



User Behavior Study

In AI Shopping, the Product Card Is Your Storefront

Getting recommended is where the competition starts. This study follows what shoppers did next, from the product card to the offer card.

40 U.S. participants · 224 recorded tasks in ChatGPT and Google AI Mode

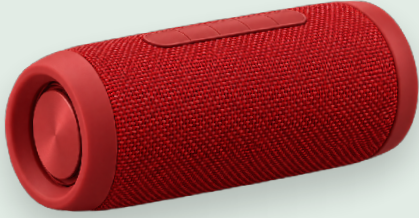
Eric Van Buskirk
Research and analysis
Clickstream Solutions

Scot Wingo
Commerce strategy
ReFiBuy

Brian Chapman
Editorial
ReFiBuy



Compact Speaker
★★★★☆ 4.2
\$79.99



Waterproof Portable Speaker
★★★★★ 4.8
\$129.99



Mini Clip Speaker
★★★★★ 4.6
\$49.99

In stock
Free 2-day delivery

\$129.99

Visit

Contents

01	AI shopping has two shelves. The decision starts before the click.	6
02	How we studied this	8
03	The product card is the shelf, and clicks miss part of the shopping	10
04	The rank advantage is concentrated at position 1	14
05	The offer cards are where competition tightens	16
06	What shoppers weighed in the product decision: specs, price and brand	17
07	Most shoppers don't use keyword prompts in ChatGPT or AI Mode	21
08	Sponsored results: relevance matters	24
09	Trust does not end inside the AI interface	27
10	Questionnaire	30

RECOMMENDATIONS

8 actionable recommendations, in brief	5
What to do now	31
Agentic commerce is reshaping where products compete	33

METHODS AND NOTES

What held up, and what to treat as directional	34
Appendix: study methods and context	35
Notes on the findings	37
About this study	39

Key findings

Getting recommended in AI shopping is just the start. From there, shoppers chose twice: which product to pick from the product cards, then where to buy it from the offer cards. Brands and retailers have to win both.

Forty U.S. participants shared their screens and narrated as they shopped. We coded hovers, clicks, prompts, product and seller choices, and the reasons behind them: **32 research fields across 224 tasks, more than 5,500 coded observations.**

The recordings captured evaluation that surveys and referral traffic alone cannot show. Here's what we found across the Research, Find, Buy journey:

Buying decisions happen in the product cards.

Product cards carried **85.3%** of final product choices, and **83.7%** even where no one was told to use them. Offer cards presented a separate decision: which seller to choose.

83.7%

of final product choices came from a product card

Categories with no product card instruction · 87 of 104 ChatGPT tasks

Shoppers went to the product cards first.

The written answer explained tradeoffs and set criteria, but product cards drew first meaningful engagement in **75.3%** of tasks, about 3.5× the written answer's **21.5%**.²

75.3%

of tasks started at the product cards

165 of 219 tasks

Position pays, and it pays twice.

When there was more than one product card, **43.4%** chose the first, about 1.5× the **29.2%** equal-chance benchmark. At the offer cards, **75.9%** took the first offer in lists with more than one.^{8,9}

43.4%

of selections went to the first product card

56 of 129 tasks

Product card clicks understate the impact of product cards.

In **36.2%** of tasks, shoppers hovered over product cards, clicked none, and still decided. Referral traffic alone cannot show that evaluation.³

36.2%

of tasks where shoppers hovered, clicked no product card and still decided

79 of 218 tasks with click data

Shoppers trusted specs, price and brand far more than reviews.

Specs, price and brand were the primary trust reason in **82.1%** of tasks; reviews in **2.2%**. Yet afterward, **75%** of participants named reviews as important, in a questionnaire that allowed multiple answers.⁵

82.1%

of main trust reasons were specs, price or brand

184 of 224 tasks

[ReFiBuy](#) commissioned this study. [Clickstream Solutions](#) designed the research, coded the sessions, and ran the analysis. ReFiBuy contributed the Agentic Commerce Optimization (ACO) lens: what the findings mean for brands and retailers, and what to do about them. The study measured selections, not completed purchases.

8 actionable recommendations for brands and retailers

ReFiBuy built these from the findings. The study did not test whether they improve placement or sales. Full recommendations on page 31.

1 Win eligibility before you chase clicks.

85.3% of final product choices came from a product card (191 of 224 tasks).

2 Treat position 1 as a priority.

43.4% chose the first product card, against 29.2% by equal chance (56 of 129 tasks).

3 Optimize seller offers separately.

75.9% took the first offer in lists with more than one (44 of 58 tasks).

4 Make specs, price and brand easy to compare.

82.1% of primary trust reasons were specs, price or brand (184 of 224 tasks).

5 Treat imagery as product data, not decoration.

One shopper chose a product without a useful image, then disliked it when she finally saw it.

6 Audit price and variant accuracy across product cards, offer cards, and your page.

Shoppers caught prices on product cards that did not match the seller's page.

7 Make sponsored placements relevant before you make them prominent.

Of 20 recorded reactions to sponsorship, 11 were negative and five positive; in four, the shopper could not tell.

