water	6	30	0.044	0.044	-0.000	0.968
qwen25-32b	wheat	6	30	0.001	0.001	-0.000	0.901
deepseek-v2-lite	water	6	30	0.096	0.073	+0.023	0.358
deepseek-v2-lite	wheat	6	30	0.000	0.000	-0.000	0.438
gemma3-27b	water	6	30	0.098	0.096	+0.002	0.863
gemma3-27b	wheat	6	30	0.032	0.029	+0.003	0.879
llama32-1b	water	6	30	0.367	0.373	-0.006	0.961
llama32-1b	wheat	6	30	0.004	0.022	-0.018	0.074
llama32-3b	water	6	30	0.115	0.119	-0.003	0.832
llama32-3b	wheat	6	30	0.883	0.878	+0.005	0.947
llama31-8b	water	6	30	0.281	0.310	-0.029	0.397
llama31-8b	wheat	6	30	0.145	0.140	+0.006	0.948
Experiment 4, names: average demand, same region vs. different region							
qwen25-0.5b	water	6	30	1.000	1.000	-0.000	0.389
qwen25-0.5b	wheat	6	30	0.940	0.953	-0.013	0.688
qwen25-1.5b	water	6	30	0.014	0.017	-0.003	0.146
qwen25-1.5b	wheat	6	30	0.000	0.000	+0.000	0.317
qwen25-32b	water	6	30	0.132	0.133	-0.001	0.808
qwen25-32b	wheat	6	30	0.000	0.000	+0.000	–
deepseek-v2-lite	water	6	30	0.151	0.134	+0.018	0.414
deepseek-v2-lite	wheat	6	30	0.004	0.003	+0.001	0.547
gemma3-27b	water	6	30	0.098	0.094	+0.004	0.534
gemma3-27b	wheat	6	30	0.148	0.152	-0.005	0.855
llama32-1b	water	6	30	0.002	0.000	+0.001	0.405
llama32-1b	wheat	6	30	0.000	0.000	+0.000	0.452
llama32-3b	water	6	30	0.179	0.174	+0.006	0.461
llama32-3b	wheat	6	30	0.417	0.444	-0.027	0.612
llama31-8b	water	6	30	0.896	0.835	+0.061	0.170
llama31-8b	wheat	6	30	0.109	0.070	+0.039	0.479

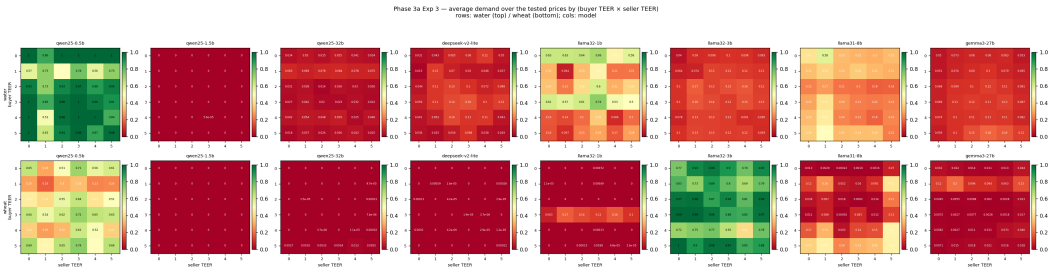


Figure 12: Average demand over the tested prices per (buyer TEER, seller TEER) cell. Rows: water (top), wheat (bottom). Columns: model. The metric is shown for every cell, including constant and near-constant cells.

