# ENURI: EVIDENCE NEGOTIATION USING REASONING OVER IMAGES

*Did the agent miss the defect, or misprice it?*

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

A language-model agent acting on visual evidence must notice the evidence that matters (Detection) and weigh it correctly (Valuation), but outcome-level evaluations score only the final result and cannot tell which of the two failed. We introduce ENURI, a bilateral price-negotiation benchmark that extends the diagnostic TERMS-BENCH environment with an image channel in which an item's defects appear only in its photo and carry a known ground-truth price impact. Placing defects on a visibility $\times$ alignment grid (Obvious/Subtle $\times$ Aligned/Misaligned) makes the correct response to evidence verifiable, so every failure can be attributed to one stage or the other. We evaluate 13 frontier, open-weight, and sub-frontier LLM/VLM agents and three rule-based baselines over 6,400 episodes, with rankings stable across four price-impact mappings. The two stages fail along different axes, and neither failure is visible in the outcome score. Detection recall falls from 0.551 to 0.278 on Subtle defects for all 13 agents, whereas Valuation accuracy holds between 0.92 and 0.99 in three cells of the grid and falls to 0.73 in Subtle-Misaligned alone. Claude Opus 4.6 captures the most surplus of any agent (normalized surplus 0.593) yet shows the largest Detection drop, from 0.612 recall on Obvious defects to 0.217 on Subtle ones, while GPT-4o-mini cites most precisely yet scores the weakest overall. ENURI thus shows not only which agents fail but which stage fails. We release the benchmark, dataset, and per-episode traces.

## 1 INTRODUCTION

Multimodal agents are increasingly asked to act on what they see, not only to describe it. A web agent reads a price off a screenshot, a claims agent assesses pictured damage, a shopping assistant judges a used item from its photos. Each has to notice the evidence that matters (Detection) and then decide how much that evidence should change its action (Valuation) (Figure 1).

Existing benchmarks address the two stages one at a time. Benchmarks built for perception ask questions about an image (Antol et al., 2015; Yue et al., 2024), so they do not check whether the perceived detail moved an action. Benchmarks built for agents score the final outcome (Zhou et al., 2024; Froger et al., 2026), so they do not check what the agent saw. When an agent gets something wrong, neither line of work can say whether it failed to notice the evidence or noticed it and misjudged it.

Bilateral price negotiation makes both stages measurable at once. A negotiation ends in an agreed price or a walk-away, and that outcome can be scored against a reservation value. When a defect's true effect on price is fixed in advance, an experimental control analogous to fixing reservation values, the weight an agent gave to what it saw becomes checkable too. Existing negotiation benchmarks leave that property unused. NEGOTIATIONARENA (Bianchi et al., 2024), TERMS-BENCH (Zhang et al., 2026), and AGENTICPAY (Liu et al., 2026) are language-mediated end to end, so perception never enters the measurement and a mispriced item cannot be traced to Detection rather than Valuation. Section 2 surveys the wider field.

ENURI (Evidence Negotiation Using Reasoning over Images) makes that attribution possible by placing the difficulty of the evidence itself under experimental control. We extend the TERMS-



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.

BENCH bilateral price-negotiation kernel with a defect that exists only in the item's photo and never in its text description, so the agent can learn about it only by looking. We vary how visible each defect is (Obvious or Subtle) and whether its apparent severity matches its true price impact (Aligned or Misaligned). A conspicuous but cosmetic stain and a barely visible structural crack are both Misaligned, in opposite directions. Because the agent must name any defect it believes it sees in a structured field each turn, Detection is recorded independently of the price it offers, and Valuation is scored only where a real defect was named. Visibility targets Detection and alignment targets Valuation by design, so a failure can be located rather than merely counted.

**Contributions** This paper makes three contributions.

- **ENURI benchmark:** We introduce a bilateral price-negotiation benchmark in which defect information reaches the agent only through the item image, with defects placed on a visibility (Obvious/Subtle) $\times$ alignment (Aligned/Misaligned) grid that separates Detection failures from Valuation failures.
- **Robustness-checked evaluation:** We score 13 LLM/VLM agents and three rule-based baselines over 6,400 episodes under four independently constructed price-impact mappings, check that rankings are stable under that choice, and control for a structural role asymmetry in the environment that would otherwise be read as an agent capability difference (Appendix H).
- **Two dissociable failure modes:** Detection degrades on Subtle defects for all 13 agents (recall falls from 0.551 to 0.278), whereas Valuation holds in three cells of the grid and fails only where a hard-to-see defect also misrepresents its own severity, a residual beyond what either axis predicts on its own. Neither failure is visible in the outcome score, where Claude Opus 4.6 captures the most surplus while showing the largest Detection drop of any agent, and fabricating a defect carries no measurable benefit.

Taken together, these results show that an outcome-only leaderboard can rank an agent highly while it is failing to see, which is an argument for scoring perception separately from the action it drives.

## 2 RELATED WORK

**Perception benchmarks and agent benchmarks** Multimodal understanding benchmarks such as VQA (Antol et al., 2015) and MMMU (Yue et al., 2024) ask a model questions about an image. They isolate perception cleanly, but the answer is where the task ends, so nothing downstream depends on how much the perceived detail was worth. Agent benchmarks run the other way. WEBARENA (Zhou et al., 2024) scores whether a task was completed, GAIA2 (Froger et al., 2026) verifies every state-changing write action against an oracle, and VISUALWEBARENA (Koh et al., 2024) adds the missing modality to grounded web tasks while still rewarding only task completion. All three verify the action, not the evidence behind it.

**Text-only negotiation benchmarks** NEGOTIATIONARENA (Bianchi et al., 2024), Xia et al. (2024), NEGOTIATIONTOM (Chan et al., 2024), PIEARENA (Zhu et al., 2026), and Abdelnabi et al. (2024) have LLM agents bargain and score the resulting deals. TERMS-BENCH (Zhang et al., 2026) goes furthest in making the environment itself diagnostic, fixing the counterpart as a parameterized kernel so outcome differences can be attributed to the agent rather than counterpart variability. We adopt and extend that kernel (§3.3). AGENTICPAY (Liu et al., 2026) scales market structure instead, from bilateral to many-to-many. None of these carry an image channel, so any evidence they use is already stated in the prompt, and evidence difficulty can be neither controlled nor scored. The same holds for the static human-collected corpora several are built on (CASINO (Chawla et al., 2021), CRAIGSLISTBARGAIN (He et al., 2018)), whose listing photos ENURI builds on but which are neither controlled nor scored there.

**Vision-grounded strategic benchmarks** VS-BENCH (Xu et al., 2026) evaluates VLM agents across ten vision-grounded multi-agent environments and scores perception, strategic reasoning, and decision-making separately, reporting that models perceive well but decide far below optimal, a pattern close to our Detection/Valuation split. Its environments are games rather than negotiation, though, and its decision score is task-general episode return rather than a quantity with a known

correct value. ENURI's Valuation metric instead checks a falsifiable quantity against a known answer, whether the cited defect moved the price by the amount it is actually worth (§3.5), so a Valuation failure is provably wrong, not merely worse.

No benchmark surveyed above combines agent-generated negotiation dialogue with a visual-evidence channel. ENURI is simultaneously bilateral-negotiation, vision-grounded, evidence-controlled, and valuation-scored, a combination absent from the benchmarks surveyed here (column-by-column comparison in Appendix A, Table 6).

## 3 The ENURI Benchmark

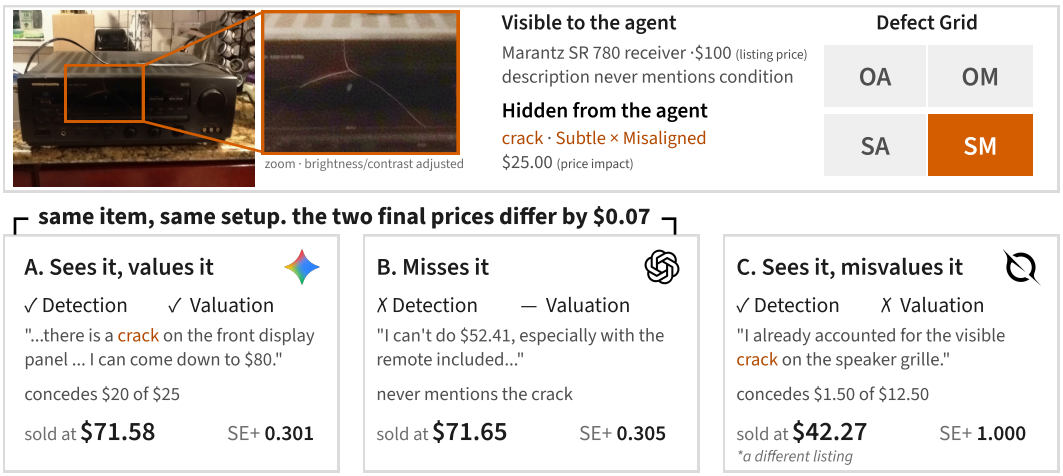



Figure 1: **The same outcome can hide different failures.** Three real, unedited episodes under mapping B, all with the agent as *seller*, where naming a defect costs it money. A (Gemini 3.1 Pro Preview) and B (GPT-5.5) share a receiver. C (Grok 4.5) is a different listing, four speakers and an amp at $50. An outcome-only score calls A and B equivalent and gives C, the worst valuation of the three, a perfect 1.000. $SE^+$ is the surplus share (§3.5), and valuation marks use the concession-to-impact ratio (calibrated 2/3–3/2, Table 14). Full transcripts in Appendix E.

### 3.1 Task

An evaluated agent is assigned the role of buyer or seller and negotiates a single scalar price for a used item against a fixed, environment-simulated counterpart over up to $K$ alternating-offer rounds, following the bilateral price-negotiation protocol of Zhang et al. (2026). Each round the agent emits a decision $d_k \in \{\text{Offer}, \text{Accept}, \text{Reject}\}$, a price $p_k$ when offering, a message, and a set of cited defect types, given the item's listing metadata, its own reservation value $r_A$, the price bounds $[p_{\min}, p_{\max}]$, and the negotiation history so far (full prompt in §3.5). The counterpart's private reservation value, urgency, and stance are drawn from a fixed prior and remain hidden from the agent throughout. ENURI's departure from this base protocol is the *evidence channel*. Each item may carry a defect that is rendered into the item image and is *never* described in the item's text metadata. The prompt discloses only the closed vocabulary of possible defect *types* (§3.5), never which type applies, so the agent can use a defect only by citing it in a negotiation turn.

### 3.2 The Visibility × Alignment Defect Design

ENURI crosses two design axes when each defect is constructed. Visibility (Obvious vs. Subtle) is how easy the defect is to notice in the image. Alignment (Aligned vs. Misaligned) is whether how conspicuous the defect looks matches how much it should actually affect price. Together they give four quadrants OA/OM/SA/SM with counterpart evidence weight $w[q]$ (right, with rationale in §3.3). Both axes are ground-truth labels fixed at construction time. The agent is never told

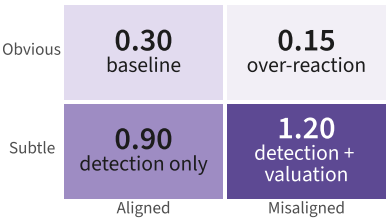



Figure 2: Quadrant weight $w[q]$.

a defect's visibility or alignment class and must infer, from
the rendered image alone, both whether a defect is present (Detection) and how much it should move the price (Valuation). The intersection, Subtle-Misaligned, requires both Detection *and* Valuation to succeed. The defect must first be noticed despite being hard to see, then correctly judged to matter more than its appearance suggests. Obvious-Misaligned runs the opposite valuation test. It contains a conspicuous defect that should *not* move the price as much as it looks like it should, so an agent that over-reacts to it fails Valuation in the other direction. Only Subtle-Misaligned combines a hard Detection case with this harder direction of Valuation, which is why it turns out to be the benchmark's hardest cell (§7).

### 3.3 Counterpart Kernel and the Evidence Term

Following the TERMS-Bench counterpart formulation, the counterpart's acceptance probability is gated by Individual Rationality. An offer that is not individually rational for the counterpart ($\bar{\delta} < 0$) is rejected outright, independently of $g$, so a high evidence weight $w[q]$ cannot rescue an individually-irrational offer. Otherwise the acceptance probability is $\sigma(g)$, where $g$ combines five negotiation-dynamics terms with an *evidence term* that is the only place the quadrant design enters the counterpart's decision:

$$g = \underbrace{\alpha\bar{\delta} + \beta\kappa - \gamma\bar{\tilde{D}} + \rho_F(\eta_B)\, c + \xi_F(\eta_B)\, \text{rigid}}_{\text{negotiation dynamics}} + \underbrace{s_A \cdot w[q]}_{\text{evidence term}}, \tag{1}$$

where $\bar{\delta}$ is offer favorability, $\kappa$ urgency, $\bar{\tilde{D}} = 1 - \sqrt{k/K}$ time remaining, $c$ the agent's concession speed, rigid flags a recent small concession, and $\eta_B$ the counterpart's hidden stance. $(\alpha, \beta, \gamma)$ are fixed weighting constants inherited from TERMS-Bench, while $\rho_F(\eta_B)$ and $\xi_F(\eta_B)$ are stance-dependent weights looked up from $\eta_B$, also inherited from TERMS-Bench (full notation in Appendix B). Citing a real defect helps a buyer's offer and hurts a seller's, so $s_A = +1$ for buyer / $-1$ for seller, and $w[q]$ is the quadrant evidence weight of Figure 2. It is a baseline value for Obvious-Aligned (easy to notice, so citing it is not very informative), tripled for Subtle-Aligned (noticing it is real evidence of careful looking), minimized for Obvious-Misaligned (a citation there should not outweigh correctly priced evidence), and maximized for Subtle-Misaligned, where detection and correct severity-reading must both succeed. A citation only enters this term if it names a defect actually present, so fabricated defects contribute 0 by construction and the kernel economically ignores hallucinated evidence at the point of decision (§7). Only our post-hoc metrics score whether the agent moved the price by the economically correct amount.

### 3.4 Price-Impact Mapping Robustness

Each defect type's price impact is price impact(type) $= M$[type] $\times$ listing price, where $M$ is one of four independently constructed mappings, namely conservative (A), moderate (B), aggressive (C), and nonlinear (D). All four preserve the same relative severity ordering across defect types. Cosmetic types (scratch, stain) are cheapest, structural types (crack, dent, rust, tear) are mid-tier, and missing part is costliest. The zone of possible agreement is anchored at fair price = listing price $-$ $\sum_d$ price impact$(d)$. Scoring all 16 agents on the same items under all four mappings lets us check whether a ranking is an artifact of one particular price curve (§5). Mappings A, B, and C preserve the tier ordering while raising the overall discount intensity in sequence, testing whether an agent's ranking is stable as the stakes of citing evidence grow, but the relative gap between tiers is not. The structural-to-cosmetic ratio runs 3.00/2.50/2.67 in A/B/C, and D pushes it to 3.75, so robustness under D tests whether a ranking survives a real change in tier structure (percentages by tier in Appendix B, Table 7).

### 3.5 Diagnostic Metrics

Every episode is reduced to nine diagnostics. Four **negotiation-outcome** metrics are inherited unchanged from the TERMS-Bench protocol (Zhang et al., 2026) we extend and are defined for every agent, including the fixed-concession baselines. $SE^+$ is the mean share of the zone-of-possible-agreement (ZOPA) width the agent captures, $AGR^+$ is its agreement (closing) rate on feasible episodes, $CSE^+$ is its mean surplus conditional on agreement (so $SE^+ = AGR^+ \cdot CSE^+$, closing rate times per-deal value), and CritViol% is its rate of price-range or individual-rationality protocol

violations (closed-form definitions in Appendix B). The remaining five **visual-evidence-use** metrics (Table 1) are this paper's core measurement contribution. They score whether a citation is accurate, acted on, and correctly sized. $C_i/D_i$ are the defect types cited and actually present in episode $i$, $O_i$ is the agent's `OFFER` turns citing a defect and $U_i \subseteq O_i$ the subset that concede, and $x_t/y_t$ are a cited defect's true price impact and the realized price move on turns citing exactly one real defect (full notation in Appendix B).