8 Measure the stages before the click.

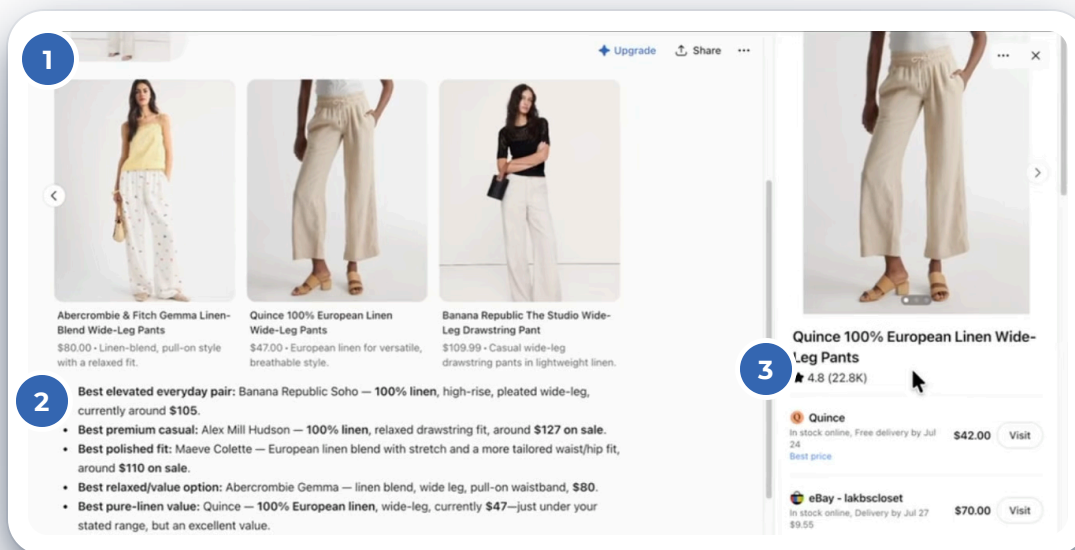
In 36.2% of tasks, shoppers hovered over product cards and clicked none (79 of 218 tasks with click data).

AI shopping has two shelves. The decision starts before the click.

Most conversations about agentic commerce stop at whether your product gets recommended. That is the first competitive step, not the whole race.

Two decisions inside AI shopping

Watch someone shop inside ChatGPT or Google AI Mode, and you see two decisions: which product, on the product cards, and which seller, on the offer cards. The written answer frames the choice.



1 Product cards

A product card presents a product: an image, a title, a price, and the details the AI chose to show. It appears inside or alongside the written answer, where shoppers decide which product fits.

2 Written answer

Frames the options. The AI's narrative response explains tradeoffs and can recommend products by name.

3 Offer cards

An offer card presents sellers of the chosen product: merchant name, price, availability, and the path to the merchant's page. It is where shoppers decide where to buy.

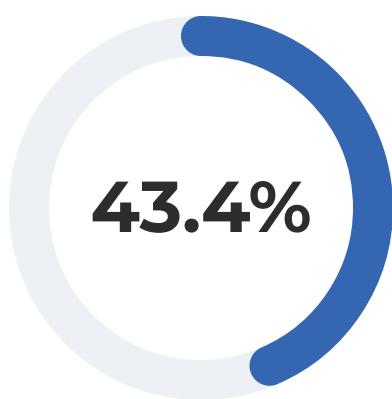
Note. One recorded study session in ChatGPT, combined from two frames of the same recording; numbers added. Source: Clickstream Solutions study recording.

First position mattered at both shelves.

Position 1 drew **43.4%** of product card choices in groups with more than one option, about 1.5× the **29.2%** equal-chance benchmark. In a separate analysis of multiple-offer lists, **75.9%** chose the first seller offer. Sections 4 and 5 examine these different choice sets.^{8,9}

FIGURE 1

Share of selections that went to the first position

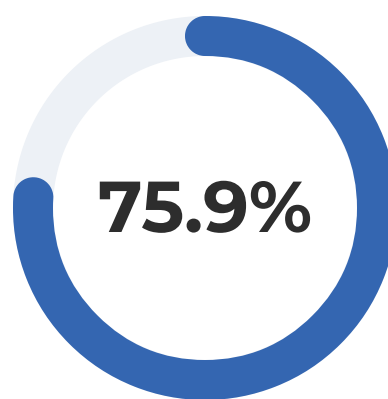


Product cards

went to the first product card

About 1.5× the 29.2% equal-chance benchmark

56 of 129 tasks with two or more product cards



Offer cards

went to the first offer

About 2× the 37.8% equal-chance benchmark

44 of 58 tasks with two or more offers

Note. Groups with one product card, lists with one offer, and groups or lists of unknown size are excluded. Different subsets; position was not randomized. Source: Clickstream Solutions task annotations; see notes 8 and 9.

So a product competes three times. It has to earn the AI shopping agent's recommendation. Its product card has to hold up against the other product cards shown. And the seller's offer card has to compete with other offers for that product.

A retailer can stock the chosen product and still lose the shopper to another seller. A brand can win the product choice while a retail partner gets the sale. The product card deserves the same care as your product page because shoppers are already evaluating the item there.

How we studied this

02

Forty U.S. participants took part in remote, unmoderated desktop shopping sessions using ChatGPT and Google AI Mode. After consent, [UXtweak](#), a usability research platform, captured their screens and voices as they narrated their actions. The analysis covers **224 annotated tasks** across six categories; we excluded incomplete tasks and recordings with technical problems.

In each task, participants chose a subcategory they would genuinely consider buying within six months. Price bands defined the budget, while the tasks stayed open enough for participants to bring their own needs and preferences. The category and platform breakdown appears below.

Instructions varied by category and shaped the behavior we observed. Shoes, pants, and personal care required a final choice from a product card if one appeared. Pants and personal care also required a seller-site visit, without checking out. The other three categories had no card-choice instruction and were all completed in ChatGPT.

FIGURE 2

Task instructions and platform by category

224 annotated tasks. Instructions differed by category on purpose; platform, category, and instruction varied together.