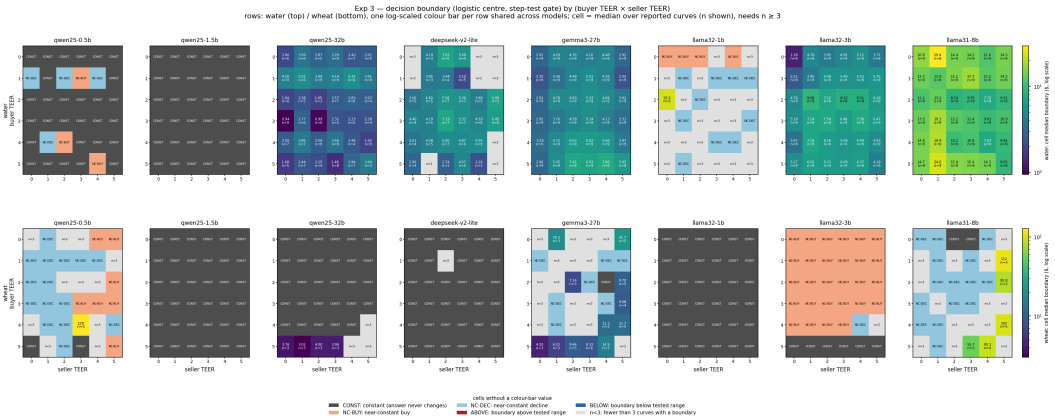


Figure 13: Decision boundary $\hat{t}$ per (buyer TEER, seller TEER) cell: the median over the cell's curves with a boundary, shown when at least three exist (n printed). Rows: water (top), wheat (bottom); columns: model; one log-scaled color bar per commodity row, shared across models. Cells without a boundary carry discrete colors: constant, near-constant buy, near-constant decline, or fewer than three curves.

The released outputs contain 6×6 region-pair heatmaps for the decision boundary, turbulence, refusal rate, average demand over the tested prices, and turbulent fraction. The full 60×60 individual-name heatmaps are placed here because their labels require full-page display or digital zoom. The main text instead reports the distribution across name pairs. Raw within-panel maximum-to-minimum boundary ratios reach 237 for deepseek-v2-lite on water; with thousands of pairs in a panel, these extremes are controlled by individual pairs, which is why the main analysis reports the 5th-95th-percentile spread.

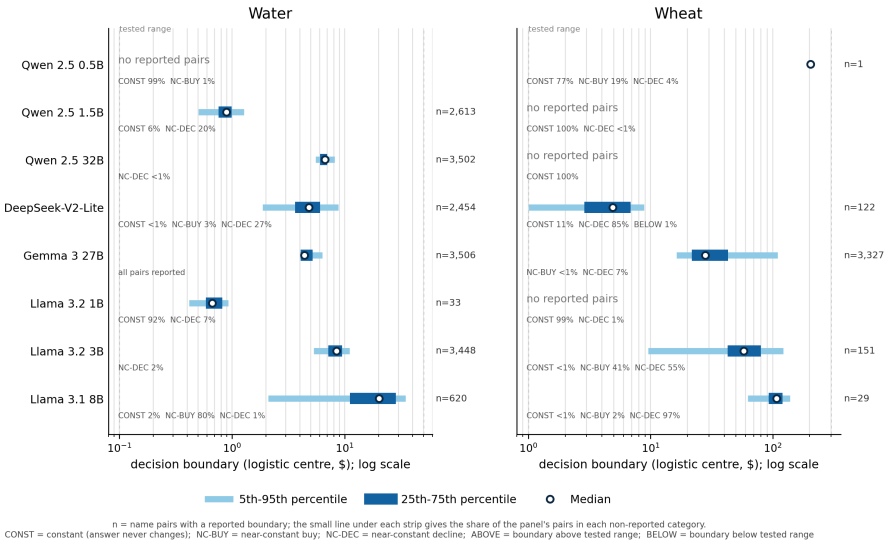


Figure 14: Distribution of the decision boundary $\hat{t}$ across name-pair curves that report one. Light intervals show the 5th-95th percentiles, dark intervals the 25th-75th percentiles, and circles the median; n is the number of reporting pairs, and the line under each strip gives the share of the panel's pairs in each non-reporting category. Panels with no reporting pair are marked explicitly.

Q IDENTITY CUES: PROFESSION AND NAME DETAIL

The following material gives the cell-level detail behind Section 5.5.

Naming the parties Per-model detail behind Section 5.4: llama31-8b buys water at 24% of the tested prices with generic roles and the median named pair at 94%, and buys wheat at 98% against a median pair at 3%, with 99% of pairs outside the phrasing range in both cases; qwen25-32b on water moves from 10% to a median of 13% with 99% of pairs outside a phrasing range only 0.3 points wide; llama32-1b on water moves from 20% to 0%; llama32-3b on water from 12% to 17% with two thirds of pairs outside; and for deepseek-v2-lite on water the median pair equals the generic curve while 69% of pairs lie outside the phrasing range on one side or the other.

Profession cues Profession cues can substantially change decisions about otherwise identical trades. Across the 16 model-commodity panels, the median within-panel range in average demand over the tested prices is 16.9 percentage points across TEER pairs. The largest range is 67.4 points for llama32-1b on water, from 0.064 to 0.738.

Aggregating the three occupations in each of six TEER categories produces a 6×6 buyer-seller grid for each model and commodity, or 576 cells. Of the 5,184 profession-pair curves, 1,519 report a boundary; 2,065 are constant, 1,598 near-constant, and 2 cross outside the tested range. With at least three such curves required per cell, 194 of the 576 cells report a median boundary; 212 are constant, 94 near-constant, and 76 contain only one or two curves with a boundary. Average demand over the tested prices remains defined in all 576 cells. Appendix Figure 12 reports this all-cell outcome.