Table 1: The five visual-evidence-use metrics, this paper's core measurement contribution.

| Metric | Definition | Captures |
|---|---|---|
| cite-prec. | $\underset{i}{\text{mean}} \dfrac{\lvert C_i \cap D_i \rvert}{\lvert C_i \rvert}$ | Share of cited defects that are real |
| cite-cov. | $\underset{i}{\text{mean}} \dfrac{\lvert C_i \cap D_i \rvert}{\lvert D_i \rvert}$ | Share of real defects that get cited (recall) |
| halluc. rate | $\dfrac{1}{N} \sum_i \mathbf{1}[\, C_i \setminus D_i \neq \emptyset \,]$ | Share of episodes with a fabricated citation |
| Util. Prox. | $\underset{i}{\text{mean}} \dfrac{\lvert U_i \rvert}{\lvert O_i \rvert}$ | Share of citations that led to a concession |
| calib. $\rho$ | Spearman$(\{x_t\}, \{y_t\})$ | True severity vs. size of the price move |

**Agent interface.** The agent acts via a forced function call (`negotiate_action`) exactly once per turn, so decision, price, message, and citations are all machine-parseable. When the item carries an image it is attached to the prompt, which discloses only the closed vocabulary of seven defect *type* tags (never which defect, if any, is present). All 13 LLM/VLM agents receive an identical prompt and tool schema, so citation-behavior differences across agents (§7, §5) cannot be attributed to prompt wording (full verbatim prompt and schema in Appendix C).

### 3.6 Dataset Construction and Quality Control

Items are real CraigslistBargain listings (He et al., 2018) restricted to four categories with a natural notion of physical damage (bike, furniture, electronics, car), filtered to items whose listing text does not already mention damage, so defect information is only ever recoverable from the image. Each item is assigned to one of the four quadrants or to a no-defect control, subject to a compatibility catalog restricting which defect types may realize which quadrant (full catalog in Appendix B). Each candidate's photo is then edited to render the defect and is accepted only after an independent, blind verification call rediscovers the damage from the edited photo alone (Figure 3, with full prompts and QA criteria in Appendix B and Appendix D). The pipeline yields 324 accepted items spanning all four quadrants plus a no-defect control (breakdown in Appendix B).

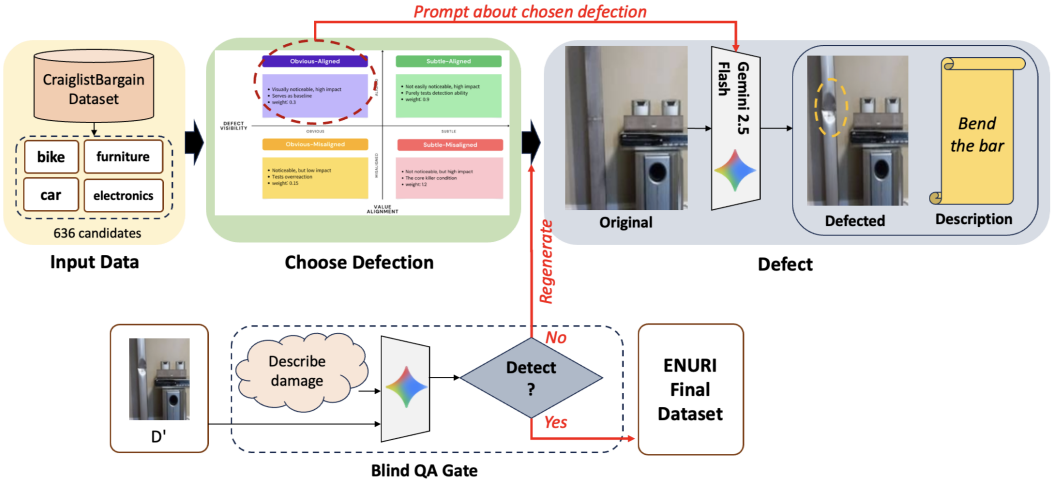



Figure 3: Dataset construction pipeline. A generation call edits the photo and writes the ground-truth description, and a separate blind call must independently rediscover the damage before the item is accepted.

## 4 EXPERIMENTAL SETUP

We evaluate 16 agents, comprising 13 LLM/VLM systems spanning frontier closed-source (Claude Opus 4.6 (Anthropic, 2026a), Claude Sonnet 4.6 (Anthropic, 2026b), Gemini 3.1 Pro Preview (Google DeepMind, 2026a), GPT-5.5 (OpenAI, 2026), Grok 4.5 (xAI, 2026), Qwen3.6-Plus (Qwen Team, 2026)), open-weight (Gemma-4-31B (Google DeepMind, 2026b), Qwen3-VL-32B-Instruct (Bai et al., 2025), Kimi K3 (Moonshot AI, 2026), ThinkingMachines-Inkling (Thinking Machines Lab, 2026), Nemotron-3-Nano-Omni (NVIDIA, 2026)), and sub-frontier (GPT-4o (OpenAI, 2024a), GPT-4o-mini (OpenAI, 2024b)) tiers, plus three fixed-concession baselines (FC-1/10/30) that concede a fixed percentage per round without reading the image. The roster is chosen for vendor and tier diversity under a native image-input requirement (pairing rationale and API access details in Appendix B). Each of the 16 agents plays 100 seeded episodes under each of the 4 price-impact mappings, for $16 \times 4 \times 100 = 6{,}400$ negotiation episodes total, over a fixed 100-item sample drawn from the 324-item pool of §3.6 (identical across all agents and mappings). $21\%$ of this sample carries no defect (versus $16.7\%$ of the pool) and serves as a hallucination control (§5).

## 5 TRUSTING THE MEASUREMENT

Three preconditions must hold before any comparison is meaningful. The measurement must be about the image, the ranking must survive the choice of price-impact mapping, and the environment must not structurally favor some episodes over others.

**Image dependence.** Is citation behavior actually driven by the item image, or could an agent produce plausible-looking citations from category and listing-text priors alone? A causal image on/off ablation on three Qwen models (right) removes the image entirely. Citation coverage collapses by $0.29$–$0.52$, to near zero for all three, yet $SE^+$ moves by only $0.014$ to $0.032$ ($1.4$ to $3.2$ points). The outcome score barely registers the loss of the whole visual channel, which directly supports the claim of §6 that outcome does not reveal what an agent saw. A lower-cost proxy over the full 13-agent roster agrees. On the $21\%$ of items with no defect, where any citation is a hallucination, the citation-rate gap between defective and non-defective items ranges from $0.053$ (Claude Opus 4.6) to $0.707$ (Nemotron-3-Nano-Omni; Appendix G).

Table 2: Image on/off ablation for the Qwen family under the moderate (B) mapping, 100 episodes/agent/condition. Values are with image, and $\Delta$ is No-Image $-$ Image.

| Agent | $SE^+$ | $\Delta SE^+$ | cite-cov. | $\Delta$cite-cov. |
|---|---|---|---|---|
| qwen3-vl-8b | 0.187 | $-0.018$ | 0.312 | $-0.294$ |
| qwen3-vl-32b | 0.152 | $-0.014$ | 0.418 | $-0.389$ |
| qwen3.6-plus | 0.503 | $-0.032$ | 0.557 | $-0.516$ |

**Mapping robustness.** Figure 4A shows $SE^+$ for every (agent, mapping) cell. Rank order is broadly preserved. Pairwise Spearman correlation of the $SE^+$ ranking across all $\binom{4}{2} = 6$ mapping pairs ranges from $\rho = 0.60$ to $0.87$ (mean $0.76$), and is even more stable for utilization-proximity ($\rho_{\text{pair}} \in [0.87, 0.97]$). Severity-calibration $\rho$ is the least stable diagnostic ($\rho_{\text{pair}} \in [0.27, 0.79]$, Appendix F). One outlier episode (Gemini 3.1 Pro Preview under mapping A, $SE^+ = -35.3$) is clipped in the color scale and marked $*$ rather than discarded. Figure 4B breaks $SE^+$ out by defect quadrant instead. Even the image-blind fixed-concession baselines vary by quadrant (FC-30 scores $0.23$, $0.21$, $0.21$, $0.27$ on OA, SA, OM, SM), so these differences reflect role and item mix as much as perception (Appendix H).

**Environment fairness: role asymmetry.** The environment has a structural buyer/seller asymmetry. 12 of the 13 LLM agents score lower as buyer than as seller, a property of the counterpart's price bounds, not of any evaluated agent. Because role is assigned from the same fixed seed for every agent, the rankings in §6–§7 are unaffected, but statistics pooled *within* an agent across quadrants inherit it. Appendix H gives the mechanism and role-conditional results, which is why Table 4's within-agent comparison is the more trustworthy diagnostic.

## 6 THE OUTCOME SCORE DOES NOT REVEAL WHAT AN AGENT SAW

Figure 5 and Table 3 report all nine headline diagnostics under the main mapping, moderate (B). Claude Opus 4.6, Grok 4.5, and Kimi K3 lead on $SE^+$ ($0.59$), followed by a cluster of three more

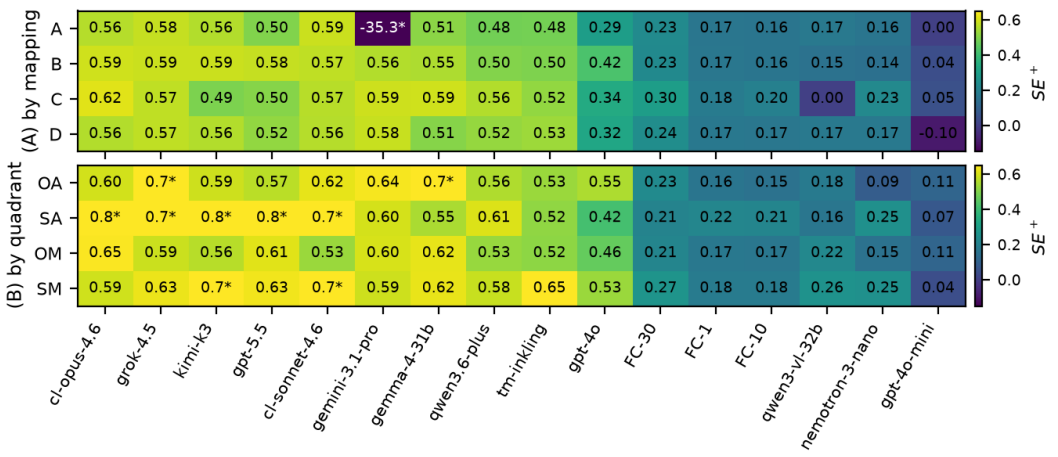



Figure 4: $SE^+$ by (A) price-impact mapping and (B) defect quadrant, same 16 agents. Values outside the color range are clipped and marked $*$.

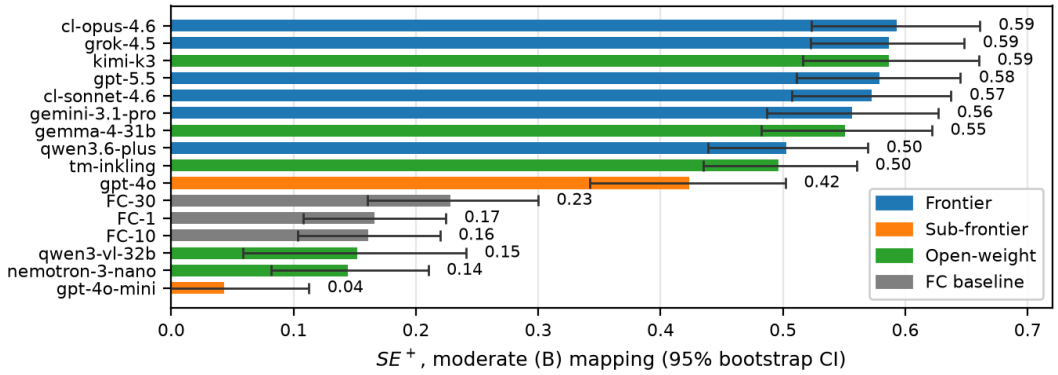



Figure 5: Agent ranking by $SE^+$ under the moderate (B) mapping, 100 episodes/agent. Bar-end numbers are point estimates, whiskers are 95% bootstrap CIs.

agents (0.56–0.58) with overlapping bootstrap CIs. Three of the 13 LLM agents score *below* all three fixed-concession baselines, with GPT-4o-mini lowest overall ($SE^+ = 0.04$, near-random pricing).

That single number already misleads about what separates the top and bottom of this ranking. Claude Opus 4.6 leads on $SE^+$ while its hallucination rate (0.410) and Subtle recall (§7) are both among the roster's worst. GPT-4o-mini has the *highest* citation precision of any LLM agent (0.670) while scoring lowest overall on $SE^+$. $SE^+$'s own two factors also dissociate. GPT-5.5, Qwen3.6-Plus, and ThinkingMachines-Inkling tie for the roster lead on closing rate ($AGR^+ = 0.940$), but only GPT-5.5 converts that into terminal value ($CSE^+ = 0.616$, against 0.535 and 0.528), while Kimi K3 closes fewer feasible deals ($AGR^+ = 0.810$) yet extracts the best terms when it does ($CSE^+ = 0.724$). A single $SE^+$ number hides both axes.

Whether citing the real defect helps at all further shows that outcome and perception are different axes. Across the 1,027 feasible episodes with a real defect present (mapping B, 13 LLM agents, role-pooled), episodes where the agent ever cites the real defect ($n = 464$) average $SE^+ = 0.474$, against 0.512 for episodes where it never does ($n = 563$). This is an observational comparison, not a randomized test, but citing shows no average advantage for the agent making it. That does *not* make citing pointless, since a firm counterpart or a favorable ZOPA can dominate $SE^+$ regardless of what the agent noticed or said (Appendix E.1 gives a paired example).

## 7 Detection and Valuation Are Dissociable Failures

When an agent misprices a defect, is that because it never noticed the defect, or because it noticed but priced it wrong? Pooling all 13 LLM agents onto the same visibility×alignment grid (§3.2) makes this dissociation measurable. Figure 6 lays out Detection rate (citation recall) and a grounding hit-rate

Table 3: All nine headline diagnostics per agent under the moderate (B) mapping, 100 episodes/agent (sorted by $SE^+$). **Bold**/*italic* = best/second-best among the 13 LLM/VLM agents. Baselines never read the item image, so for them cite-prec., Util. Prox., and calib. $\rho$ are undefined (–) and cite-cov. and halluc. rate are 0 by construction.

| | Negotiation outcome | | | | Visual evidence use | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Agent | $SE^+$ | $AGR^+$ | $CSE^+$ | CritViol% | cite-prec. | cite-cov. | halluc. rate | Util. Prox. | calib. $\rho$ |
| cl-opus-4.6 | **0.593** | 0.860 | *0.689* | 0.020 | 0.458 | 0.456 | 0.410 | *0.699* | **0.803** |
| grok-4.5 | *0.587* | 0.890 | 0.659 | 0.030 | 0.585 | 0.468 | 0.270 | 0.690 | 0.659 |
| kimi-k3 | *0.587* | 0.810 | **0.724** | **0.000** | 0.417 | 0.304 | 0.300 | 0.657 | 0.656 |
| gpt-5.5 | 0.579 | **0.940** | 0.616 | 0.040 | *0.639* | 0.430 | 0.200 | 0.662 | 0.701 |
| cl-sonnet-4.6 | 0.572 | *0.900* | 0.636 | 0.040 | 0.457 | *0.544* | 0.490 | 0.653 | 0.692 |
| gemini-3.1-pro | 0.557 | 0.840 | 0.663 | **0.000** | 0.608 | *0.544* | 0.290 | **0.727** | 0.728 |
| gemma-4-31b | 0.551 | 0.840 | 0.656 | *0.010* | 0.547 | 0.228 | **0.150** | 0.543 | 0.659 |
| qwen3.6-plus | 0.503 | **0.940** | 0.535 | 0.020 | 0.551 | **0.557** | 0.360 | 0.498 | *0.758* |
| tm-inkling | 0.496 | **0.940** | 0.528 | 0.040 | 0.394 | 0.494 | 0.560 | 0.574 | 0.756 |
| gpt-4o | 0.423 | 0.870 | 0.486 | 0.170 | 0.569 | 0.519 | 0.320 | 0.643 | 0.591 |
| FC-30 *(baseline)* | 0.228 | 0.490 | 0.466 | 0.000 | – | 0.000 | 0.000 | – | – |
| FC-1 *(baseline)* | 0.166 | 0.410 | 0.405 | 0.000 | – | 0.000 | 0.000 | – | – |
| FC-10 *(baseline)* | 0.161 | 0.410 | 0.393 | 0.000 | – | 0.000 | 0.000 | – | – |
| qwen3-vl-32b | 0.152 | 0.840 | 0.181 | 0.390 | 0.469 | 0.418 | 0.370 | 0.324 | 0.453 |
| nemotron-3-nano | 0.145 | 0.590 | 0.245 | 0.240 | 0.617 | 0.519 | 0.260 | 0.145 | 0.749 |
| gpt-4o-mini | 0.043 | 0.600 | 0.072 | 0.360 | **0.670** | 0.392 | *0.160* | 0.226 | 0.377 |

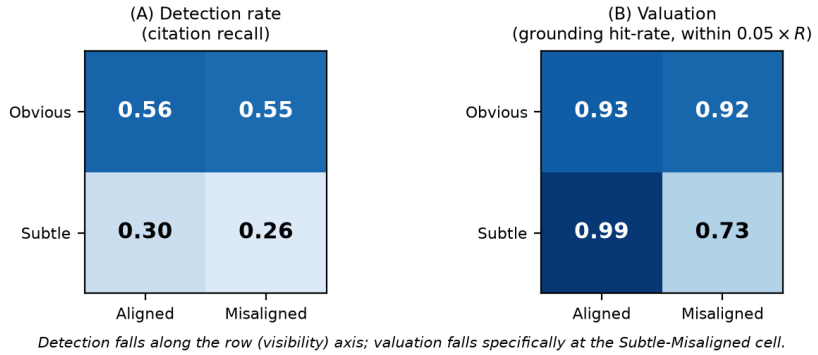



Figure 6: Detection rate (A) and grounding hit-rate (B) on the same visibility×alignment grid, pooled across all 13 LLM agents.

on identical $2 \times 2$ axes. The grounding hit-rate is the fraction of single-citation offer turns whose price concession lands within $0.05 \times R$ of the cited defect's true price impact, where $R = p_{\max} - p_{\min}$.