CATEGORY	PLATFORM MIX		TASKS	CHOOSE FROM A CARD IF ONE APPEARS	VISIT THE SELLER'S SITE
	ChatGPT	Google AI Mode			
Watches			24 / 0	☐	☐
Consumer electronics			40 / 0	☐	☐
Bags			40 / 0	☐	☐
Shoes			13 / 27	☒	☐
Pants			40 / 0	☒	☒
Personal care			1 / 39	☒	☒
All tasks	ChatGPT 158 · Google AI Mode 66		158 / 66	120 tasks	80 tasks

Note. Task instructions and platform by category across the 224 annotated tasks. Each bar shows the category's split between ChatGPT and Google AI Mode. An empty circle means the action was not required, not that it was prohibited. Source: Clickstream Solutions study protocol and task annotations.

For the card-based tasks, participants received a plain-language definition and visual examples. The instructions called these “shopping cards”; this report uses “product cards.”

FIGURE 3

The definition and examples shown before the card-based tasks

WHAT PARTICIPANTS WERE TOLD

A shopping card shows one specific product with at least a product image and a product name. It may also include a price, a store, a rating, reviews, a description, or other information.

Shopping cards may appear alone (as one) or in a row, list, grid, or scrollable section in a group with other shopping cards.

PRODUCT GROUP



Product name
Price
Store
★ Rating



Product name
Price
Store
★ Rating



Product name
Price
Store
★ Rating



SINGLE PRODUCT CARD



Product name
Price
Store
★ Rating (count)

Description or other details

Participants were asked to hover over whatever held their attention, giving us an engagement signal that clicks alone would not. A recorded hover lasted at least two seconds; it is not eye tracking. Trained annotators coded every recording across the 32 research fields summarized above.

Product card rank refers to the selected product card’s position within its recorded group; calculation details are in the notes.

The questionnaire findings appear in section 10; full methods and limitations are in the appendix.

The product card is the shelf, and clicks miss part of the shopping

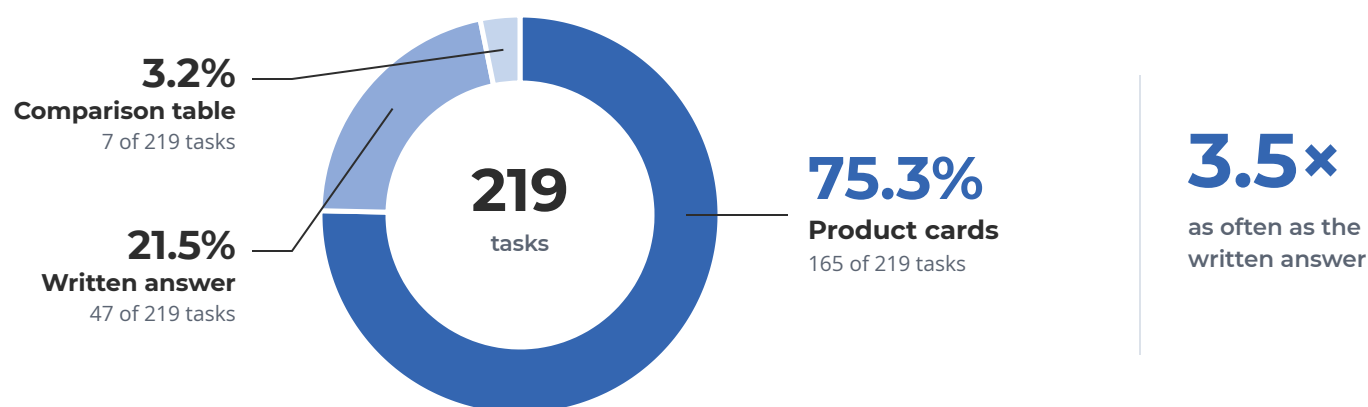
Product cards appeared in **220 of 224 tasks (98.2%)** and carried the final product choice in **191 (85.3%)**.¹ In watches, electronics, and bags, where no one was told to choose from a product card, they still carried **87 of 104 final choices (83.7%)**. Those categories were all completed in ChatGPT, and the categories differ, so this is not a controlled test of the instruction. It does show that card-based selection extended well beyond the tasks that asked for it.

Product cards also drew attention first. Among the 219 tasks with a recorded first meaningful engagement, product cards came first in **165**, the written answer in 47, and a comparison table in seven.²

FIGURE 4

Product cards drew first meaningful engagement in 3 of 4 tasks

First element in the ordered engagement sequence. Base: 219 tasks with meaningful engagement (5 excluded).



Note. Not eye tracking; first recorded engagement. Product cards drew first meaningful engagement about 3.5× as often as the written answer. Base: 219 tasks with a recorded first meaningful engagement; five tasks excluded. Source: Clickstream Solutions task annotations, ordered engagement sequence.

The written answer still mattered. It framed criteria, explained tradeoffs, and helped narrow the options. But the choice rarely landed there.

Card engagement extended beyond clicks. Where product cards appeared, at least one was hovered in **210 of 220 tasks (95.5%)**, and the card set was coded as considered, fully or partially, in **198 of 219**. Only **133 of 218** tasks with usable annotation data included a product card click (61.0%).³

FIGURE 5

Product card clicks understate the impact of product cards

Each measure uses its own eligible base, shown under each value.



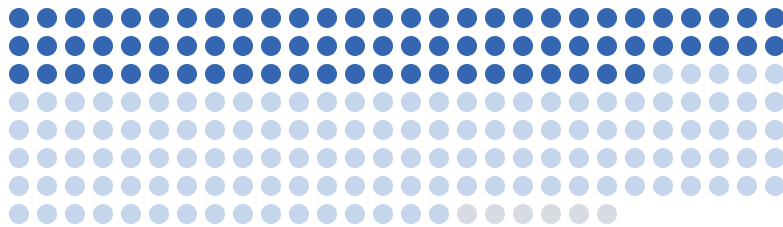
Note. Hover was instructed; two-second threshold. These are distinct measures, not successive steps in a funnel. Source: Clickstream Solutions task annotations.