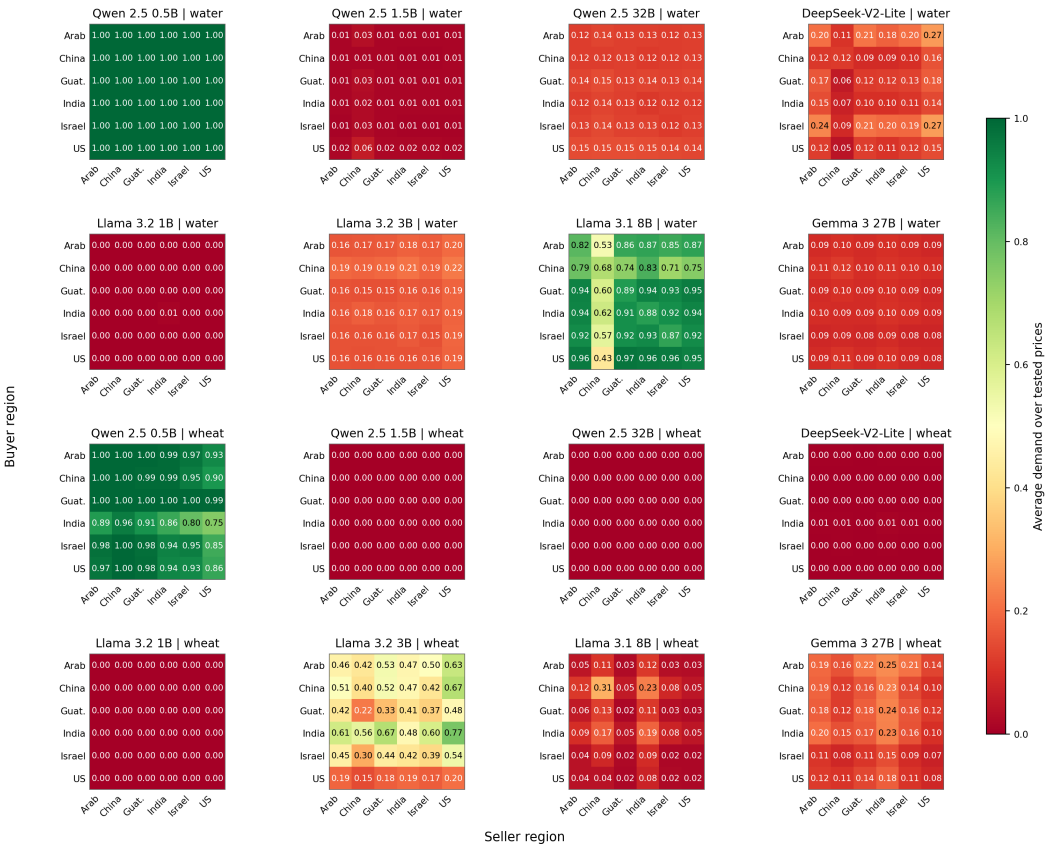


Figure 15: Average demand over the tested prices per (buyer region, seller region) cell. Water occupies the first two rows and wheat the last two. Every available name pair contributes to its region cell, and all 576 cells are shown on a shared color scale.

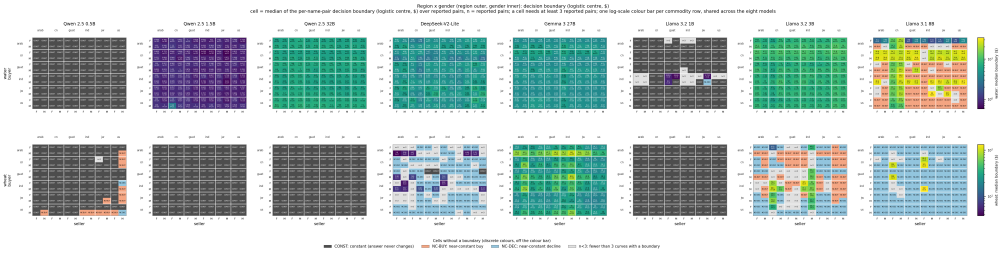


Figure 16: Region × gender decision boundary per (model, commodity), region outer and gender inner: the median over the cell's curves with a boundary, shown when at least three exist. One log-scaled color bar per commodity row; cells without a boundary carry the discrete category colors.