**Detection fails along visibility, for every agent.** Detection (panel A) shows a large row effect (Obvious 0.56 vs. Subtle 0.28) and a negligible column effect (0.56 vs. 0.55 within Obvious). Failing to notice a defect tracks visibility almost exclusively. Pooling episodes across all four mappings, mean citation recall drops from 0.551 on Obvious defects to 0.278 on Subtle ones, and the per-agent drop ranges from 0.143 to 0.396. Claude Opus 4.6, the $SE^+$ leader (§6), has the largest drop, from 0.612 on Obvious to 0.217 on Subtle defects. After Benjamini–Hochberg correction (Benjamini & Hochberg, 1995) ($\alpha = 0.05$, bootstrap CIs (Efron & Tibshirani, 1993), $B = 2{,}000$), all **13/13** agents remain significant (largest corrected $p < 0.01$), and the effect varies by defect type (Appendix G.1). This pooled test mixes roles. In buyer-only episodes, where citing is always in the agent's interest, the drop stays positive for all 13 agents but its CI excludes zero for only 5 (Figure 19), so part of the pooled effect reflects sellers' incentive not to disclose. Visibility changes *whether* a defect gets noticed, not how long noticing takes, since the mean turn of first citation is 0.193 of the round budget $K$ for Obvious defects versus 0.195 for Subtle (Appendix G).

**Valuation fails only where Detection must also succeed.** Valuation (panel B) is the mirror image. Three of the four cells sit at 0.92–0.99, and only Subtle-Misaligned drops to 0.73. The other three cells predict 0.97 for it additively (OM + SA − OA), so the interaction contrast is −0.24, a residual beyond what either axis explains. Agents who correctly detect a Subtle-Misaligned defect still under-adjust price relative to a Subtle-Aligned defect of matched visibility, with a significantly negative Misaligned-minus-Aligned difference for 10 of the 11 agents that have single-citation turns in both

Table 4: Subtle-Misaligned grounding hit-rate vs. the other three quadrants pooled, paired by agent (10 of 13 LLM agents with $\geq 10$ single-citation Subtle-Misaligned turns).

| Agent | hit-rate (SM) | $n_{\text{SM}}$ | hit-rate (other) | $n_{\text{other}}$ | diff |
|---|---|---|---|---|---|
| cl-sonnet-4.6 | 0.565 | 23 | 0.928 | 222 | $-0.363$ |
| kimi-k3 | 0.615 | 13 | 0.945 | 163 | $-0.329$ |
| grok-4.5 | 0.625 | 32 | 0.913 | 218 | $-0.288$ |
| gpt-4o | 0.611 | 36 | 0.887 | 230 | $-0.276$ |
| tm-inkling | 0.727 | 11 | 0.972 | 141 | $-0.244$ |
| gpt-5.5 | 0.733 | 15 | 0.974 | 193 | $-0.241$ |
| gemini-3.1-pro | 0.780 | 41 | 0.934 | 197 | $-0.154$ |
| qwen3.6-plus | 0.800 | 15 | 0.944 | 179 | $-0.144$ |
| gemma-4-31b | 0.929 | 14 | 0.953 | 85 | $-0.024$ |
| nemotron-3-nano | 1.000 | 13 | 0.929 | 28 | $+0.071$ |

Subtle cells (BH-corrected, Appendix G). Table 4 confirms this with more power, pairing each agent's Subtle-Misaligned grounding hit-rate against its own hit-rate on the other three quadrants pooled. **9/10** agents with enough data are lower on Subtle-Misaligned (sign test $p = 0.022$). The three excluded agents (Claude Opus 4.6, Qwen3-VL-32B, GPT-4o-mini) have fewer than 10 single-citation Subtle-Misaligned turns. Mis-valued defects are one-sided, with under-adjustment outnumbering over-adjustment more than 30 to 1 (Appendix G.1). The $0.05R$ tolerance is loose relative to typical price impacts, so it registers the Subtle-Misaligned breakdown but not the pervasive under-adjustment of Table 14, where even Aligned citations under-adjust in 77% of turns. We therefore read the SM residual as a relative effect.

**Fabricated evidence yields no benefit.** Hallucinated claims are a well-documented LLM failure mode (Ji et al., 2023). The kernel gives fabricated citations zero evidence weight by construction (§3.3), so whether they cost anything at the level of terminal outcome is the empirical question. Across the 13 LLM agents of Table 3, hallucination rate is not significantly correlated with $SE^+$ (Spearman $\rho = 0.15$) or $CSE^+$ ($\rho = 0.07$). Within an agent, comparing episodes citing only fabricated defects against episodes citing nothing, an uncorrected 95% bootstrap CI flags 4/13 agents with a significant negative $\Delta SE^+$. After Benjamini–Hochberg correction only **2/13** remain significant (marked $^\dagger$ at right, with the full roster in Appendix H), driven mostly by failed agreement rather than worse deals, and no agent shows a significant *positive* $\Delta SE^+$ when pooled across roles, although several role-specific cells do (Appendix H, Table 18). These pooled comparisons exclude episodes with a logged protocol violation. What gets fabricated tracks category-typical defect priors (scratch for cars and bikes, crack for furniture) more than anything visible in the item's photo (Appendix G.1).

Table 5: Citing only fabricated defects vs. nothing, $\Delta SE^+$. $^*$ = 95% CI excludes 0, $^\dagger$ = significant after BH correction (full roster in Appendix H).

| Agent | $\Delta SE^+$ |
|---|---|
| qwen3-vl-32b$^\dagger$ | $-0.174^*$ |
| gemini-3.1-pro$^\dagger$ | $-0.164^*$ |
| cl-opus-4.6 | $-0.111^*$ |
| kimi-k3 | $-0.107^*$ |

## 8 Limitations and Conclusion

ENURI makes answerable a question outcome-only evaluation cannot ask. When an agent misprices an item, did it fail to see the evidence, or see it and misjudge it? Across all 16 agents and 6,400 episodes, Detection and Valuation fail along different axes. Detection collapses with visibility for every one of the 13 LLM/VLM agents (the three baselines never read the image), Valuation holds everywhere except the single cell, Subtle-Misaligned, that requires both skills at once, and fabricated citations earn no measurable benefit, none of which the outcome score alone reveals, since the agent with the highest surplus shows the largest Detection drop. This dissociation is ENURI's central contribution, a diagnostic split that credits or blames the right stage of an agent's decision rather than only its final price. Four limitations bound this claim. The visibility/alignment quadrant is a construction-time label rather than a measured property. The image-driven-citation check of §5 is mainly observational, with a randomized ablation on three Qwen models, two of which are in the main roster. All episodes also use a single counterpart configuration (the Candid behavior family in the overlap regime), so the dissociation is established against one counterpart profile rather than the six families TERMS-Bench defines. Citation also conflates perception with willingness to disclose, since a seller has an incentive not to name a defect it has seen. We release the benchmark, dataset, and per-episode traces so the Detection/Valuation split can be reused as an audit template for other agents acting on visual evidence.

### AI USE STATEMENT

We used generative AI tools to (i) generate the synthetic defect images used to construct the ENURI dataset (§3.6), (ii) assist in writing code, including the benchmark environment and analysis pipeline, and (iii) draft and revise sections of this paper's text, including fixing non-native and awkward English phrasing. We did not use generative AI tools to develop the theoretical or conceptual framework underlying the benchmark, to formulate or prove any mathematical claims, to propose or refine the paper's hypotheses, or to interpret experimental results. These were done by the authors. We have reviewed all AI-assisted work. All AI-suggested code was tested and verified by the authors, all AI-edited text was checked against our intended meaning, and every citation was checked against its source. We take responsibility for the final content of this work.

### ETHICS STATEMENT

Beyond the usual dual-use disclaimer, two of our findings are properties of the simulated counterpart specifically and should not be assumed to transfer to human counterparts. First, the Subtle-Misaligned result of §7 (defects detected but under-priced) describes our counterpart's blind spot, not necessarily a human's. Second, the finding of §7 that fabrication yields no benefit reflects our kernel's design, which gives fabricated evidence zero economic weight by construction. A human counterpart may be more persuadable by a confident false claim. §5's environment asymmetry is a related reminder that evaluation results can reflect properties of the benchmark rather than the agent, which matters most when such results are used to inform real deployment decisions.

### REPRODUCIBILITY STATEMENT

The counterpart kernel and its parameters are given in full in §3.3 and Appendix B. The four price-impact mappings are given in Appendix B, Table 7, and the verbatim agent prompt and tool schema are given in Appendix C. Dataset construction, including the image-generation and QA-gate prompts and the compatibility catalog restricting which defect types may realize which quadrant, is given in full in Appendix B and Appendix D. All statistical tests (bootstrap CIs, Benjamini–Hochberg correction) are specified with their resampling counts and test families in §7. We release the benchmark code, the 324-item dataset with per-item quadrant and price impact labels, and per-episode negotiation traces for all 16 agents under all four mappings.

## REFERENCES

Sahar Abdelnabi, Amr Gomaa, Sarath Sivaprasad, Lea Schönherr, and Mario Fritz. Cooperation, competition, and maliciousness: LLM-stakeholders interactive negotiation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Anthropic. Introducing Claude Opus 4.6. `https://www.anthropic.com/news/claude-opus-4-6`, 2026a.

Anthropic. Introducing Claude Sonnet 4.6. `https://www.anthropic.com/news/claude-sonnet-4-6`, 2026b.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015.

Shuai Bai et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300, 1995.

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 (ICML)*, 2024.

Chunkit Chan, Cheng Jiayang, Yauwai Yim, Zheye Deng, Wei Fan, Haoran Li, Xin Liu, Hongming Zhang, Weiqi Wang, and Yangqiu Song. NegotiationToM: A benchmark for stress-testing machine theory of mind on negotiation surrounding. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024.

Kushal Chawla, Jaysa Ramirez, Rene Clever, Gale Lucas, Jonathan May, and Jonathan Gratch. CaSiNo: A corpus of campsite negotiation dialogues for automatic negotiation systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pp. 3167–3185, 2021.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1993.

Romain Froger, Pierre Andrews, Matteo Bettini, Amar Budhiraja, Ricardo Silveira Cabral, Virginie Do, Emilien Garreau, Jean-Baptiste Gaya, Hugo Laurençon, Maxime Lecanu, Kunal Malkan, Dheeraj Mekala, Pierre Menard, Gerard Moreno-Torres Bertran, Ulyana Piterbarg, Mikhail Plekhanov, Mathieu Rita, Andrey Rusakov, Vladislav Vorotilov, Mengjue Wang, Ian Yu, Amine Benhalloum, Grégoire Mialon, and Thomas Scialom. Gaia2: Benchmarking LLM agents on dynamic and asynchronous environments. In *International Conference on Learning Representations (ICLR)*, 2026.

Google DeepMind. Gemini 3.1 Pro. `https://deepmind.google/models/gemini/pro/`, 2026a.

Google DeepMind. Gemma 4. `https://deepmind.google/models/gemma`, 2026b.

He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. Decoupling strategy and generation in negotiation dialogues. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

Jing Yu Koh, Robert Lo, Lawrence Jang, Vikram Duvvur, Ming Chong Lim, Po-Yu Huang, Graham Neubig, Shuyan Zhou, Ruslan Salakhutdinov, and Daniel Fried. VisualWebArena: Evaluating multimodal agents on realistic visual web tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Xianyang Liu, Shangding Gu, and Dawn Song. AgenticPay: A multi-agent LLM negotiation system for buyer-seller transactions. In *ICLR 2026 Workshop AIMS*, 2026.

Moonshot AI. Kimi K3. `https://www.moonshot.ai/`, 2026.

NVIDIA. Nemotron 3 nano omni: Efficient and open multimodal intelligence. *arXiv preprint arXiv:2604.24954*, 2026.

OpenAI. GPT-4o system card. `https://openai.com/index/gpt-4o-system-card/`, 2024a.

OpenAI. GPT-4o mini: Advancing cost-efficient intelligence. `https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/`, 2024b.

OpenAI. Introducing GPT-5.5. `https://openai.com/index/introducing-gpt-5-5/`, 2026.

Qwen Team. Qwen3.6-Plus: Towards real world agents. `https://qwen.ai/blog?id=qwen3.6`, 2026.

Thinking Machines Lab. Inkling: Our open-weights model. `https://thinkingmachines.ai/news/introducing-inkling/`, 2026.

xAI. Introducing Grok 4.5. `https://x.ai/news/grok-4-5`, 2026.

Tian Xia, Zhiwei He, Tong Ren, Yibo Miao, Zhuosheng Zhang, Yang Yang, and Rui Wang. Measuring bargaining abilities of LLMs: A benchmark and a buyer-enhancement method. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.

Zelai Xu, Zhexuan Xu, Xiangmin Yi, Huining Yuan, Mo Guang, Kaiwen Long, Xinlei Chen, Yi Wu, Chao Yu, and Yu Wang. VS-Bench: Evaluating VLMs for strategic abilities in multi-agent environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Erica Zhang, Fangzhao Zhang, Aneesh Pappu, Batu El, Jose Blanchet, Susan Athey, Jiashuo Liu, and James Zou. TERMS-bench: Diagnosing LLM negotiation agents beyond deal rate. *arXiv preprint arXiv:2605.13909*, 2026.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. WebArena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations (ICLR)*, 2024.

Chris Zhu, Sasha Cui, Will Sanok Dufallo, Runzhi Jin, Zhen Xu, Linjun Zhang, and Daylian Cain. PieArena: Ranking and profiling language agents in realistic negotiation scenarios. *arXiv preprint arXiv:2602.05302*, 2026.

# A Extended Related Work Comparison

Table 6 gives the column-by-column comparison behind §2's summary: **Negotiation** asks whether the task is bilateral or multilateral bargaining specifically. **Vision** asks whether the agent receives an image at all. **Controlled evidence** asks whether the difficulty of that image is varied by design. **Valuation** asks whether the benchmark scores how far the evidence should have moved the agent's action, not only whether it was noticed.

Table 6: Comparison with prior negotiation and strategic-decision benchmarks. A dash marks an axis that does not apply, since a benchmark with no image channel can neither control evidence difficulty nor score evidence valuation.

| Benchmark | Negot. | Vision | Ctrl. evid. | Valuation |
|---|---|---|---|---|
| NegotiationArena (Bianchi et al., 2024) | ✓ | ✗ | – | – |
| Xia et al. (2024) | ✓ | ✗ | – | – |
| NegotiationToM (Chan et al., 2024) | ✓ | ✗ | – | – |
| PieArena (Zhu et al., 2026) | ✓ | ✗ | – | – |
| Abdelnabi et al. (2024) | ✓ | ✗ | – | – |
| TERMS-Bench (Zhang et al., 2026) | ✓ | ✗ | – | – |
| AgenticPay (Liu et al., 2026) | ✓ | ✗ | – | – |
| VS-Bench (Xu et al., 2026) | ✗ | ✓ | ✗ | ✗ |
| **ENURI (ours)** | ✓ | ✓ | ✓ | ✓ |

# B Benchmark Construction Details

This appendix gives the full quantitative specification of the benchmark referenced from §3: the defect quadrant weights, the price-impact mappings, the closed-form definition of all nine diagnostic metrics, and the dataset construction pipeline.

## B.1 Kernel Notation

Figure 2 (§3.2) gives the counterpart's evidence weight $w[q]$ for each quadrant. Its magnitude is a same-turn proxy for how much genuine signal a citation in that cell carries, not a valuation check, which is instead assessed post-hoc (§3.5, §7). Two notational details about the kernel (Eq. 1) are worth making explicit, since the bar notation could otherwise be read as a temporal average. $\bar{\delta}$ is *not* averaged over turns. It is recomputed from scratch every turn using only the agent's most recent offer, and the bar instead denotes normalization by the price range $R$. $\kappa$, in contrast, genuinely is fixed for the whole episode, drawn once at episode start from $\mathrm{Beta}(2, 2)$ and not resampled each turn. Similarly, $c$ is signed *velocity*, not magnitude. It is the mean of the agent's last (up to three) consecutive offer-price deltas, sign-normalized by role so that conceding is always positive regardless of whether the agent is buyer or seller.