The gap between hovers and clicks is the measurement finding. In **79 of 218 tasks (36.2%)**, participants hovered over product cards and clicked none of them, then reached a decision anyway. The recordings captured evaluation that card-click counts missed, two seconds at a time. Hovers exceeded clicks in every category and on both platforms.¹³

FIGURE 6

In 36% of card-present tasks, shoppers hovered over product cards and clicked none

Each dot is one task · 218 tasks



36.2%

Hovered, clicked none

79 of 218 tasks

61.0% clicked a product card

133 of 218 tasks

2.8% neither

6 of 218 tasks

Note. Base: 218 tasks with product cards present and usable click counts. Card clicks only; an external visit may still have occurred. A product card click captured only part of recorded card engagement. Hovers were instructed and lasted at least two seconds. Source: Clickstream Solutions task annotations.

This does not mean participants clicked nothing elsewhere or never visited a seller. For a brand watching referral traffic, the distinction matters: traffic reports show arrivals, not the hovers, comparisons, and choices that happened before them or without them.

Where product card design visibly broke: imagery

The qualitative review recorded **11 negative comments and three positive comments** about imagery across 10 sessions. These are comments from selected sessions, not a measure of how common image problems were across the study.

One shopper worked through a task without useful product images and narrated the consequence:



I'm honestly really annoyed that there's no pictures.

STUDY PARTICIPANT

Then, after clicking through:

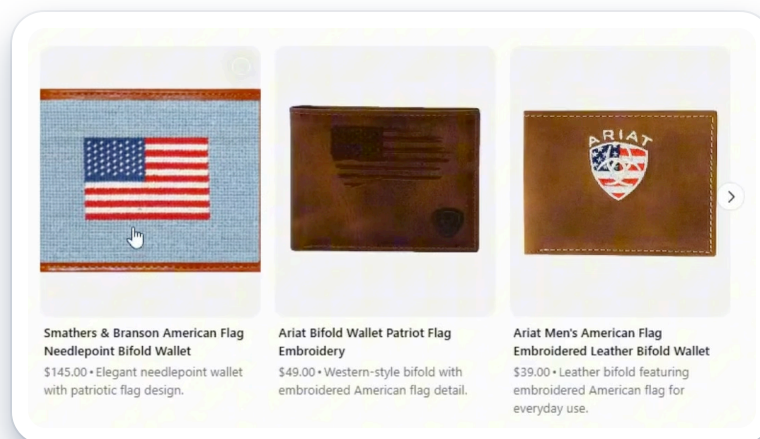


Ew, I actually hate those. Would have known that if there was a picture of them.

STUDY PARTICIPANT

She finished the task having selected a product she disliked the moment she saw it. That is one reason a recorded choice should not be read as a purchase, and one reason imagery belongs in the evaluation itself.

The positive cases point in a specific direction. Images earned praise when they answered a question: multiple angles, a look inside a bag, available color variants, a product shown on a person. One shopper chose a wallet specifically because the product card let him see inside it. Another rejected a heavily discounted item because the image quality was poor.



Recorded session. Wallet product cards in ChatGPT: image, product name, price and a one-line description. An example of product card imagery, not the task described above. Source: Clickstream Solutions study recording.

What this means for brands and retailers.

Provide images that answer product questions, then inspect what the AI interface actually displays. Missing or misleading imagery can originate in source assets, platform processing, or presentation. The study records the shopper's experience; it does not establish the technical cause of each problem.

The rank advantage is concentrated at position 1

04

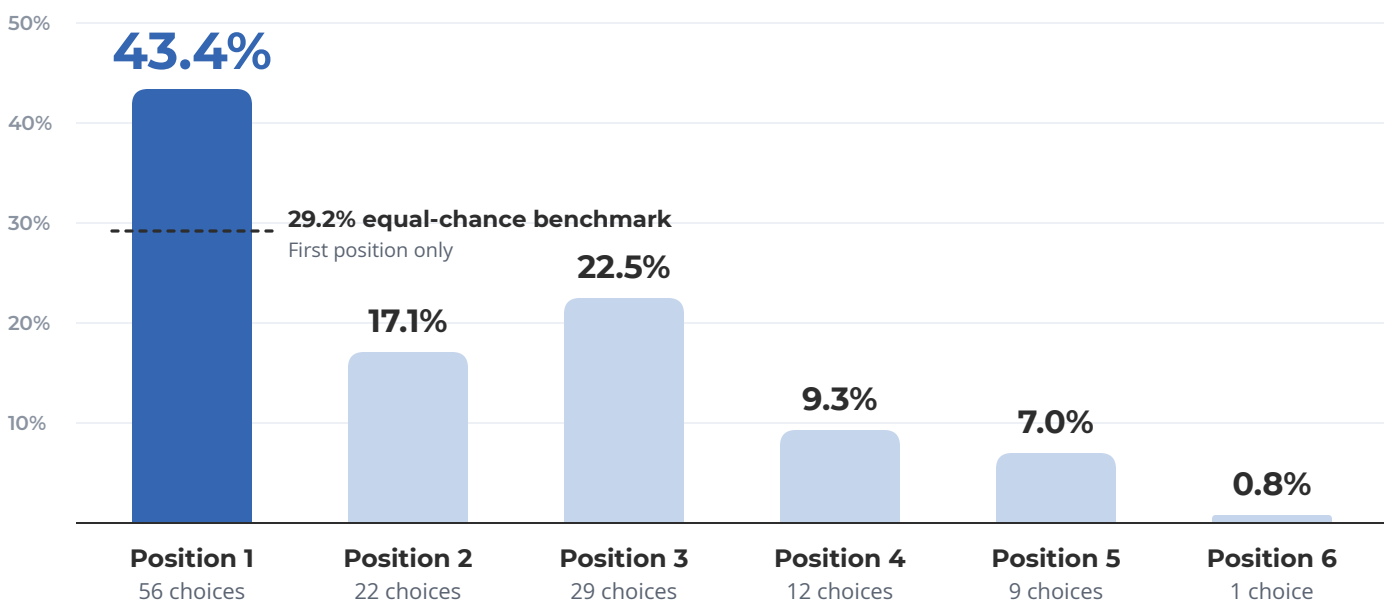
Position 1 was the most frequently chosen position by a wide margin.