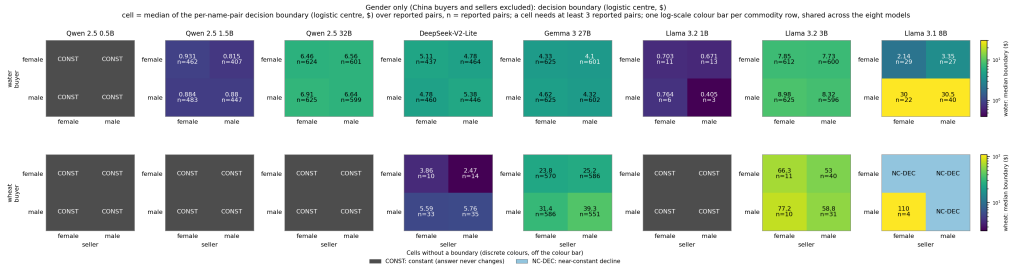


Figure 17: Gender-only decision boundary per (model, commodity), first name set, Chinese-associated names excluded because the survey validated their regional association but not their gender tags; median over the cell's curves with a boundary, at least three required; same color rules as the region $\times$ gender figure.

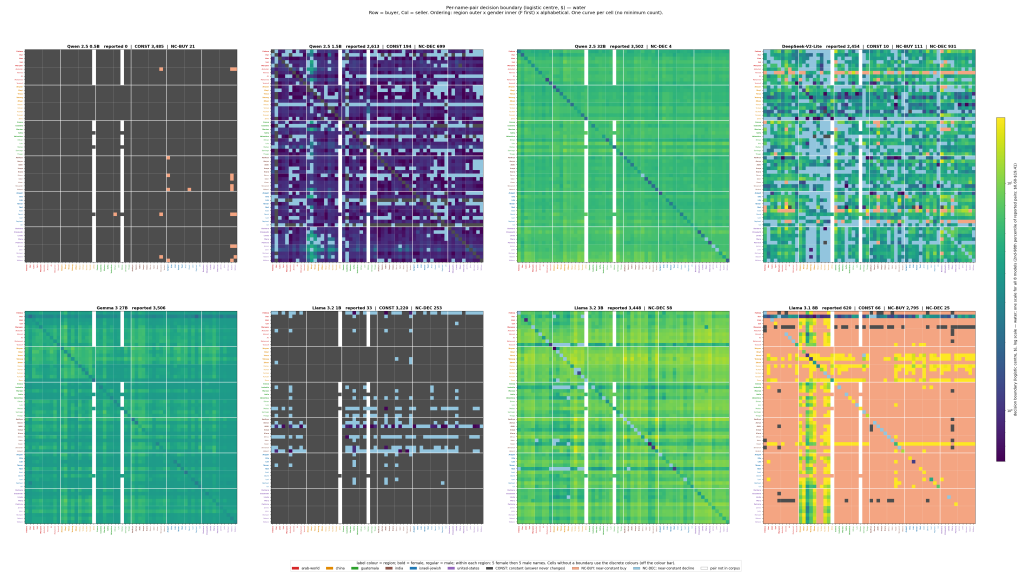


Figure 18: Per-name-pair decision boundary for water across eight models (one curve per cell, so no minimum applies). Names are ordered by region, then gender, then alphabetically; one log-scaled color bar shared across models; pairs without a boundary carry the discrete category colors, and white marks pairs absent from the corpus.



Figure 19: Per-name-pair decision boundary for wheat, as the water figure.

Reported cells occur in 10 of the 16 model-commodity panels, and four water panels (qwen25-32b, gemma3-27b, llama32-3b, and llama31-8b) report all 36 cells (35 for qwen25-32b in the self-pair-corrected comparison, which drops one diagonal cell with too few cross pairs). Across the eight panels with at least two reported cells, the median ratio between the highest and lowest cell boundary is 3.63, and seven panels have a greater than twofold range; the largest ratios arise in sparse panels whose cells rest on three to nine curves. The corresponding heatmaps are in Appendix N. The self-pair-corrected in-group comparison (buyer and seller in the same TEER tier, excluding identical professions, against different tiers; Appendix Table 14) finds a significant same-tier boundary difference in one of the five testable panels (deepseek-v2-lite on water, $+\$1.07$, $p = 0.032$, resting on three same-tier cells) and no significant same-tier difference in average demand in any of the 16 panels. This cell-level Welch test compares six cells with thirty and has little power; the pair-level test of Appendix R supersedes it.