## B.2 Price-Impact Mappings

Table 7 gives the price-impact percentage by tier for each of the four mappings (§3.4).

Table 7: Price-impact percentage by tier and mapping ($M[\text{type}]$).

| Tier (types) | A | B | C | D |
|---|---|---|---|---|
| Cosmetic (scratch, stain) | 5% | 10% | 15% | 8% |
| Structural (crack, dent, rust, tear) | 15% | 25% | 40% | 30% |
| Missing part | 20% | 35% | 50% | 45% |

## B.3 Diagnostic Metric Definitions

Table 8 gives closed-form definitions of the nine headline diagnostics summarized in §3.5. Let episode $i$ have terminal outcome $f_i$ (a price, or $\perp$ for disagreement), buyer/seller reservation values

$r^{(i)}_{\text{buyer}}, r^{(i)}_{\text{seller}}$, $\Delta_i := r^{(i)}_{\text{buyer}} - r^{(i)}_{\text{seller}}$ (feasible iff $\Delta_i > 0$), and $u_A(f_i)$ the evaluated agent's realized utility. Write $\mathcal{I}^+ := \{i : \Delta_i > 0\}$ and $\mathcal{I}^+_{\text{agr}} := \{i \in \mathcal{I}^+ : f_i \neq \perp\}$. For the citation-accuracy and evidence-utilization rows, $C_i$ is the set of defect types the agent cites at least once in episode $i$ and $D_i$ the set of defect types actually present in item $i$ (hidden from the agent). $O_i$ is the agent's `OFFER` turns citing at least one defect and $U_i \subseteq O_i$ the subset that are concessions relative to the agent's own immediately preceding offer (sign-normalized by role). $x_t$ = price impact$(d(t))$ and $y_t = |p_t - p_{t-1}|/R$ for every `OFFER` turn $t$ citing exactly one real defect $d(t)$. Quadrant-level detection rate (§7) is cite-cov. restricted to $D_i \cap \{d : \text{quadrant}(d) = q\}$ in place of $D_i$. The Subtle-Misaligned caught-vs-missed gap (Appendix H only) partitions Subtle-Misaligned episodes into *caught* ($C_i$ contains the defect) and *missed*, and reports $CSE^+(\text{caught}) - CSE^+(\text{missed})$ (Eq. 2), an observational, not causal, comparison. $v_i = 1$ flags a protocol violation (an accept before any offer, an out-of-range offer, or an accept/offer violating individual rationality) and 0 otherwise. CritViol% does not require the agent to have read the item image, so it is defined (and typically near 0) even for the fixed-concession baselines. We use Spearman rather than Pearson for calib. $\rho$ because $x_t$'s scale is mapping-dependent (§3.4), and rank correlation is invariant to any monotone rescaling of price impact.

Table 8: The nine headline diagnostic metrics, grouped by axis, with closed-form definitions. "Negotiation outcome" metrics are defined for every agent, including the fixed-concession baselines. The baselines never read the item image, so among the "Visual evidence use" metrics cite-cov. and halluc. rate are 0 by construction, while cite-prec., Util. Prox., and calib. $\rho$ are undefined ($-$, Table 3).

| Axis | Metric | Definition |
|---|---|---|
| Negotiation outcome | $SE^+$ | $\frac{1}{\lvert\mathcal{I}^+\rvert} \sum_{i \in \mathcal{I}^+} \frac{u_A(f_i)}{\Delta_i}$ |
| | $AGR^+$ | $\frac{\lvert\mathcal{I}^+_{\text{agr}}\rvert}{\lvert\mathcal{I}^+\rvert}$ |
| | $CSE^+$ | $\frac{1}{\lvert\mathcal{I}^+_{\text{agr}}\rvert} \sum_{i \in \mathcal{I}^+_{\text{agr}}} \frac{u_A(f_i)}{\Delta_i}$; $SE^+ = AGR^+ \cdot CSE^+$ |
| | CritViol% | $\frac{1}{N} \sum_{i=1}^{N} v_i$ |
| Visual evidence use | cite-prec. | $\underset{i:\, C_i \neq \emptyset}{\text{mean}} \frac{\lvert C_i \cap D_i\rvert}{\lvert C_i\rvert}$ |
| | cite-cov. | $\underset{i:\, D_i \neq \emptyset}{\text{mean}} \frac{\lvert C_i \cap D_i\rvert}{\lvert D_i\rvert}$ |
| | halluc. rate | $\frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[\, C_i \setminus D_i \neq \emptyset\,]$ |
| | Util. Prox. | $\underset{i:\, O_i \neq \emptyset}{\text{mean}} \frac{\lvert U_i\rvert}{\lvert O_i\rvert}$ |
| | calib. $\rho$ | Spearman$(\{x_t\}, \{y_t\})$, pooled or per-quadrant |

$$\text{gap} = CSE^+(\text{caught}) - CSE^+(\text{missed}) \tag{2}$$

### B.4 Dataset Construction Pipeline

Items are real CraigslistBargain listings (He et al., 2018) restricted to four categories with a natural notion of physical damage (bike, furniture, electronics, car, giving 636 candidates with a usable photo), further filtered to 546 items whose original listing text does not already mention damage, preserving the invariant that defect information is only ever recoverable from the image. Each surviving item is assigned to one of the four quadrants or to a no-defect control (targeted at 10% of items, with the rest quadrant-balanced via the largest-remainder apportionment method), and a compatibility catalog restricts which defect *types* may realize which quadrant. Surface-type defects (scratch, stain, tear) are excluded from Subtle-Misaligned (tear is priced in the structural tier of Table 7 but grouped with the surface types here), since "invisible but structurally serious" is not a plausible story for a purely cosmetic defect, while missing_part is excluded from both Subtle cells, since a missing part is obvious by definition. Only the three inherently structural types (dent, rust, crack) can realize the benchmark's hardest cell.

**Image synthesis.** For each item, a single call to Gemini 2.5 Flash Image is given the item photo and asked to (i) choose, itself, which part of *this specific photo* is clearly visible enough to damage, and (ii) edit that part and describe what it did, in the same call, so the resulting text description and the rendered image are guaranteed to agree (both come from one generation context). Pilot iteration surfaced several failure modes and the prompt constraints that fixed each, documented in Appendix D.3.

**QA gate (LLM-as-judge).** Because a generation model can silently fail to render a convincing defect (most likely for Subtle items, where under-rendering and correct rendering look similar to a casual glance), every synthesized image is passed through a second, *blind* verification call before it is accepted. The same model is shown only the edited photo, with no defect description, location, or quadrant label, and asked "Does this photo show any visible damage, defects, or flaws? If yes, describe what you see; if no, say NONE." Any non-`NONE` response counts as a pass (a Subtle defect only needs to be find*able* on close inspection, not obvious). `NONE` triggers a regeneration, up to three attempts total, after which the item is discarded rather than kept with an unverified defect. This judge call is deliberately *decoupled* from the generation call and from any ground-truth description, so it cannot pass an item merely because the generation model insists it succeeded, because it is asked to independently rediscover the damage.

**Yield.** The pipeline produced $324$ accepted items (bike $97$, furniture $121$, electronics $68$, car $38$) spanning Obvious-Misaligned ($80$), Obvious-Aligned ($79$), Subtle-Misaligned ($59$), Subtle-Aligned ($52$), and no-defect ($54$). The $54$ no-defect controls are about $10\%$ of the $546$ assigned items and cannot fail generation or QA, whereas all $222$ discards fell on defect items, so the accepted pool is $54/324 = 16.7\%$ no-defect and the fixed $100$-item evaluation sample contains $21$. Defect types skew toward scratch ($105$) and dent ($61$), with rust ($54$), crack ($23$), stain ($15$), tear ($10$), and missing_part ($2$) rarer. Of $222$ discarded candidates, $73$ failed only due to transient API credit exhaustion (recoverable by retry, not a QA failure) and $149$ failed QA or generation outright, with a category-level pattern to the discards discussed in Appendix D.3.

### B.5 Agent Roster

The roster (§4) is built for vendor diversity and tier coverage rather than exhaustiveness. Native image-input support is a hard requirement, and within that constraint each provider contributes a frontier/sub-frontier or closed/open-weight pair where available, plus a single representative for vendors we could only cover once. OpenAI contributes a frontier-vs-sub-frontier comparison slot (GPT-5.5) against a weak-model floor anchor (GPT-4o/GPT-4o-mini). Anthropic contributes a high-capability reasoning model vs. a lighter everyday model (Claude Opus 4.6 vs. Claude Sonnet 4.6). Qwen contributes a general-purpose, multimodal-capable flagship vs. a vision-specialized open-weight model (Qwen3.6-Plus vs. Qwen3-VL-32B-Instruct). Google contributes a commercial model vs. an open-weight model (Gemini 3.1 Pro Preview vs. Gemma-4-31B). Kimi K3, Grok 4.5, ThinkingMachines-Inkling, and Nemotron-3-Nano-Omni each stand in as a single representative for a vendor we could not cover with a pair. Table 9 gives the short label used in tables and figures, the full model name, and the exact identifier called for every agent in the main roster.

Table 9: Model identifiers as called. Short labels are the names used in tables and figures. Nemotron-3-Nano-Omni was served through OpenRouter's free tier, which may route to a different provider or quantization than the paid variant.

| Short label | Model | Provider | Model identifier |
|---|---|---|---|
| gpt-5.5 | GPT-5.5 | OpenAI | `gpt-5.5` |
| gpt-4o | GPT-4o | OpenAI | `gpt-4o` |
| gpt-4o-mini | GPT-4o-mini | OpenAI | `gpt-4o-mini` |
| cl-opus-4.6 | Claude Opus 4.6 | OpenRouter | `anthropic/claude-opus-4.6` |
| cl-sonnet-4.6 | Claude Sonnet 4.6 | OpenRouter | `anthropic/claude-sonnet-4.6` |
| gemini-3.1-pro | Gemini 3.1 Pro Preview | OpenRouter | `google/gemini-3.1-pro-preview` |
| gemma-4-31b | Gemma-4-31B | OpenRouter | `google/gemma-4-31b-it` |
| grok-4.5 | Grok 4.5 | OpenRouter | `x-ai/grok-4.5` |
| kimi-k3 | Kimi K3 | OpenRouter | `moonshotai/kimi-k3` |
| qwen3.6-plus | Qwen3.6-Plus | OpenRouter | `qwen/qwen3.6-plus` |
| qwen3-vl-32b | Qwen3-VL-32B-Instruct | OpenRouter | `qwen/qwen3-vl-32b-instruct` |
| tm-inkling | ThinkingMachines-Inkling | OpenRouter | `thinkingmachines/inkling` |
| nemotron-3-nano | Nemotron-3-Nano-Omni | OpenRouter | `nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free` |

# C AGENT PROMPT TEMPLATE AND TOOL SCHEMA

This appendix gives the verbatim prompt and tool-calling schema sent to every evaluated LLM/VLM agent (§3.5), reproduced directly from the benchmark's agent-interface code. Curly braces mark template fields filled in per episode, and the italicized note after each field gives its source. The photo line and defect-tag vocabulary are included only when the item carries a real image (`include_image=True`), and the vocabulary is fixed across all episodes and mappings.

## C.1 SYSTEM PROMPT

```
You are negotiating the price of a used item over chat,
alternating offers with a counterpart.  You are the {role}.
(BUYER or SELLER) You must call the negotiate_action tool
exactly once per turn to record your move -- OFFER a price,
ACCEPT the counterpart's last offer, or REJECT and walk
away.
```

## C.2 USER PROMPT TEMPLATE

```
Item for sale:
- Category:  {item.category}
- Title:  {item.title}
- Description:  {item.description}
- Listing price:  ${item.listing_price}
[if include_image:  the item photo is attached as an image
content block, and the following line is appended:]
A photo of the item is attached above -- look it over before
deciding your move.  If you notice a defect, cite it in
cited_defect_ids using its type, one of:  scratch, dent,
rust, crack, tear, stain, missing_part.
Your role:  {role}
Your reservation price (the most you will pay / least you
will accept):  ${r_A}
Price bounds for this negotiation:  ${p_min} - ${p_max}
Round {k} of {K} (the negotiation ends in disagreement if no
deal is reached by round {K}).

Negotiation so far:
{formatted transcript, or ``(No one has made an offer yet)''
on round 1}

What is your move this round?
```

Each line of the transcript is rendered as, e.g., `Round 2 [Counterpart] OFFER $41.44 -- "Thank you for your continued flexibility;..."`, with the speaker label written from the listener's own perspective (`You` / `Counterpart`), and the counterpart's hidden type $t_B = (r_B, \kappa_B, \eta_B)$ and the item's ground-truth defect list are never included in either prompt, enforcing the information-hiding contract of §3.

## C.3 TOOL SCHEMA (FORCED FUNCTION CALL)

```
{
  "name": "negotiate_action",
  "description": "Decide the action to take
    in this negotiation turn.",
  "parameters": {
    "type": "object",
    "properties": {
      "decision": {
        "type": "string",
        "enum": ["OFFER", "ACCEPT", "REJECT"]
      },
      "price": {
```

```
            "type": ["number", "null"],
            "description": "Required only when
              decision is OFFER. Null otherwise."
          },
          "message": {
            "type": "string",
            "description": "The message sent to
              the counterpart."
          },
          "cited_defect_ids": {
            "type": "array",
            "items": {"type": "string"},
            "description": "Defect type(s) you
              actually observed in the photo and
              are citing this message, using the
              defect type list given in the
              prompt. Empty array if none."
          }
        },
        "required": ["decision", "price",
          "message", "cited_defect_ids"],
        "additionalProperties": false
      }
    }
```

The API call sets `tool_choice` to force exactly this function on every turn, so the agent's move is always structured output rather than free text requiring a separate parser. If a response fails to parse (malformed JSON, missing required field, or an unusable price on an `OFFER`), the turn falls back to a fixed rule (accept if the counterpart's last offer is individually rational, otherwise reject) rather than crashing the episode. We do not separately report how often this fallback fires per agent in this paper, and doing so is a natural addition for future work on protocol-compliance diagnostics.

## D Defect-Image Generation and QA Prompts

This appendix gives the verbatim prompts behind the dataset-construction pipeline of §3.6. They are the single-call generation prompt sent to Gemini 2.5 Flash Image for every synthesized defect, and the blind QA-gate prompt used to accept or reject the result. Curly braces mark fields filled in per item, and {`types`} is the compatibility-filtered list of defect types allowed for that item's category and quadrant family (§3.6).

### D.1 Image Generation Prompt

```
You are creating a synthetic damage photo for a research
benchmark on visual evidence in negotiation.

Here is a photo of a {category} item listed for sale.  Add
exactly ONE realistic defect to it, choosing a type from:
{types}.

Requirements:
- [visibility requirement, one of two variants below]
- [alignment requirement, one of four variants below]
- Choose a part of the item that is clearly, prominently,
and unambiguously visible in THIS SPECIFIC photo.  Do NOT
use a part that is absent, blurry, tiny, far away, or only
visible through a reflection/window.

First output exactly two lines of text:
DEFECT_TYPE: <the type you chose, one of {types}>
DESCRIPTION: <one sentence describing exactly what you added
and on which specific visible part, written as ground-truth
fact>
```

```
Then generate the edited photo with that defect added.  This
must be a LOCALIZED, MINIMAL edit, not a full repaint or
re-imagining of the scene:  do not alter the background,
walls, floor, or surroundings in any way; do not change
the state, pose, position, count, or configuration of the
item or of any other object; do not introduce anything that
did not already exist in the original photo, other than
the texture/damage of the single defect itself.  The ONLY
difference between the original and edited photo should
be the damage texture of the one defect described above,
confined strictly to the part of the item you named.
```

**Visibility variants** (*one* is substituted, per the item's visibility label). Obvious: "The defect must be OBVIOUS – large, high-contrast, and immediately noticeable at a glance. Render it strongly and unmistakably; do not render it faintly." Subtle: "The defect must be SUBTLE – small and easy to miss at a glance, blending into the item's existing color/texture. SUBTLE DOES NOT MEAN INVISIBLE OR ABSENT: someone who looks carefully, specifically checking for damage, MUST be able to clearly find and confirm it . . . render it with enough real texture/contrast . . . that close inspection unambiguously confirms it – just make it small and low-key rather than prominent."

**Alignment variants** (*one* is substituted, per the item's alignment label). Obvious-Aligned: "It should look exactly as serious as it really is: obviously damaged, and it genuinely is a serious problem." Subtle-Aligned: "It should look exactly as minor as it really is: barely noticeable, and it genuinely is only a minor cosmetic issue." Obvious-Misaligned: "Even though it LOOKS dramatic and serious, describe it as being only cosmetic/superficial and NOT a real functional or structural problem." Subtle-Misaligned: "Even though it looks minor/barely noticeable, describe it as actually indicating a real underlying structural or functional problem beneath the surface."