Among the **129 tasks** where the final choice came from a product card group with at least two visible product cards, the first product card was selected in **56 (43.4%)**.⁸ If choices had been random within each recorded group, first position would have been chosen about **29.2%** of the time. The observed rate is roughly 1.5× that benchmark.

FIGURE 7

First position led at about 1.5× chance, and did not lock the pick

Share of final product card choices by recorded position. Base: 129 tasks with two or more product cards.



Note. Excludes 51 single-card groups and 7 with unknown totals; 5 lower-bound totals (3+, 5+) kept at their visible count. Benchmark: mean of 1/group size. The 29.2% benchmark applies to first position only; later positions were not available in every group. Source: Clickstream Solutions task annotations.

Exploring more did not visibly erode the first product card's advantage. Among tasks where the participant hovered over zero or one product card, **20 of 44** chose the first (45.5%). Among tasks with two or more product cards hovered, **36 of 85** did (42.4%). This is a directional finding given the sample sizes, and hover counts describe the whole task rather than the moment of choice.

Across the 129-task rank analysis, **73 choices (56.6%)** went to another product card in the group, with the second position chosen less often than the third. Shoppers stayed inside the set the AI built, but they made their own selection within it.

Choosing past the first product card

The session accounts show what moving past the first product card looked like:



Although [the AI] has ranked this one towards the bottom, I think that it would be best for my purposes having the laptop sleeve included in the back.

STUDY PARTICIPANT

Another shopper abandoned a brand he already knew after reading a specification on a cheaper, unfamiliar competitor's product card, then chose the unfamiliar one. These accounts illustrate the reasoning behind individual choices; they are not additional estimates from the 129-task analysis.

Position was not randomly assigned. These results describe where choices concentrated; they do not isolate position's effect from product relevance, price, or the AI's judgment about fit.

What this means for brands and retailers.

There are two jobs. Earn inclusion and a strong position. Then make sure the comparable evidence on the product card gives shoppers reasons to choose your product.

The offer cards are where competition tightens

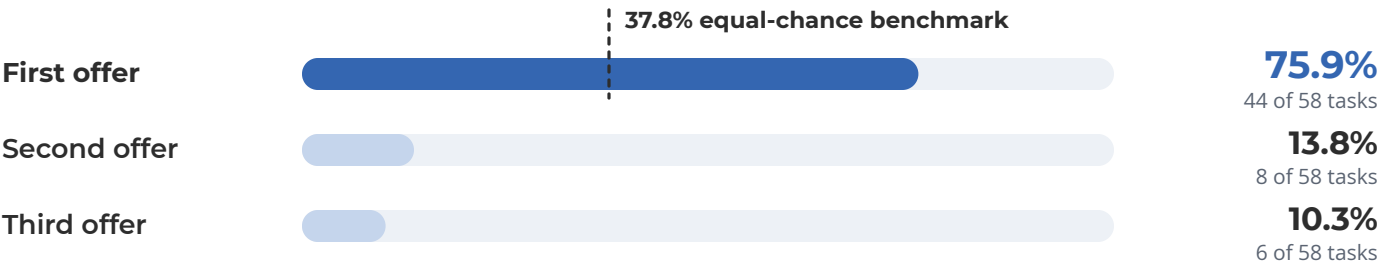
Once a shopper has chosen a product, the competing sellers present a separate decision. In the recorded multi-offer lists, choices were concentrated at position 1.

Among **58 tasks** with an exact recorded list of two or more seller offers, **44** selected the first offer (**75.9%**). That is about 2× the **37.8%** equal-chance benchmark.⁹ Including the 14 lists that showed only one offer, the figure is **58 of 72 (80.6%)**. Rates were similar on both platforms.

FIGURE 8

Three in four seller selections went to the first offer, about 2× chance

Share of seller offer choices by recorded position. Base: 58 tasks with two or more offers.



Note. Excludes single-offer and unknown-size lists. Benchmark: mean of 1/list size. Source: Clickstream Solutions task annotations; note 9.

Offer lists were shorter than product card groups, so the two results are not directly comparable measures of position’s effect. Seller-site visits were required in two categories, which supplied 39 of these 58 tasks. These counts do not establish that position caused the choice or that shoppers stopped looking afterward.

What this means for brands and retailers.

Offer card optimization is a different job from product card optimization, with different inputs: merchant identity, current price, availability, fulfillment, delivery timing, returns, and the destination for the selected variant. A product can win the shopper’s choice while another seller gets the next step. Stocking the chosen product does not ensure your offer is selected.