Name cues Name cues also change decisions about otherwise identical trades. Across the 16 model-commodity panels, the median within-panel range in average demand over the tested prices is 4.4 percentage points across region pairs, and the largest is 61.9 points for llama32-3b on wheat. Across the 11 panels in which at least 20 name pairs report a boundary, the ratio between the 95th and 5th percentiles of $\hat{t}$ has a median of 2.5 and reaches 16.5 for llama31-8b on water.

The 60 names carry six regions and two genders, allowing comparisons at the individual-name, region, region×gender, and gender levels without stating a demographic label in the prompt. The retained corpus contains 56,704 available name-pair curves[1]. Aggregating them into a 6×6 (buyer region, seller region) grid for each model and commodity produces 576 cells: 286 report a median boundary from at least three curves, 207 are constant, 55 near-constant, and 28 contain only one or two curves with a boundary. Average demand over the tested prices is defined in all 576 cells. Appendix Figure 15 reports this all-cell outcome.

Across the 56,704 available name-pair curves, 19,806 report a boundary, 20,861 are constant, 16,011 are near-constant (10,877 decline, 5,134 buy), and 26 have a fitted crossing below the tested range. Name pairs with a boundary occur in 12 of the 16 model-commodity panels; the other four report none. With up to 3,600 pairs per panel, Appendix Figure 14 summarizes the 5th-95th-percentile spread rather than the two most extreme name pairs; nine panels exceed a twofold spread, and the two largest spreads (llama31-8b on water and llama32-3b on wheat) arise in panels where most pairs are near-constant buy and only the minority with a clean step report a boundary (Appendix B). Full

[1]Each model-commodity panel contains 3,506 (water) or 3,582 (wheat) of the 3,600 possible buyer-seller pairs; Appendix B documents the build irregularity behind the shortfall.

individual-name, region×gender, and gender-only heatmaps appear in Appendix P. The self-pair-corrected in-group comparison (buyer and seller from the same region, excluding identical names, against different regions; Appendix Table 14) finds no significant same-region difference in $\hat{t}$ in any of the ten testable panels, nor in average demand in any of the 16 panels. The same caveat applies; the pair-level test of Appendix R finds diagonal deviations of both signs.

R NAME-LEVEL PERMUTATION TESTS

All tests use the per-name mean of average demand (or the per-name median of the log boundary) as the unit, 2,000 permutations, seed 20260925, and exclude self-pairs; panels whose mean demand lies within 0.02 of 0 or 1 are saturated and cannot show an effect. Full tables are released with the code (`regional_bias/`, `dyadic/`, `intersectional/`).

Region and gender The spread of average demand across buyer regions is compared with the spread produced by reassigning the region labels among the names, ten per region. In the first name set the observed spread exceeds the 95th percentile of this null in 6 of the 9 non-saturated panels, and in 5 of 10 in the second set; the region labels explain a median 19% (first set) and 21% (second set) of the variance between buyer names in the non-saturated panels, measured as η^2 over per-name means, against about 9% expected for a random grouping of ten names into six or seven regions. The spread exceeds the SD of demand across the four phrasings in most panels but the phrasing max-min range in only about half. The ranking of the six shared regions (mapping in Appendix B) agrees between the sets with a median Spearman correlation of 0.71, and the lowest-demand buyer region agrees in 11 of 16 panels; it is the united-states-tagged names in 6 of 16 panels in the first set and the western-white-tagged names in 7 of 16 in the second. A buyer-gender gap exceeds the within-region null in 4 of 16 panels in every corpus (0.8 expected by chance), the same few models each time, most clearly gemma3-27b on wheat (male-tagged names $+9$ points) and deepseek-v2-lite on water (female-tagged names $+7$ points); a seller-gender gap in 5 panels, with female-tagged sellers receiving more purchases in every one.

Buyer-seller pairings A two-way fixed-effects fit removes each buyer name's and each seller name's own effect (median $R^2 = 0.82$); the residual is the dyadic component. Grouping the residuals by buyer region × seller region explains more than a random regrouping of the names in 27 of the 28 non-saturated panel-corpus combinations (nine or ten panels in each of three corpus variants: both name sets and the second back-filled), but a median 10% of a residual that is itself a fifth of the total variance, so region-structured pairing effects are one to two percent of the pair-level variance. The residuals sum to zero within each name, so the cells are dependent and the false-discovery correction treats them as separate tests. Fifty-seven cells are significant after that correction with the same sign in both name sets; recurring ones are an Indian-associated buyer with an Indian-associated seller ($+2$ to $+5$ points in four models, reversed in llama32-3b on wheat) and an Indian-associated buyer with a Western-associated seller (-1.5 to -2.5 points in four models). The same-region diagonal is significant in most panels but positive in some models and negative in others, so no uniform in-group premium exists. Gender × gender pairings are below one point and change sign across models.