### D.2 Blind QA-Gate Prompt

```
Look closely at this photo of an item for sale.  Does it
show any visible damage, defects, or flaws (e.g. scratch,
dent, crack, tear, stain, rust, missing part)?  If yes,
describe exactly what you see and where.  If no, say NONE.
```

This call is shown only the edited photo, with no defect description, location, or quadrant label, and passes on any non-`NONE` response (§3.6).

### D.3 Pilot Failure Modes

An early version of the generation prompt (§3.6) pre-specified, per category, which part of the item to damage, and this once assigned a defect to a side mirror in a photo taken from inside the vehicle, where no mirror was visible, so the prompt was changed to let the model choose its own damageable part from *that specific photo* instead (Appendix D). Three further failure modes were observed and fixed by prompt constraint. The background/walls were sometimes subtly redrawn (fixed by requiring a pixel-identical background and a strictly localized edit), object state occasionally changed, e.g. a closed laptop shown open (fixed by explicitly forbidding state, pose, or configuration changes), and "Subtle" was sometimes rendered as effectively invisible (fixed by clarifying that subtle $\neq$ invisible, so it must remain blind-confirmable on close inspection).

At the QA gate (§3.6), one representative case illustrates a failure mode the gate is specifically designed to catch. An electronics item (rigid, glossy packaging) never passed blind verification for a crack defect across repeated prompt strengthening and was correctly discarded. Some materials structurally resist a renderable-yet-real defect, and the QA gate surfaces this rather than silently shipping an item whose "ground-truth" defect a VLM agent could never plausibly detect either. At the category level, the car category's discard rate was notably higher than the other three, and we hypothesize but have not confirmed that this relates to reflective paint, glass, and leather surfaces resisting a renderable-yet-verifiable defect more often than matte bike, furniture, and electronics surfaces do.

# E QUALITATIVE EXAMPLES

The diagnostics in §5–§7 are aggregate statistics, and this section grounds them in raw episodes so the reader can see what a citation, a miss, or a fabrication actually looks like in the transcript. All examples below are real, unedited episodes (mapping B) with the item's listing price, ground-truth defect, and price impact as generated for the benchmark. Nothing here is constructed for exposition.

## E.1 EXAMPLE NEGOTIATIONS

**Detection succeeds vs. fails on the same item.** Both episodes below place the agent as SELLER of the same "Marantz SR 780 receiver with remote" (listing $100), which has one subtle-misaligned defect, a hairline crack on the front panel, price impact = $25. GEMINI-3.1-PRO finds it, and GPT-5.5 does not.

**gemini-3.1-pro (detects)**

```
T1 agent OFFER $100.00 -- ``This Marantz receiver is a great
unit, putting out 80 watts per channel.  It comes with the
remote as well.''
T2 counterpart OFFER $52.41
T3 agent OFFER $80.00 -- ``While there is a crack on the
front display panel, the receiver works perfectly and
includes the remote, so $52.41 is a bit too low.  I can
come down to $80.00.''    [cites:  crack]
T4 counterpart OFFER $71.58
T5 agent OFFER $75.00 -- ``You drive a hard bargain!  If you
can meet me at $75.00, we have a deal.''
T6 counterpart OFFER $71.58
T7 agent ACCEPT $71.58
```

**gpt-5.5 (misses)**

```
T1 agent OFFER $100.00 -- ``It's a solid unit with 80 watts
per channel, so I think the listing price is fair.''
T2 counterpart OFFER $52.41
T3 agent OFFER $85.00 -- ``I can't do $52.41, especially
with the remote included and this being a solid Marantz
receiver.''    [cites:  none]
T4 counterpart OFFER $71.65
T5 agent OFFER $78.00
T6 counterpart OFFER $71.65
T7 agent ACCEPT $71.65
```

Note the two episodes close within $0.07 of each other. GEMINI names the crack and cites it as its stated reason for a real concession ($100→$80), while GPT-5.5 never mentions the defect at all and describes the unit as "solid." A firm counterpart absorbed most of the difference in this instance, which is why we score detection and valuation (§7) directly from the transcript rather than from final price alone, which would have called these two episodes equivalent.

**Detection without valuation.** GROK-4.5, as SELLER of "Four Speakers and Stereo Amp" (listing $50, subtle-misaligned crack on a speaker grille, price impact = $12.50), does cite the defect, but the concession attached to that citation is $1.50, far short of the $12.50 the defect is actually worth.

```
T1 counterpart OFFER $27.34
T2 agent OFFER $45.00 -- ``$27 is too low for this working
amp and four speakers...  I can come down to $45.''
[cites:  none]
T3 counterpart OFFER $42.27
T4 agent OFFER $43.50 -- ``I do see a visible crack on one
of the speaker grilles in the photo.  Still, everything
f[unctions]...''    [cites:  crack]
T5 counterpart OFFER $42.27
T6 agent OFFER $43.00 -- ``I already noted the crack on the
speaker grille...  let's split the dif[ference].''    [cites:
```

```
crack]
T7 counterpart OFFER $42.27
T8 agent ACCEPT $42.27 -- ``I already accounted for the
visible crack on the speaker grille.''    [cites:  crack]
```

The agent names the defect three separate times and explicitly claims to have "accounted for" it, but its own offer moves only \$1.50 (\$45.00→\$43.50) at the point of citation, a 6% discount on a defect worth 25% of the listing price. This is the Detection-without-Valuation failure mode central to §7 made concrete, since the agent's language asserts calibration that its numbers do not deliver.

**Fabrication over the real defect.** GPT-5.5, as BUYER of a "JVC Digital Theater System TH-C5" (listing \$299) with one obvious-aligned defect, a large, unmistakable dent on the center speaker's faceplate (price impact = \$74.75), never mentions the dent at all. Instead it invents a defect that is not in the image.

```
T1 counterpart OFFER $1433.04
T2 agent OFFER $150.00 -- ``Since the remote is missing and
I'd need to source one, I could offer $150 for the system if
it's working as described.''    [cites:  missing_part]
T3 counterpart OFFER $1271.91
T4 agent OFFER $190.00 -- ``Given the missing remote
and that I'd need to test/source parts...''    [cites:
missing_part]
⋮
T8 agent OFFER $248.00 -- ``Since the remote is missing, my
absolute best and final offer is $248 cash...''    [cites:
missing_part]
T9 counterpart ACCEPT $248.00
```

The dent is the obvious-aligned quadrant, the easiest case in the benchmark's own design (§3.2), yet the agent leverages an unverifiable claim ("the remote is missing") instead of the defect actually in front of it, across every citing turn of the episode. This is the fabrication failure mode of §7, a plausible-sounding, negotiation-relevant claim that has no basis in the item shown.

### E.2 Example Defect Images

Figure 7 shows one original/edited pair per quadrant, the same four items referenced in the dialogue examples above (plus the Linksys router and corner drink center used nowhere else in the main text). Obvious defects are meant to be found by any agent that looks; subtle defects require close inspection but must remain confirmable (Appendix D gives the exact generation constraint). Misaligned defects are the harder valuation case regardless of visibility, since what is pictured does not, by construction, tell the agent how much it should matter.

## F Mapping Robustness: Calibration and Hallucination by Mapping

Figure 4 (§5) gives $SE^+$ by mapping and by quadrant. Figure 8 gives the same breakdown for the two diagnostics discussed but not tabulated in the main text. These are severity-calibration $\rho$, the least stable of the three mapping-independent diagnostics, and hallucination rate, which is mapping-independent by construction (citations do not depend on which price-impact curve is in effect) and is included here as a stability sanity check.

Figure 9 gives the %-discordant-pairs breakdown behind the rank-order stability reported there, and Figure 10 gives severity-calibration $\rho$ separately for each of the four mappings.

## G Supporting Diagnostic Figures

Figures 14–15 give supporting detail for §5 and §7. Figures 11 and 12 give the per-agent detection and valuation breakdowns underlying the pooled numbers of §7, and Table 11 gives the per-quadrant grounding hit-rates behind them. Table 10 gives the full version of Table 2 (§5), including $AGR^+$, halluc. rate, and Util. Prox.

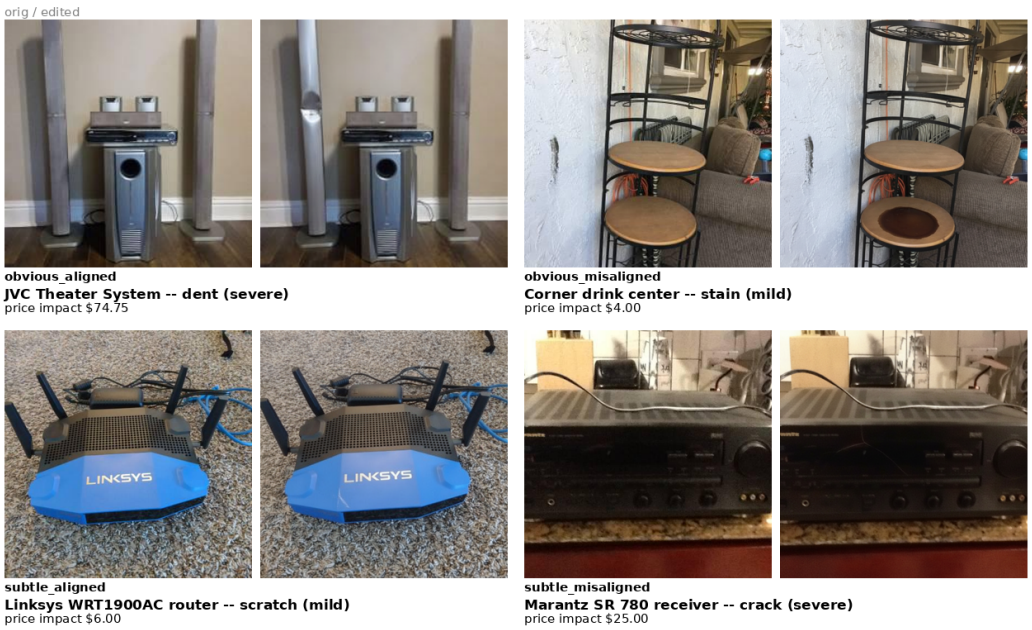



Figure 7: One real original/edited pair per quadrant, with the ground-truth defect and price impact used in scoring. The JVC and Marantz items are the same episodes shown in Appendix E.1.

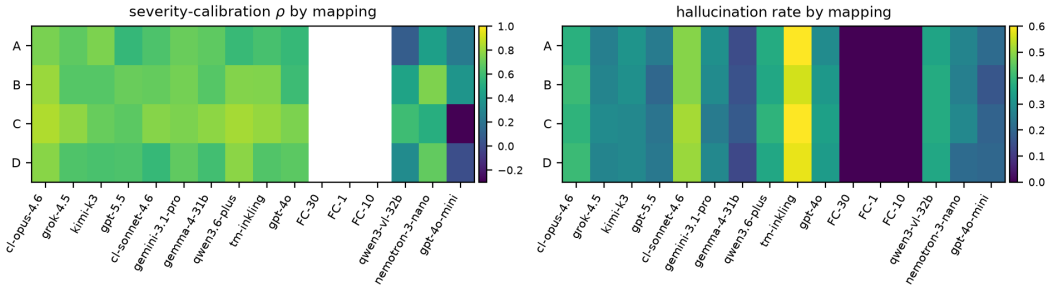



Figure 8: Severity-calibration $\rho$ (left) and hallucination rate (right) by agent and mapping.

### G.1 Error Analysis by Defect Type

The results above are pooled across defect types (scratch, dent, crack, stain, rust, tear). This subsection breaks three of the paper's headline diagnostics down by type instead, supporting §7. All three use the moderate (B) mapping with all 13 LLM agents pooled and no episodes excluded.

Table 12 gives citation recall by defect type, split by visibility, since pooling across type conflates the two. For example, crack looks mediocre pooled (0.505) only because 71% of crack instances happen to be Subtle in this pool, not because crack itself is hard, and at Obvious visibility it is the single easiest type (0.923). Two types stand out once visibility is controlled for. Dent is the hardest type to catch at Subtle visibility (0.175, worst cell in the table), and scratch is uniquely hard even when Obvious (0.407, the only type below 0.46 at that visibility level).

Table 13 shows what agents fabricate when they hallucinate a defect (§7): 31.8% of episodes (414/1,300) contain at least one fabricated citation, and the fabricated type is not uniform across item categories. Scratch is the single most common fabrication overall (39.8% of all fabricated citations) and dominates specifically for cars (68% of that category's fabrications) and bikes (48%), furniture fabrications are led by crack instead (39%), and electronics is the only category where "missing_part", not a visually groundable defect type at all, appears at meaningful share (14%). This matches the qualitative example in Appendix E.1 where an agent invents a missing remote rather than citing the real, obvious dent in the photo. Fabrication therefore tracks category-typical defect *priors* (cars get scratched, furniture cracks) more than it tracks anything specific to the image shown.

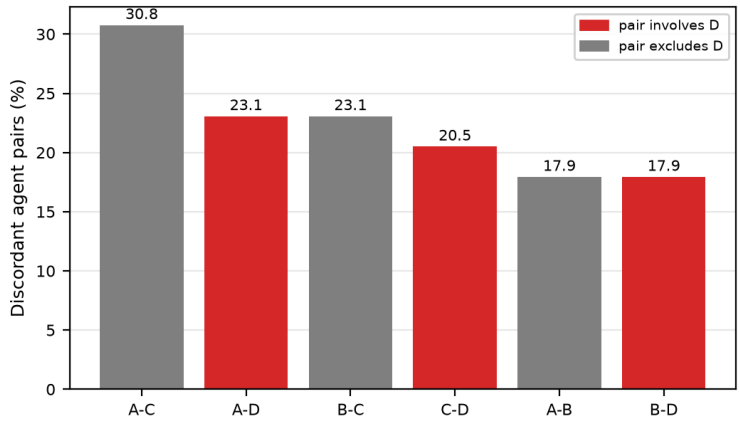



Figure 9: % of the 13 LLM agents' $\binom{13}{2} = 78$ pairs that swap relative $SE^{+}$ rank, one bar per pair of price-impact mappings.

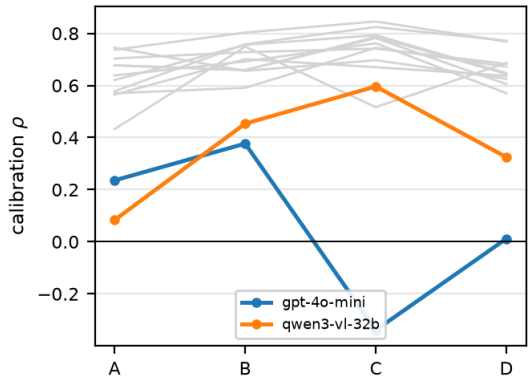



Figure 10: Severity-calibration $\rho$ across the four price mappings, with gray lines for the other 11 LLM agents.

Table 14 classifies every single-citation OFFER turn by the ratio of the actual price concession to the cited defect's true price impact (the same turns severity-calibration $\rho$ is computed from, §7, recast here as a direction rather than a rank correlation): under-adjusted (ratio $< 2/3$), roughly calibrated ($2/3$–$3/2$), or over-adjusted (ratio $> 3/2$). The distribution is heavily one-sided, with under-adjustment outnumbering over-adjustment more than 30 to 1 (median ratio $0.40$, so the typical concession is well under half of what the defect is actually worth). This is not specific to the Misaligned cells where under-pricing is the documented failure mode (§7), since Aligned citations under-adjust essentially as often. Agents that do cite a real defect almost never overcorrect for it, so the risk this benchmark's valuation metrics are built to catch is one-directional in practice. This ratio-based view is also the more conservative reading of the same turns. The grounding hit-rate of §7 scores them against a $\pm 0.05 \times R$ tolerance that is wide relative to the price impacts involved, since $R$ is a category-level price range rather than an item-level one. The ratio direction reported here and the bias contrast of Figure 12 reach the same Subtle-Misaligned conclusion without relying on that tolerance.