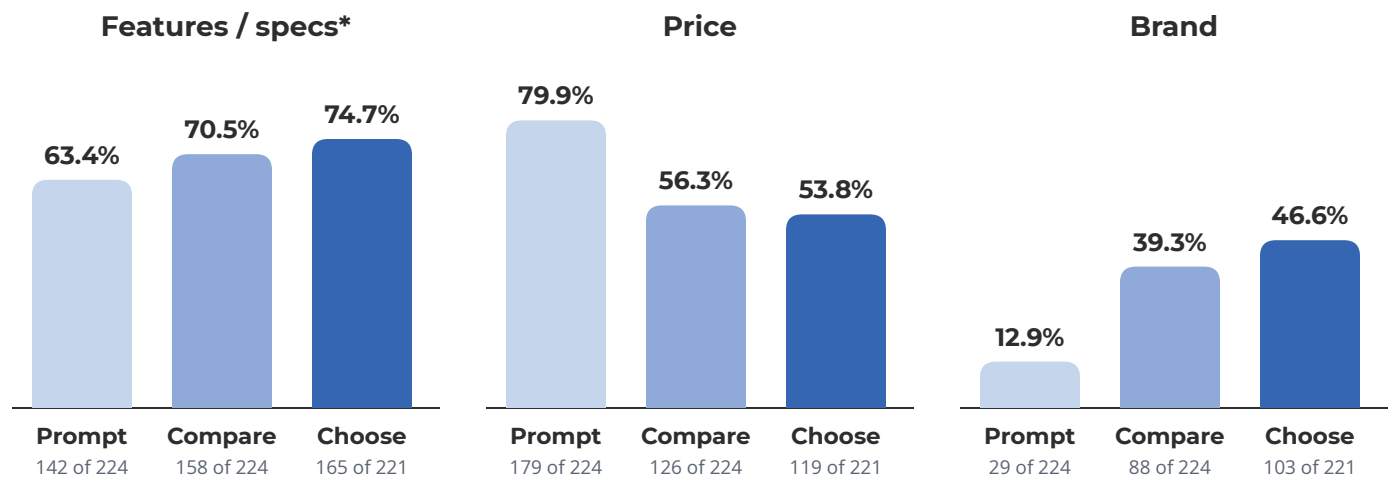
What shoppers weighed in the product decision: specs, price and brand

Across the tasks, the same three attributes emerge: what shoppers asked for, what they compared on, and what they cited when they chose.

FIGURE 9

What shoppers asked for, compared on, and cited when they chose

Share of tasks in which each attribute was coded at each stage. Prompt and comparison: 224 tasks. Final rationale: 221 tasks. A task can carry several attributes.



Note. *Final rationale includes features, specs, style and fit. Comparison field blank in 41 tasks. Each attribute counted once per task. Source: Clickstream Solutions task annotations.

Prompt and comparison figures use all 224 tasks; final-choice rationales use the 221 tasks with an identifiable rationale. Each attribute is counted once per task, and a task can carry several.⁴

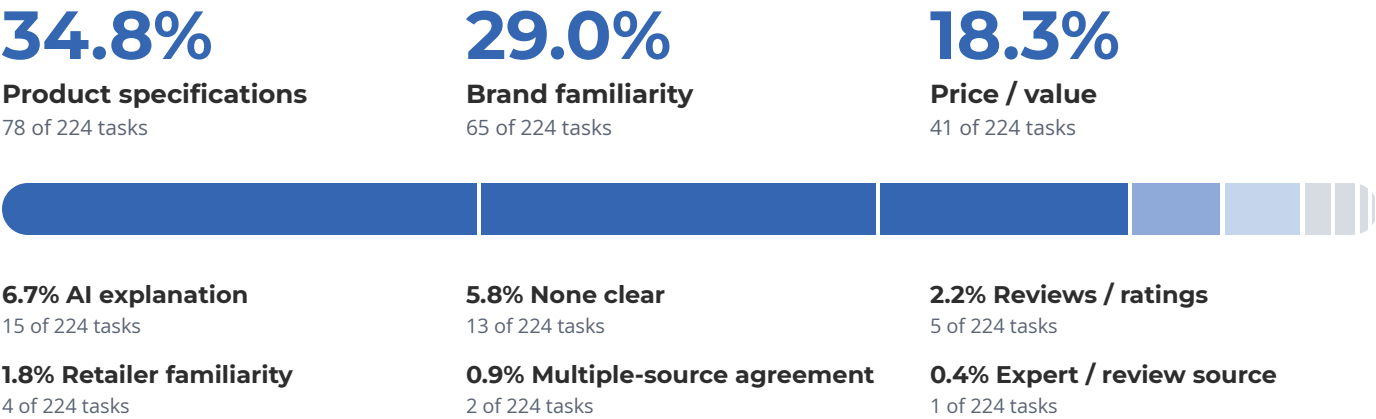
Price led the opening prompts, shaped in part by the budgets we set. Brand was rarely named at the start but climbed through comparison and finished in nearly half of final rationales. Features and specifications rose throughout and finished first.

The trust coding tells the same story from another angle. Product specifications were the primary trust reason in **34.8%** of tasks, brand familiarity in **29.0%**, and price or value in **18.3%**. Together, they made up **82.1%**.⁵

FIGURE 10

Specs, brand familiarity, and price were the primary trust reason in 82% of tasks

One primary trust code per task. Base: 224 tasks.



Note. Task-wide code, not every reason spoken. Codes reflect evidence across the task, including the written answer and external pages. Source: Clickstream Solutions task annotations.

Price and specs can appear on the product card; brand familiarity is the one shoppers bring with them. These are task-wide measures; they do not isolate something read on a product card from something read in the written answer or on a seller’s page.

The recordings supply that detail: shoppers comparing a specification across two product cards, checking a price against a budget, recognizing a brand they already trusted.