Region × gender With five names per cell, an interaction test that keeps both main effects (gender labels permuted within region) rejects in 1-3 of 16 panels per corpus and side against 0.8 expected by chance, and 0-2 survive false-discovery correction. A power simulation that injects a shift into one cell's five names finds the smallest detectable single-cell shift at 80% power to be about 2.7 within-cell standard deviations, which corresponds to an interaction deviation of about 1.1; the largest observed deviations are 0.8-1.5, at the limit of what the design can detect. The two panels that reject in both sets rest on the ten Chinese-associated names, which are identical in the two sets. Intersectional effects are therefore neither established nor excluded by these data; a set with roughly thirty names per cell would be required.

Turbulence by identity The adjacent-price flip rate and the turbulent fraction of each name-pair curve were computed from the raw sweeps. Their means depend on the buyer's or seller's region in 32 of 60 tests and on gender in 25 of 60, but the dependence follows the demand level: names whose curves lie between always-buy and always-decline are the names whose curves flip. No

model separating the two was fitted, so identity-specific turbulence beyond the demand level is neither established nor excluded.

S REPLICATION WITH A SURVEY-CONFIRMED NAME SET

The second name set contains 70 names in seven regions (arab-world, china, french, indian, jewish, latin, western-white), five tagged female and five tagged male per region, assembled by asking members of each group which given names they associate with it. Twenty-six names are shared with the first set, and pairs among the shared names were not re-run, so each panel contains 4,224 of 4,900 possible pairs and the China×China cell is empty; Figure 21 shows the region-cell boundaries with those pairs back-filled from the first sweep, which used identical prompts. Table 15 repeats Table 2 for this set and Figure 20 places the two sets side by side: the six panels with at least 1,000 reporting pairs in both sets reproduce their median boundary within 15% and their 95th-to-5th spread within 30%, whereas the sparsely reporting panels do not (llama31-8b on water, with 620 and 558 reporting pairs, has a spread of 16.5 in the first set against 3.0 in the second), the share of pairs with a boundary is 35% in the first set and 36% in the second, and the region rankings agree as reported in Appendix R. Reducing the ambiguity of the names did not strengthen the region-level effect.

Table 15: Table 2 for the second name set (4,224 pairs per panel).

		Average demand		Named pairs		Boundary		Turbulence	
Model	Good	generic	named 5–95%	boundary	constant	median ($)	95/5	median	95th pct
qwen25-0.5b	water	1.00^{C}	1.00–1.00	0%	97%	–	–	0.000	0.000
qwen25-0.5b	wheat	1.00^{C}	0.21–1.00	1%	74%	9.47	11.59	0.000	0.132
qwen25-1.5b	water	0.03	0.00–0.04	84%	1%	0.95	3.20	0.003	0.021
qwen25-1.5b	wheat	0.00^{C}	0.00–0.00	0%	100%	–	–	0.000	0.000
qwen25-32b	water	0.10	0.10–0.16	100%	0%	6.39	1.51	0.005	0.011
qwen25-32b	wheat	0.00^{C}	0.00–0.00	0%	100%	–	–	0.000	0.000
deepseek-v2-lite	water	0.11	0.02–0.70	76%	0%	4.84	4.87	0.041	0.179
deepseek-v2-lite	wheat	$0.01^{\mathrm{NC-D}}$	0.00–0.02	5%	17%	5.92	10.40	0.001	0.013
gemma3-27b	water	0.08	0.08–0.14	100%	0%	4.50	1.66	0.011	0.021
gemma3-27b	wheat	$0.69^{\mathrm{NC-B}}$	0.05–0.38	93%	0%	26.73	7.18	0.031	0.102
llama32-1b	water	$0.20^{\mathrm{NC-D}}$	0.00–0.12	1%	74%	0.73	40.37	0.000	0.079
llama32-1b	wheat	0.00^{C}	0.00–0.00	0%	94%	–	–	0.000	0.000
llama32-3b	water	0.12	0.13–0.30	99%	0%	9.64	2.24	0.027	0.044
llama32-3b	wheat	0.34	0.05–0.99	3%	1%	48.03	4.64	0.080	0.134
llama31-8b	water	0.24	0.49–1.00	13%	3%	27.22	3.04	0.028	0.106
llama31-8b	wheat	$0.98^{\mathrm{NC-B}}$	0.00–0.34	4%	0%	88.09	2.36	0.037	0.141

T ALIGNMENT AGAINST A RANDOM-LOCATION BASELINE

The permutation null of the non-monotonic-choice analysis is applied to the coalesced extremum catalog: for every (model, commodity, phrasing version, kind) the observed number of extrema is preserved and the same number of locations is drawn uniformly without replacement from that curve's price grid; alignment is recomputed for 1,000 such draws. Table 16 reports block means and Figure 22 the per-pair matrix of observed minus null. Generated by `phase2_alignment_null.py`.