## H Role-Conditional Results

§5 shows a structural asymmetry between playing buyer and seller and recommends reading role-pooled comparisons with caution. The mechanism is that the price lower bound is fixed at $p_{\min} = 0$ while $p_{\max}$ is set well above listing price, so a counterpart playing seller has far more room to open aggressively than one playing buyer. Pooling all agents, the counterpart's first-offer deviation from its own reservation value is $0.362$ as seller versus $0.060$ as buyer, a $\sim 6\times$ asymmetry that is a property of the environment, not of any evaluated agent. This confound is not neutral: quadrant assignment is not role-balanced (Obvious-Aligned is $60.0\%$ buyer-role episodes vs. Subtle-Misaligned at $35.3\%$), and the Subtle-Misaligned caught-vs-missed gap (Eq. 2, an appendix-only diagnostic that §7 does not rely

Table 10: Image on/off ablation for the Qwen family under the moderate (B) mapping, 100 episodes/agent/condition, full version of Table 2. Util. Prox. is undefined (−) under No-Image since citations collapse to near zero.

| Agent | Condition | $SE^+$ | $AGR^+$ | cite-cov. | halluc. rate | Util. Prox. |
|---|---|---|---|---|---|---|
| qwen3-vl-8b | Image | 0.187 | 0.720 | 0.312 | 0.410 | 0.291 |
| | No-Image | 0.169 | 0.690 | 0.018 | 0.230 | – |
| | Δ | −0.018 | −0.030 | −0.294 | −0.180 | – |
| qwen3-vl-32b | Image | 0.152 | 0.840 | 0.418 | 0.370 | 0.324 |
| | No-Image | 0.138 | 0.810 | 0.029 | 0.210 | – |
| | Δ | −0.014 | −0.030 | −0.389 | −0.160 | – |
| qwen3.6-plus | Image | 0.503 | 0.940 | 0.557 | 0.360 | 0.498 |
| | No-Image | 0.471 | 0.920 | 0.041 | 0.180 | – |
| | Δ | −0.032 | −0.020 | −0.516 | −0.180 | – |

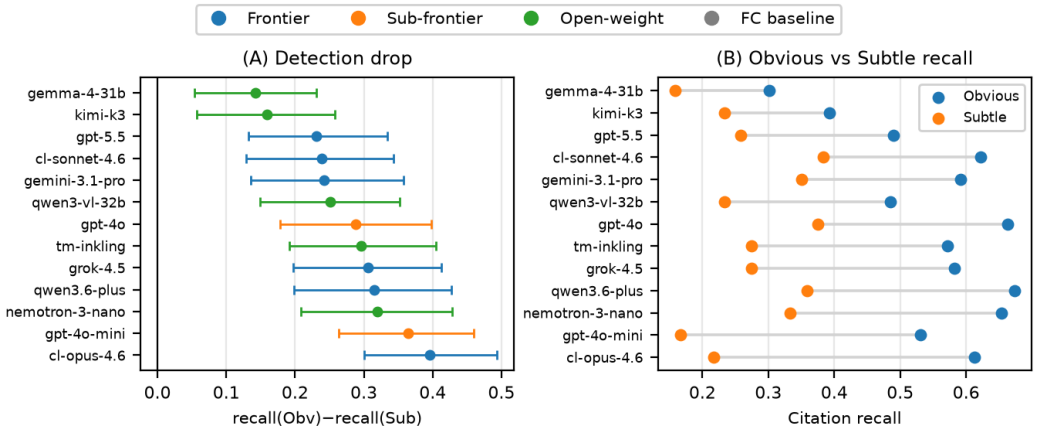



Figure 11: Detection (95% bootstrap CI, colored by capability tier). (A) Detection drop, recall(Obvious) − recall(Subtle). (B) The underlying Obvious-vs-Subtle citation recall each agent's detection drop (A) is computed from.

on) flips sign once role is controlled for (naive $-0.018$ to role-controlled $+0.038$; Figure 17 below). Rather than only asserting this, we give the role-conditional version of every headline table in the paper here, so pooled numbers in the main text can be checked directly against their BUYER-only and SELLER-only counterparts. Table 15/16 split Table 3 (all nine headline diagnostics under mapping B, $n = 50$ episodes/agent/role, half of the pooled 100), and Table 18 splits Table 17's fabricated-evidence effect. 12 of the 13 LLM agents' $SE^+$ is higher as seller than as buyer (GPT-4o-mini is the exception), consistent with Figure 16. The visual-evidence columns (cite-prec. through calib. $\rho$) are noisier per role than pooled ($n = 50$ vs. 100) and should be read as directional, not as independently significant findings. Figure 17 makes the pooling confound concrete for the Subtle-Misaligned gap specifically. It shows the caught-vs-missed $CSE^+$ gap per quadrant, computed naively (role-pooled) vs. role-controlled (averaging a BUYER-only and a SELLER-only estimate).

The remaining figures in this appendix give the role-conditional version of every other figure in the main text. Figure 16 is the role-split counterpart to Figure 5 (§6).

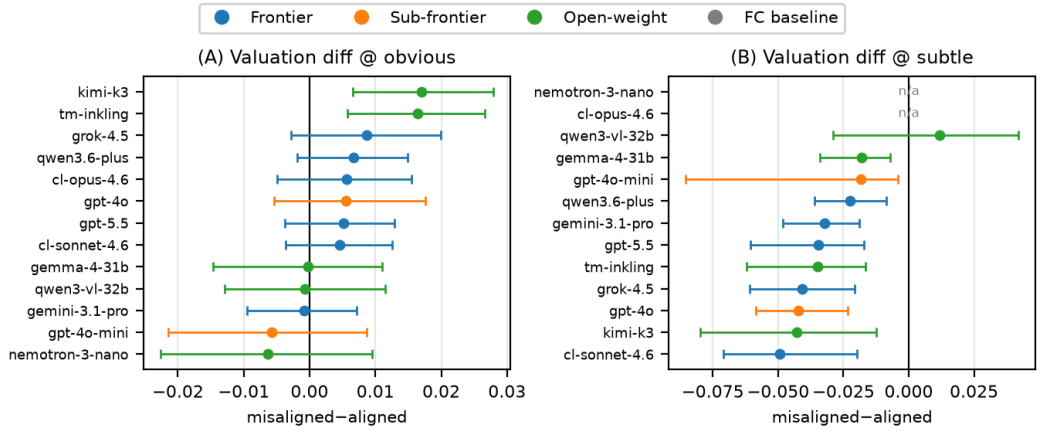



Figure 12: Valuation (95% bootstrap CI, colored by capability tier). (A) Valuation diff at Obvious visibility (Misaligned − Aligned). (B) Valuation diff at Subtle visibility; `n/a` = no single-citation data.

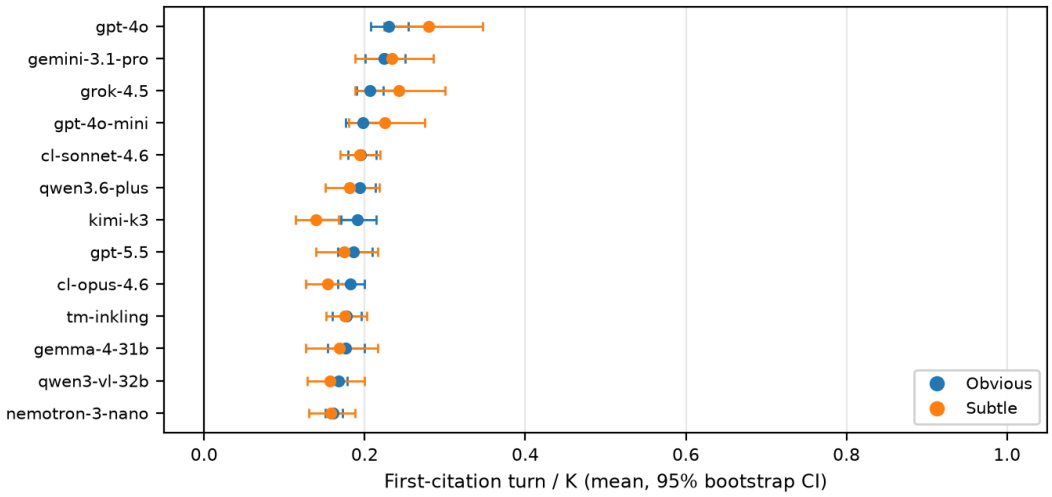



Figure 13: Mean turn of first real-defect citation, normalized by the episode's round budget $K$, Obvious vs. Subtle (episode-level 95% bootstrap CI, $B = 2{,}000$).

Table 11: Grounding hit-rate (fraction of single-citation OFFER turns whose price concession lands within $0.05 \times R$ of the cited defect's true price impact), by agent and quadrant, pooled across all four mappings, with no episodes excluded. $n$ in parentheses.

| Agent | Obv.-Align. | Sub.-Align. | Obv.-Misalign. | Sub.-Misalign. |
|---|---|---|---|---|
| cl-opus-4.6 | 0.935 (62) | 1.000 (50) | 0.911 (112) | n/a (0) |
| grok-4.5 | 0.857 (63) | 1.000 (23) | 0.924 (132) | 0.625 (32) |
| kimi-k3 | 0.850 (60) | 1.000 (26) | 1.000 (77) | 0.615 (13) |
| gpt-5.5 | 0.982 (56) | 1.000 (27) | 0.964 (110) | 0.733 (15) |
| cl-sonnet-4.6 | 0.930 (71) | 1.000 (36) | 0.904 (115) | 0.565 (23) |
| gemini-3.1-pro | 0.957 (70) | 1.000 (20) | 0.907 (107) | 0.780 (41) |
| gemma-4-31b | 1.000 (41) | 1.000 (6) | 0.895 (38) | 0.929 (14) |
| qwen3.6-plus | 0.944 (72) | 1.000 (14) | 0.935 (93) | 0.800 (15) |
| tm-inkling | 0.921 (38) | 1.000 (16) | 0.989 (87) | 0.727 (11) |
| gpt-4o | 0.883 (77) | 1.000 (15) | 0.877 (138) | 0.611 (36) |
| qwen3-vl-32b | 1.000 (22) | 0.750 (12) | 0.864 (66) | 1.000 (6) |
| nemotron-3-nano | 1.000 (8) | n/a (0) | 0.900 (20) | 1.000 (13) |
| gpt-4o-mini | 1.000 (38) | 1.000 (3) | 0.784 (51) | 0.857 (7) |

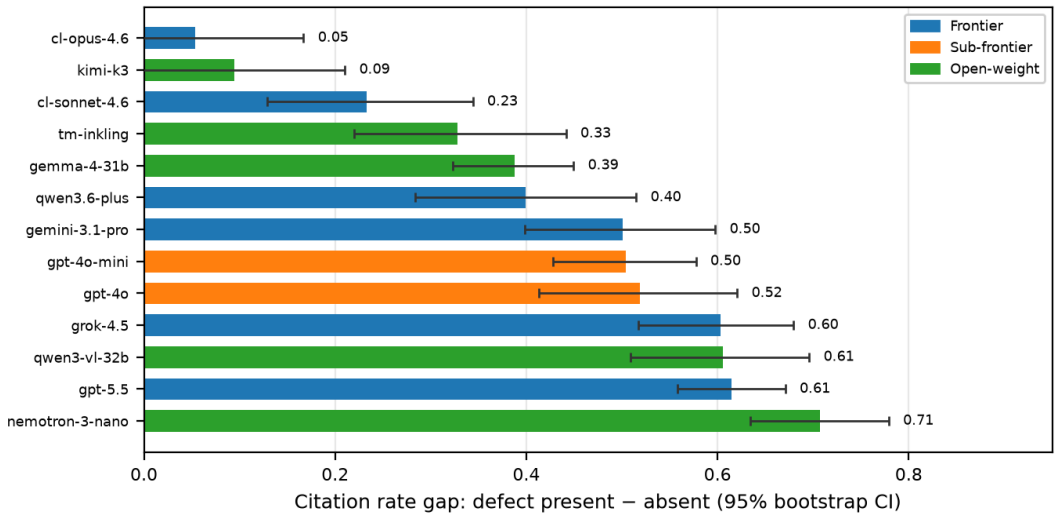



Figure 14: Citation rate on items with a real defect vs. items with none, sorted by gap (95% episode-level bootstrap CI, $B = 2{,}000$).

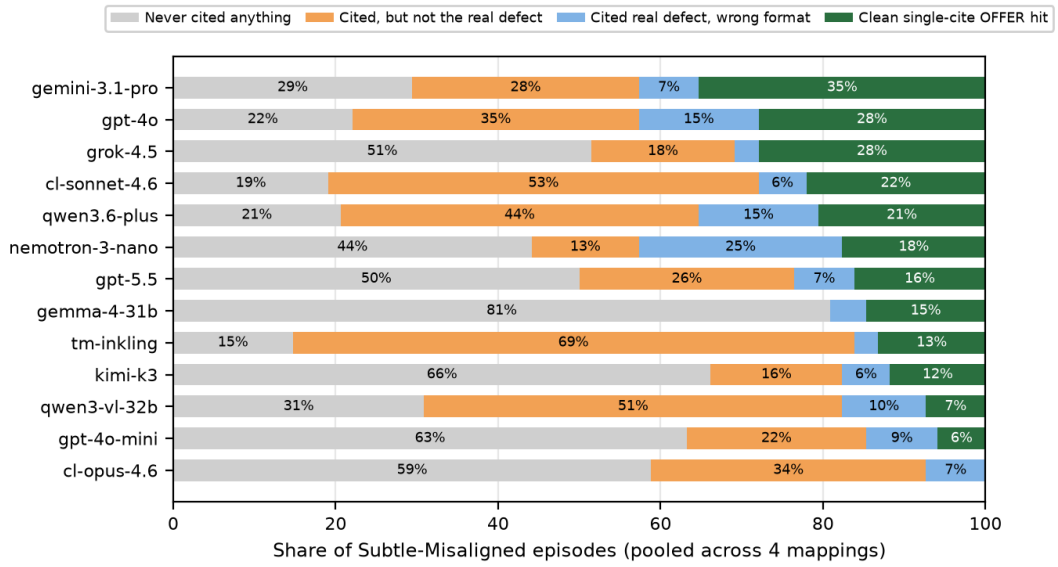



Figure 15: Where Subtle-Misaligned episodes are lost, per agent ($n = 68$/agent, pooled across 4 mappings). Segments, left to right, are never cited anything, cited something but not the real defect, cited the real defect but not as a single-citation OFFER turn (the format the valuation-bias metric of §7 requires), and a clean single-citation OFFER-turn hit. Sorted by hit rate (green, right).

Table 12: Citation recall by defect type and visibility, pooled across 13 LLM agents (mapping B). $n$ = ground-truth defect instances of that (type, visibility). Stain has no Subtle instances in this pool.

| Defect type | Obv. recall ($n$) | Sub. recall ($n$) | Pooled recall ($n$) |
|---|---|---|---|
| crack | 0.923 (26) | 0.338 (65) | 0.505 (91) |
| stain | 0.712 (52) | n/a (0) | 0.712 (52) |
| rust | 0.654 (130) | 0.308 (26) | 0.596 (156) |
| dent | 0.585 (195) | 0.175 (143) | 0.411 (338) |
| tear | 0.462 (13) | 0.231 (13) | 0.346 (26) |
| scratch | 0.407 (221) | 0.350 (143) | 0.385 (364) |

Table 13: Fabricated (non-existent) defect type, by item category, as share of that category's fabricated citations. $n$ = fabricated-citation instances in that category. Types under 5% combined omitted per row.

| Category | $n$ | Top fabricated types (share) |
|---|---|---|
| car | 109 | scratch 68%, dent 8%, rust 8%, crack 7% |
| bike | 61 | scratch 48%, rust 26%, missing_part 15%, stain 11% |
| furniture | 156 | crack 39%, scratch 26%, stain 12%, tear 11% |
| electronics | 151 | scratch 30%, tear 26%, crack 23%, missing_part 14% |

Table 14: Direction of price adjustment on single-citation OFFER turns, by the cited defect's alignment ($n = 555$ total citing turns, mapping B, 13 LLM agents pooled).

| Alignment | $n$ | Under-adj. ($<2/3$) | Calibrated | Over-adj. ($>3/2$) |
|---|---|---|---|---|
| Aligned | 223 | 76.7% | 22.9% | 0.4% |
| Misaligned | 332 | 68.4% | 28.3% | 3.3% |
| Pooled | 555 | 71.7% | 26.1% | 2.2% |