Top Max-Cushioning Picks



Asics Novablast 5

\$129.95 ~~\$150~~

ASICS & more

4.8 ★★★★★ (769)

Named the best overall cushioned running shoe by [iRunFar](#). Its thick FlyteFoam Blast+ Eco cushioning provides a deep bed of shock absorption to protect your joints, while the trampoline-inspired outsole helps roll your stride forward effortlessly.

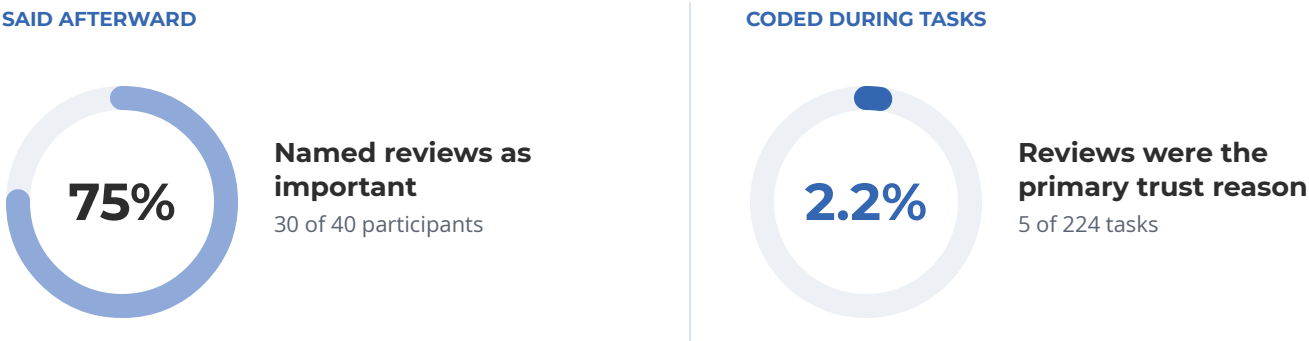
Recorded session. A product card and its recommendation text in Google AI Mode: price and brand on the product card, specifications in the written answer below it. Source: Clickstream Solutions study recording.

Reviews were rarely the primary trust reason in the moment

Reviews and ratings were the primary trust reason in **five of 224 tasks (2.2%)**. Asked afterward which factors were most important in completing the tasks, **30 of 40 participants (75%)** named reviews or testimonials.⁵

FIGURE 11

Stated afterward versus observed during the tasks. Different bases and different questions; read as a contrast, not a paired measure.



Note. The questionnaire was participant-level and allowed several answers; the trust code is one per task. Source: Clickstream Solutions post-study questionnaire and task annotations.

The contrast does not show reviews were irrelevant: participants could name several factors afterward, while the task coding identified one primary trust reason. What people say matters and what they lean on with the options in front of them can diverge. That is a reason to be careful building an AI shopping strategy from survey research alone.

I searched current listings, and the lowest new price I found is \$99.00.

Best deal



PhotoSpring 10-inch WiFi Digital Picture Frame
\$99.00 • Walmart - PhotoSpring, Inc.

Why this is the best buy:

- Price: \$99.00
- Seller: Walmart (sold by PhotoSpring, Inc.)
- Free shipping is available on the listing.

Recorded session. A single product card in ChatGPT with the price, the seller and the AI’s reasons for the pick. Source: Clickstream Solutions study recording.

Price and variant integrity: when the product card says the wrong number

This was the most concrete product-data failure we recorded, and it surfaced at the point of commitment.

Shoppers caught prices on product cards that did not match the seller's page:



It currently says [the retailer has it] in stock online for \$24.97. A lot cheaper than what was shown originally in the top shopping tile.

STUDY PARTICIPANT

They found a price attached to a different package size:



They actually have this option for \$36 ... I believe that's actually the 1.7 ounce option, that's what it defaults to. When we go to our nine ounce option, it is \$48.

STUDY PARTICIPANT

And prices that disagreed within a single interface:



I'm a little confused because this is \$140, but everything I see here looks like \$160.

STUDY PARTICIPANT

Related inconsistencies turned up in attributes. One page stated 94% naturally derived at the top and 95% further down, and the shopper noticed. Another product card rendered an SPF rating as a percentage.

The sessions document the experienced mismatch, not its root cause. It could sit in a feed, a product page, platform processing, or somewhere else in the chain.

What this means for brands and retailers.

Price, size, quantity, and variant attributes need to stay consistent between the AI's product card and the relevant destination page. A price attached to the wrong size changes the comparison the shopper is making. Make attributes explicit, current, and easy to compare, with consistent units. Persuasive copy cannot supply a missing dimension, resolve an ambiguous variant, or explain a conflicting price.

Most shoppers don't use keyword prompts in ChatGPT or AI Mode

Three in four opening requests (**172 of 224, 76.8%**) were full natural-language prompts rather than keyword strings.⁶ Participants front-loaded requirements: dimensions, fit, materials, waterproofing, budget, intended use. The assigned categories and budgets shaped those requests; some participants deliberately left room for exploration.

FIGURE 12

Three in four opening requests were full natural-language prompts

Opening request style, share of 224 tasks.