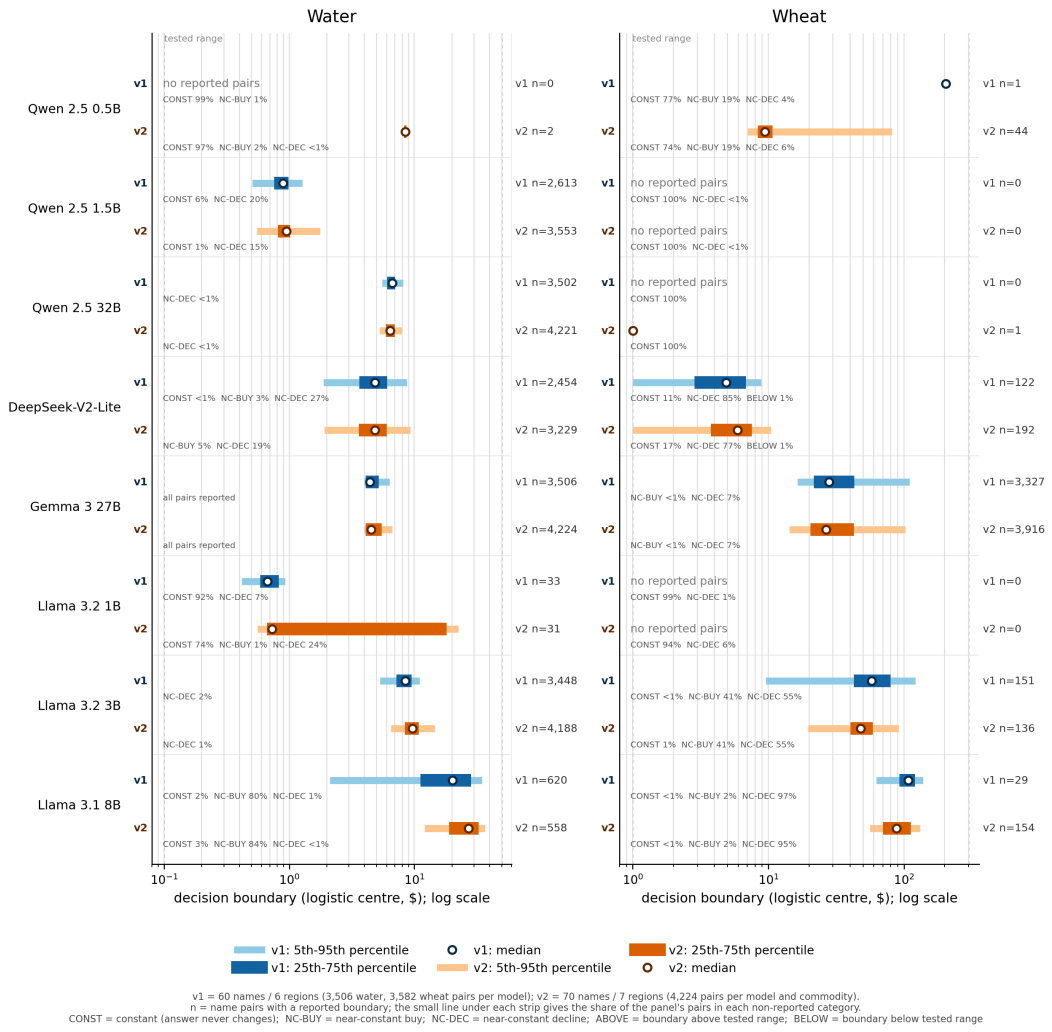


Figure 20: Distribution of the decision boundary across name pairs, first set (blue) and second set (orange) per model; light intervals 5th-95th percentiles, dark 25th-75th, circles the median, with the share of non-reporting pairs under each strip.