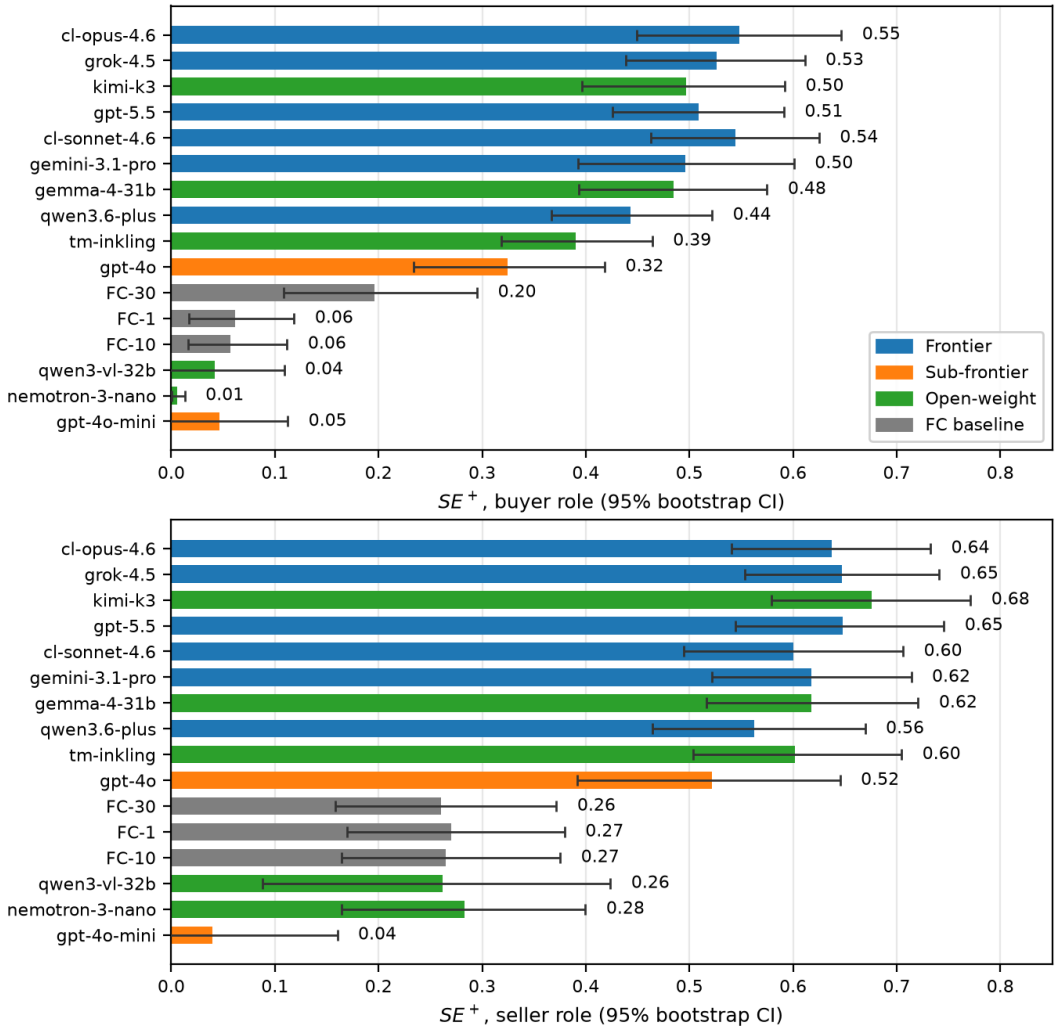



Figure 16: $SE^+$ by agent, BUYER (left) vs. SELLER (right) role (episode-level 95% bootstrap CI, $B = 2{,}000$), with the same agent order in both panels (pooled mapping B $SE^+$, descending), colored by capability tier. Every agent but GPT-4o-mini scores lower as buyer. Bar-end numbers are point estimates.

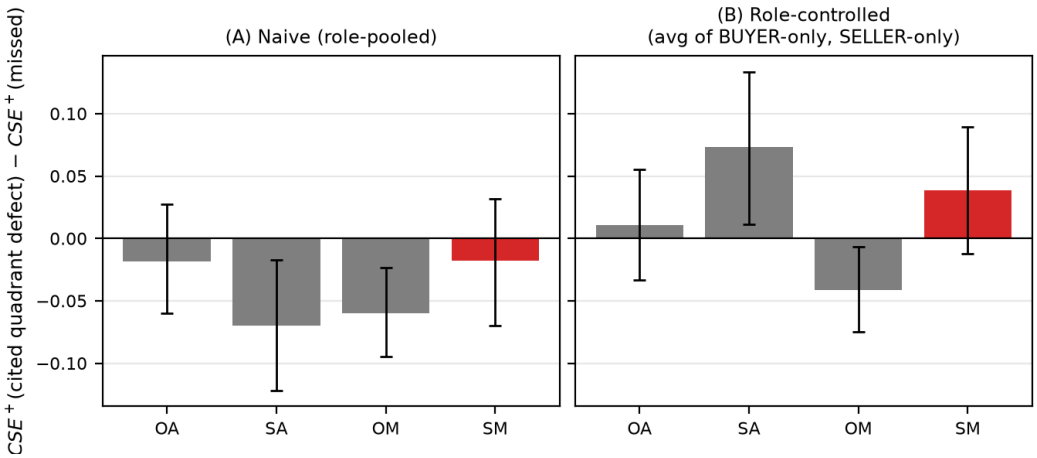



Figure 17: Caught-vs-missed $CSE^+$ gap per quadrant (Eq. 2, pooled over 13 LLM agents, protocol-violation episodes excluded, episode-level 95% bootstrap CI, $B = 2{,}000$), computed naively (role-pooled, panel A) vs. role-controlled (average of a BUYER-only and a SELLER-only estimate, panel B). OA/SA/OM/SM = Obvious/Subtle-Aligned/Misaligned.

Table 15: Table 3's nine headline diagnostics, BUYER-role episodes only (mapping B, $n = 50$/agent). Same column definitions and agent order as Table 3; no bold/italic ranking here given the halved sample.

| | Negotiation outcome | | | | Visual evidence use | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Agent | $SE^+$ | $AGR^+$ | $CSE^+$ | CritViol% | cite-prec. | cite-cov. | halluc. rate | Util. Prox. | calib. $\rho$ |
| cl-opus-4.6 | 0.548 | 0.760 | 0.722 | 0.000 | 0.401 | 0.658 | 0.680 | 0.710 | 0.883 |
| grok-4.5 | 0.526 | 0.820 | 0.642 | 0.040 | 0.537 | 0.579 | 0.380 | 0.687 | 0.637 |
| kimi-k3 | 0.497 | 0.700 | 0.710 | 0.000 | 0.399 | 0.579 | 0.580 | 0.656 | 0.801 |
| gpt-5.5 | 0.509 | 0.920 | 0.553 | 0.060 | 0.616 | 0.579 | 0.280 | 0.649 | 0.710 |
| cl-sonnet-4.6 | 0.544 | 0.880 | 0.619 | 0.020 | 0.418 | 0.658 | 0.640 | 0.631 | 0.617 |
| gemini-3.1-pro | 0.496 | 0.720 | 0.689 | 0.000 | 0.560 | 0.684 | 0.420 | 0.673 | 0.679 |
| gemma-4-31b | 0.484 | 0.760 | 0.637 | 0.000 | 0.569 | 0.447 | 0.260 | 0.548 | 0.637 |
| qwen3.6-plus | 0.443 | 0.920 | 0.482 | 0.000 | 0.508 | 0.632 | 0.460 | 0.515 | 0.691 |
| tm-inkling | 0.391 | 0.940 | 0.416 | 0.020 | 0.401 | 0.553 | 0.580 | 0.602 | 0.648 |
| gpt-4o | 0.324 | 0.860 | 0.377 | 0.180 | 0.521 | 0.632 | 0.440 | 0.576 | 0.691 |
| FC-30 *(baseline)* | 0.196 | 0.480 | 0.408 | 0.000 | – | 0.000 | 0.000 | – | – |
| FC-1 *(baseline)* | 0.061 | 0.300 | 0.205 | 0.000 | – | 0.000 | 0.000 | – | – |
| FC-10 *(baseline)* | 0.057 | 0.300 | 0.190 | 0.000 | – | 0.000 | 0.000 | – | – |
| qwen3-vl-32b | 0.042 | 0.720 | 0.058 | 0.400 | 0.458 | 0.500 | 0.440 | 0.282 | 0.225 |
| nemotron-3-nano | 0.006 | 0.320 | 0.019 | 0.220 | 0.562 | 0.500 | 0.300 | 0.060 | – |
| gpt-4o-mini | 0.047 | 0.620 | 0.075 | 0.260 | 0.661 | 0.500 | 0.200 | 0.278 | 0.110 |

Table 16: Table 3's nine headline diagnostics, SELLER-role episodes only (mapping B, $n = 50$/agent).

| | Negotiation outcome | | | | Visual evidence use | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Agent | $SE^+$ | $AGR^+$ | $CSE^+$ | CritViol% | cite-prec. | cite-cov. | halluc. rate | Util. Prox. | calib. $\rho$ |
| cl-opus-4.6 | 0.637 | 0.960 | 0.664 | 0.040 | 0.611 | 0.268 | 0.140 | 0.667 | 0.659 |
| grok-4.5 | 0.647 | 0.960 | 0.674 | 0.020 | 0.667 | 0.366 | 0.160 | 0.697 | 0.728 |
| kimi-k3 | 0.676 | 0.920 | 0.735 | 0.000 | 0.667 | 0.049 | 0.020 | 0.667 | – |
| gpt-5.5 | 0.648 | 0.960 | 0.675 | 0.020 | 0.688 | 0.293 | 0.120 | 0.688 | 0.792 |
| cl-sonnet-4.6 | 0.600 | 0.920 | 0.652 | 0.060 | 0.515 | 0.439 | 0.340 | 0.688 | 0.825 |
| gemini-3.1-pro | 0.618 | 0.960 | 0.643 | 0.000 | 0.696 | 0.415 | 0.160 | 0.826 | 0.824 |
| gemma-4-31b | 0.617 | 0.920 | 0.671 | 0.020 | 0.333 | 0.024 | 0.040 | 0.500 | – |
| qwen3.6-plus | 0.563 | 0.960 | 0.586 | 0.040 | 0.611 | 0.488 | 0.260 | 0.474 | 0.781 |
| tm-inkling | 0.602 | 0.940 | 0.640 | 0.060 | 0.388 | 0.439 | 0.540 | 0.543 | 0.841 |
| gpt-4o | 0.522 | 0.880 | 0.593 | 0.160 | 0.646 | 0.415 | 0.200 | 0.752 | 0.615 |
| FC-30 *(baseline)* | 0.260 | 0.500 | 0.521 | 0.000 | – | 0.000 | 0.000 | – | – |
| FC-1 *(baseline)* | 0.270 | 0.520 | 0.520 | 0.000 | – | 0.000 | 0.000 | – | – |
| FC-10 *(baseline)* | 0.265 | 0.520 | 0.510 | 0.000 | – | 0.000 | 0.000 | – | – |
| qwen3-vl-32b | 0.262 | 0.960 | 0.273 | 0.380 | 0.483 | 0.341 | 0.300 | 0.376 | 0.544 |
| nemotron-3-nano | 0.283 | 0.860 | 0.329 | 0.260 | 0.672 | 0.537 | 0.220 | 0.230 | 0.749 |
| gpt-4o-mini | 0.040 | 0.580 | 0.068 | 0.460 | 0.688 | 0.293 | 0.120 | 0.138 | 0.651 |

Table 17: Effect of citing only fabricated (non-existent) defects, versus citing nothing, on negotiation outcome, full 13-agent roster (episode-level 95% bootstrap CI, $B = 2{,}000$, pooled across 4 mappings; abbreviated in §7, Table 5). $^*$ = uncorrected 95% CI excludes 0. Agents marked $^\dagger$ are the 2/13 that remain significant on $\Delta SE^+$ after Benjamini–Hochberg correction (§7). Episodes with a logged protocol violation are excluded. $n_{\text{fab}}/n_{\text{none}}$ = episodes citing only fabricated defects / citing nothing.

| Agent | $\Delta SE^+$ | $\Delta AGR^+$ | $\Delta CSE^+$ | $n_{\text{fab}}$ | $n_{\text{none}}$ |
|---|---|---|---|---|---|
| qwen3-vl-32b$^\dagger$ | $-0.174$ $[-0.28, -0.06]^*$ | $-0.063$ $[-0.13, 0.00]$ | $-0.157$ $[-0.27, -0.04]^*$ | 91 | 80 |
| gemini-3.1-pro$^\dagger$ | $-0.164$ $[-0.27, -0.06]^*$ | $-0.330$ $[-0.45, -0.21]^*$ | $0.076$ $[-0.01, 0.16]$ | 85 | 154 |
| cl-opus-4.6 | $-0.111$ $[-0.20, -0.02]^*$ | $-0.194$ $[-0.29, -0.10]^*$ | $0.029$ $[-0.05, 0.10]$ | 105 | 142 |
| kimi-k3 | $-0.107$ $[-0.20, -0.02]^*$ | $-0.108$ $[-0.22, -0.01]^*$ | $-0.040$ $[-0.11, 0.03]$ | 80 | 208 |
| gemma-4-31b | $-0.077$ $[-0.18, 0.03]$ | $-0.129$ $[-0.26, 0.00]^*$ | $0.012$ $[-0.07, 0.09]$ | 54 | 262 |
| grok-4.5 | $-0.053$ $[-0.14, 0.04]$ | $-0.094$ $[-0.19, -0.01]^*$ | $0.009$ $[-0.06, 0.08]$ | 73 | 171 |
| gpt-4o | $-0.042$ $[-0.15, 0.06]$ | $-0.126$ $[-0.23, -0.02]^*$ | $0.036$ $[-0.06, 0.13]$ | 70 | 100 |
| gpt-5.5 | $-0.036$ $[-0.13, 0.06]$ | $-0.056$ $[-0.13, 0.01]$ | $-0.004$ $[-0.10, 0.09]$ | 65 | 192 |
| gpt-4o-mini | $-0.002$ $[-0.09, 0.08]$ | $0.246$ $[0.08, 0.40]^*$ | $-0.151$ $[-0.27, -0.03]^*$ | 45 | 140 |
| qwen3.6-plus | $0.026$ $[-0.06, 0.11]$ | $-0.045$ $[-0.11, 0.02]$ | $0.054$ $[-0.04, 0.14]$ | 108 | 104 |
| nemotron-3-nano | $0.055$ $[-0.06, 0.17]$ | $0.104$ $[-0.05, 0.25]$ | $0.019$ $[-0.13, 0.17]$ | 69 | 96 |
| tm-inkling | $0.067$ $[-0.01, 0.15]$ | $-0.032$ $[-0.07, 0.02]$ | $0.086$ $[0.00, 0.17]^*$ | 185 | 59 |
| cl-sonnet-4.6 | $0.100$ $[-0.00, 0.21]$ | $0.031$ $[-0.05, 0.12]$ | $0.088$ $[-0.01, 0.19]$ | 157 | 65 |

Table 18: Table 17's fabricated-evidence effect, split by role (pooled across 4 mappings, episode-level 95% bootstrap CI, $B = 2{,}000$, protocol-violation episodes excluded as in Table 17). $n_{\text{fab}}/n_{\text{none}}$ sum to the counts of Table 17, so several per-role CIs are wide (e.g. Claude Opus 4.6/buyer, $n_{\text{none}} = 4$), and directions matter more than any single $^*$. $^*$ = 95% CI excludes 0. Cells with $n_{\text{fab}} < 3$ or $n_{\text{none}} < 3$ are marked –.

| Agent | Role | $\Delta SE^+$ | $\Delta AGR^+$ | $\Delta CSE^+$ | $n_{\text{fab}}$ | $n_{\text{none}}$ |
|---|---|---|---|---|---|---|
| qwen3-vl-32b | Buyer | $-0.107$ [$-0.21$, $0.01$] | $-0.107$ [$-0.22$, $0.00$] | $-0.091$ [$-0.20$, $0.02$] | 51 | 33 |
| | Seller | $-0.130$ [$-0.28$, $0.01$] | $-0.004$ [$-0.07$, $0.06$] | $-0.130$ [$-0.27$, $0.01$] | 40 | 47 |
| gemini-3.1-pro | Buyer | $-0.093$ [$-0.24$, $0.07$] | $-0.165$ [$-0.34$, $0.03$] | $+0.033$ [$-0.08$, $0.14$] | 65 | 34 |
| | Seller | $-0.165$ [$-0.36$, $0.02$] | $-0.342$ [$-0.55$, $-0.14$]$^*$ | $+0.076$ [$-0.10$, $0.23$] | 20 | 120 |
| cl-opus-4.6 | Buyer | $-0.359$ [$-0.45$, $-0.26$]$^*$ | $-0.293$ [$-0.38$, $-0.21$]$^*$ | $-0.144$ [$-0.22$, $-0.07$]$^*$ | 92 | 4 |
| | Seller | $-0.007$ [$-0.18$, $0.15$] | $+0.065$ [$0.03$, $0.11$]$^*$ | $-0.051$ [$-0.23$, $0.10$] | 13 | 138 |
| kimi-k3 | Buyer | $+0.155$ [$-0.01$, $0.32$] | $+0.200$ [$-0.03$, $0.43$] | $+0.034$ [$-0.08$, $0.16$] | 75 | 25 |
| | Seller | $-0.269$ [$-0.50$, $-0.04$]$^*$ | $-0.113$ [$-0.53$, $0.12$] | $-0.236$ [$-0.44$, $-0.04$]$^*$ | 5 | 183 |
| gemma-4-31b | Buyer | $+0.038$ [$-0.10$, $0.16$] | $-0.037$ [$-0.21$, $0.13$] | $+0.087$ [$-0.01$, $0.18$] | 47 | 85 |
| | Seller | $-0.000$ [$-0.13$, $0.13$] | $+0.085$ [$0.05$, $0.13$]$^*$ | $-0.058$ [$-0.19$, $0.06$] | 7 | 177 |
| grok-4.5 | Buyer | $-0.022$ [$-0.13$, $0.09$] | $-0.058$ [$-0.20$, $0.07$] | $+0.019$ [$-0.07$, $0.10$] | 57 | 52 |
| | Seller | $-0.031$ [$-0.20$, $0.14$] | $-0.020$ [$-0.16$, $0.08$] | $-0.019$ [$-0.18$, $0.14$] | 16 | 119 |
| gpt-4o | Buyer | $+0.114$ [$-0.01$, $0.24$] | $-0.151$ [$-0.28$, $-0.01$]$^*$ | $+0.207$ [$0.09$, $0.33$]$^*$ | 47 | 25 |
| | Seller | $-0.022$ [$-0.20$, $0.15$] | $-0.107$ [$-0.28$, $0.05$] | $+0.057$ [$-0.10$, $0.20$] | 23 | 75 |
| gpt-5.5 | Buyer | $-0.016$ [$-0.14$, $0.11$] | $-0.053$ [$-0.13$, $0.00$] | $+0.010$ [$-0.11$, $0.13$] | 38 | 67 |
| | Seller | $+0.005$ [$-0.14$, $0.15$] | $-0.079$ [$-0.21$, $0.02$] | $+0.061$ [$-0.07$, $0.19$] | 27 | 125 |
| gpt-4o-mini | Buyer | $-0.000$ [$-0.08$, $0.08$] | $+0.236$ [$0.04$, $0.43$]$^*$ | $-0.111$ [$-0.24$, $0.02$] | 39 | 65 |
| | Seller | $+0.274$ [$0.04$, $0.47$]$^*$ | $+0.380$ [$0.02$, $0.63$]$^*$ | $+0.099$ [$-0.06$, $0.26$] | 6 | 75 |
| qwen3.6-plus | Buyer | $+0.119$ [$-0.01$, $0.24$] | $+0.007$ [$-0.10$, $0.14$] | $+0.127$ [$0.01$, $0.24$]$^*$ | 69 | 32 |
| | Seller | $+0.027$ [$-0.12$, $0.17$] | $-0.063$ [$-0.15$, $0.01$] | $+0.068$ [$-0.07$, $0.21$] | 39 | 72 |
| nemotron-3-nano | Buyer | $+0.070$ [$0.01$, $0.13$]$^*$ | $+0.165$ [$-0.03$, $0.37$] | $+0.124$ [$-0.01$, $0.26$] | 39 | 48 |
| | Seller | $+0.097$ [$-0.09$, $0.28$] | $+0.113$ [$-0.00$, $0.23$] | $+0.042$ [$-0.15$, $0.22$] | 30 | 48 |
| tm-inkling | Buyer | $+0.006$ [$-0.10$, $0.12$] | $-0.015$ [$-0.09$, $0.11$] | $+0.013$ [$-0.08$, $0.12$] | 96 | 21 |
| | Seller | $+0.157$ [$0.04$, $0.27$]$^*$ | $-0.034$ [$-0.08$, $0.00$] | $+0.179$ [$0.06$, $0.30$]$^*$ | 89 | 38 |
| cl-sonnet-4.6 | Buyer | – | – | – | 99 | 2 |
| | Seller | $+0.167$ [$0.04$, $0.30$]$^*$ | $+0.061$ [$-0.02$, $0.14$] | $+0.136$ [$0.01$, $0.26$]$^*$ | 58 | 63 |

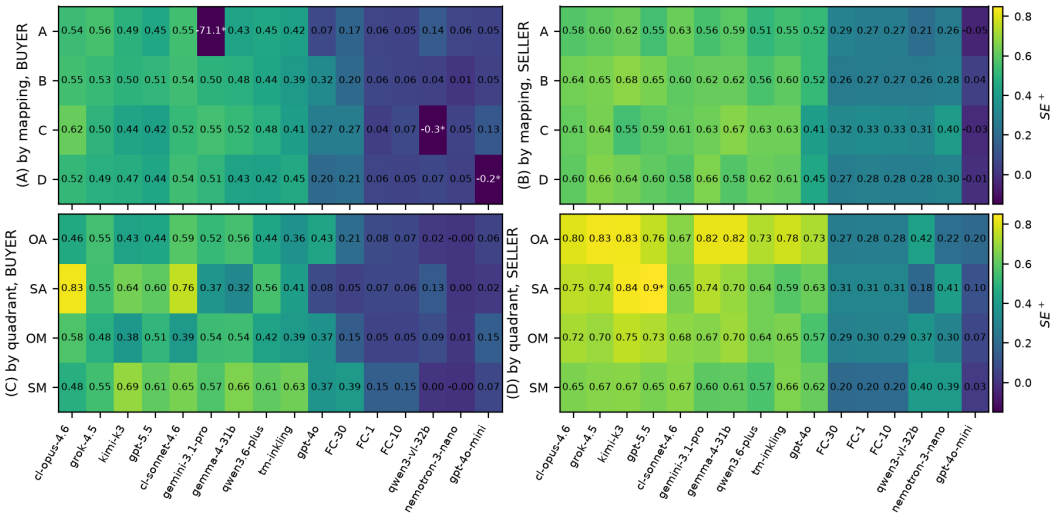



Figure 18: Role-split counterpart to Figure 4: $SE^+$ by mapping (A, B) and by quadrant under mapping B (C, D), BUYER vs. SELLER. Color scale clipped to $[-0.15, 0.85]$, off-scale cells marked $^*$ (e.g. panel A's Gemini 3.1 Pro Preview/mapping A cell reflects the same single wide-ranging outlier episode noted in §5, more extreme here since the role split roughly halves the per-cell episode count).

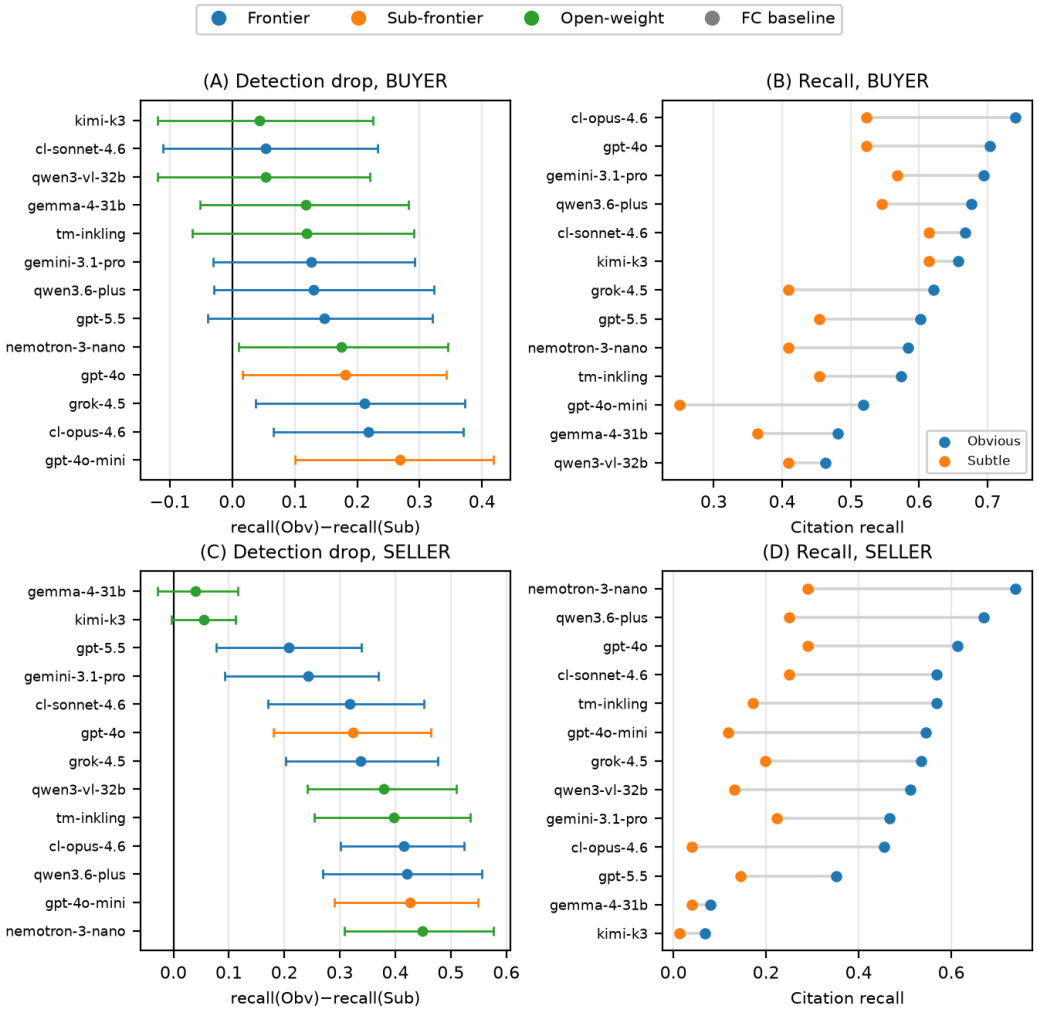



Figure 19: Role-split counterpart to Figure 11 (95% bootstrap CI, colored by capability tier).

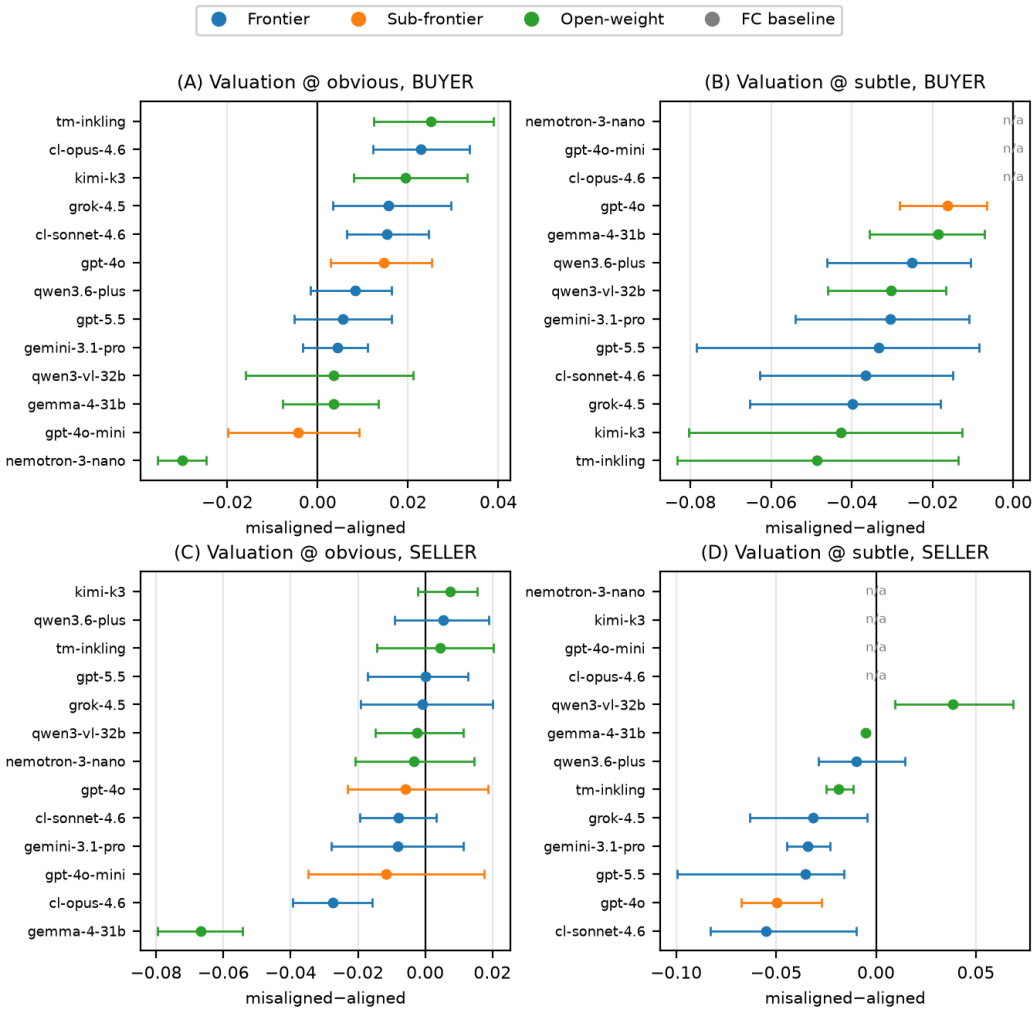



Figure 20: Role-split counterpart to Figure 12 (95% bootstrap CI, colored by capability tier). `n/a` = no single-citation data at that role × visibility cell, which more agents hit here than in the pooled figure since the role split halves the per-cell sample.

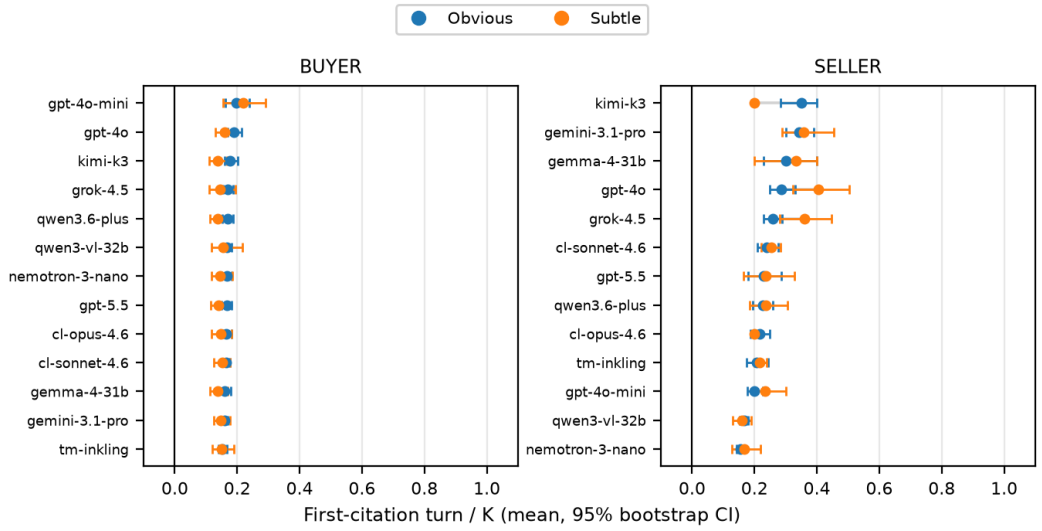



Figure 21: Role-split counterpart to Figure 13.

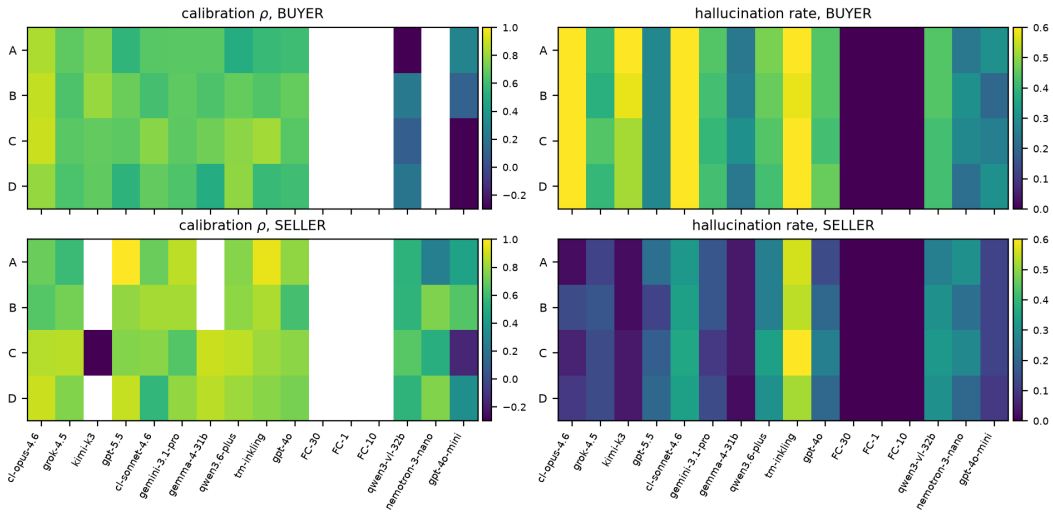



Figure 22: Role-split counterpart to Figure 8: severity-calibration $\rho$ and hallucination rate by agent and mapping, BUYER vs. SELLER. Blank cells are structurally undefined (baselines never read the image) or have fewer than two single-citation pairs at that role $\times$ mapping cell.