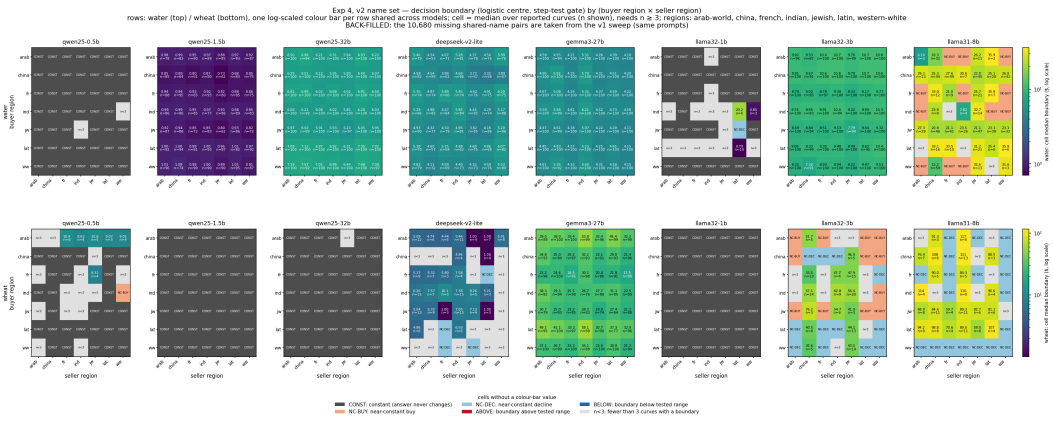


Figure 21: Second name set: median decision boundary per (buyer region, seller region) cell with at least three reporting pairs, shared pairs back-filled from the first sweep; same color rules as Appendix Figure 16.

Table 16: Cross-model alignment of coalesced extrema against the location-preserving null. Observed and null are mean alignment over ordered model pairs in the block (self-pairs excluded); the last column counts pairs whose observed alignment exceeds the null at $p < 0.05$ (upper tail).

Comparison	Ordered pairs	Observed	Null mean	Observed − null	Pairs $p < 0.05$
within per-digit block	20	0.079	0.191	-0.112	0
within chunked block	6	0.251	0.348	-0.097	0
per-digit → chunked	15	0.202	0.344	-0.141	1
chunked → per-digit	15	0.101	0.150	-0.049	1
within CN-origin	12	0.060	0.179	-0.119	0
within US-origin	12	0.197	0.313	-0.116	0
CN → US	16	0.192	0.338	-0.146	1
US → CN	16	0.092	0.125	-0.033	1
Gemma-3 → Llama	3	0.165	0.317	-0.152	0
Llama → Gemma-3	3	0.122	0.239	-0.117	0
Gemma-3 → CN per-digit	4	0.082	0.117	-0.034	0
CN per-digit → Gemma-3	4	0.131	0.300	-0.169	0

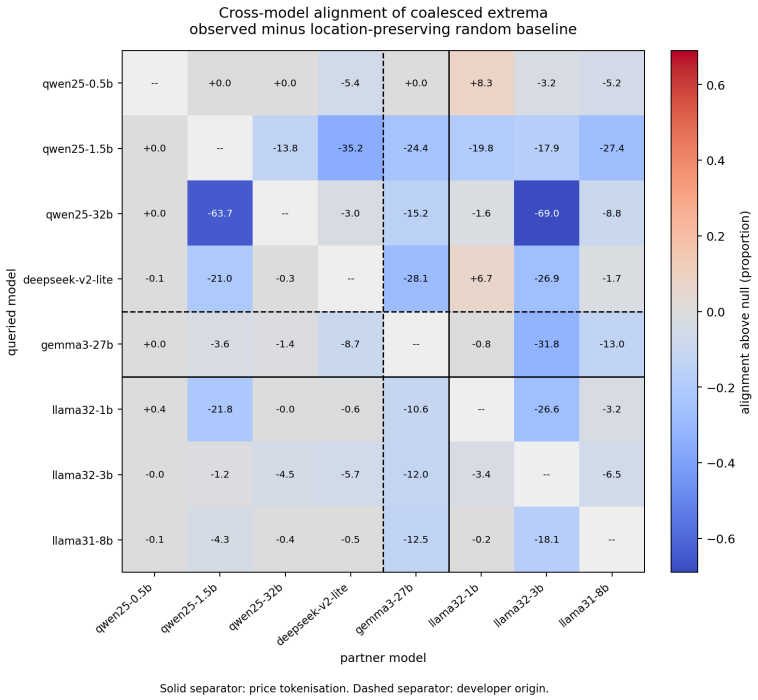


Figure 22: Observed minus null alignment for every ordered model pair (rows query columns; values in percentage points). Solid separator: price tokenization; dashed: developer origin.