76.8%

Full natural-language prompts

172 of 224 tasks



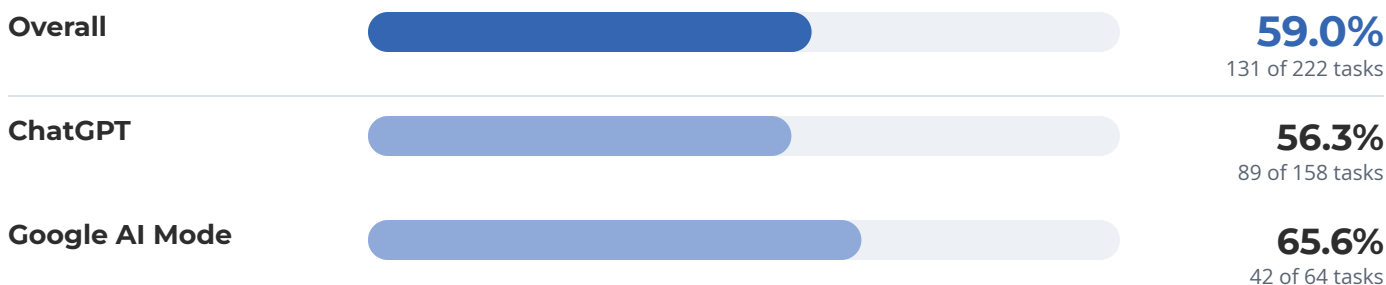
Note. Assigned categories and budgets shaped the requests. Source: Clickstream Solutions task annotations.

Among **222 tasks** with a recorded follow-up count, **131 (59.0%)** involved no follow-up prompt at all. The share was **56.3%** in ChatGPT and **65.6%** in Google AI Mode.⁷ That describes the conversation; shoppers could still compare product cards or verify elsewhere.

FIGURE 13

Most shopping tasks had no follow-up prompt

Share of tasks in which the participant sent no follow-up after the opening request, by platform.



Note. Zero follow-ups does not establish that the shopper skipped other evaluation. Base: 222 tasks with a usable count. Source: Clickstream Solutions task annotations.

Not everyone wanted to be narrowed

Some shoppers kept requests broad on purpose to see more options:



With my prompt, I really just said that I wanted to stay under \$200 ... Other than that, I just wanted [the AI] to throw me all the options it could. I didn't want to get too specific because I really wanted to do a little research and see what options were available to me.

STUDY PARTICIPANT

The same shopper declined the AI's offer to narrow further. Others specified everything up front. Product data has to serve both: identify the item clearly, and describe the conditions under which it fits.

Others delegated the retrieval but audited the sourcing. One built instructions into his prompt to consult specific communities, avoid overconfidence, and list sources. He then checked the source list and noticed when citations disappeared across follow-up turns.

When the AI added a requirement nobody asked for



I want lots of pockets. Yeah, I didn't ask for minimalist. Why did it say minimalist? I swear.

STUDY PARTICIPANT

An unrequested attribute lowered confidence in the recommendation. Specific evidence had to match the need.

Gender defaulting created a similar problem. One participant was repeatedly shown men's jeans and corrected the AI mid-session. Another discovered only after clicking through that he had been looking at a women's hiking shoe when he thought it was a men's shoe. A third pre-empted the problem on every prompt she wrote, adding "for a woman" to each request.

The behaviors that moved together, and the ones that did not

We looked for evidence that certain participants were less likely to choose the top-ranked recommendation. Across the 40-participant analysis, rank-1 selection rates had no statistically significant association with prior category knowledge, prompt style, follow-up frequency, or external-click propensity. These results do not rule out relationships a larger study might detect.¹⁷

Prior knowledge was associated with how shoppers explored. Tasks with higher prior knowledge tended to include more product card hovering and fewer external clicks. The pattern remained when the analysis was restricted to tasks without an instruction to visit a seller. Shoppers could inspect more evidence inside the AI interface while making fewer visits outside it.¹⁷

Knowledgeable shoppers still need rich evidence on the product card. A low click-through rate can coexist with close evaluation. These are associations; the study does not establish why an individual shopper stayed or left.

What this means for brands and retailers.

Make fit, dimensions, materials, compatibility, audience, and intended use explicit in your product data, then check whether the product card represents those attributes accurately for the shopper's request. Do not rely on a follow-up prompt to supply missing context.

Sponsored results: relevance matters

08

We recorded **20 distinct reactions** to advertising or sponsorship across 11 sessions in the transcript sentiment analysis: 11 negative, five positive, and four where the shopper could not tell. These are qualitative reactions, not a measure of how common each view was among shoppers.

Negative reactions clustered on mismatch: wrong product, wrong size, wrong brand, wrong destination. Positive reactions appeared when a sponsored result was precisely what the shopper wanted and shortened the path to it.

One shopper read a Sponsored label aloud and chose the item anyway because it was the right item.

A well-targeted ad could still lose to the other products shown:

“

It does have an ad at the bottom for Hoka running shoes, which is probably what I would buy in real life, but based on the ones that it has given me, these would be more my style.

STUDY PARTICIPANT

The ad matched her stated real-world brand preference and still lost to the surfaced products in this task.

Some shoppers read the product card layer itself as advertising. One had just received the study's plain-language definition of a shopping card:

“

So I assume that these must be types of ads. They don't look like the ads I've seen. The ones I've seen are strictly text-based. But I assume these must be some of the types of ads that appear in under a free account.

STUDY PARTICIPANT

One participant described expecting ads below, and clearly separated from, the model's answer:



Any ads that I've seen on a free account are always below the output of the model and clearly separated from that output, so that you know what is the actual output of the model and what the actual advertisements are.

STUDY PARTICIPANT

The after-task questions asked about advertising, which may have influenced attention to it across tasks. The accounts show why sponsorship clarity matters even when the product is relevant.

What participants told us afterward about ads.

Across **180 after-task responses**, **111 (61.7%)** reported seeing advertising or sponsored content. Among those 111, **71.2%** rated the ads as at least as useful as conventional Google Search ads. Nearly half, **48.6%**, said the ads helped them at least a little; **41.4%** reported no effect and **9.0%** said they hurt. Sponsored content influenced or expanded what they considered buying in **35.1%** of these responses.¹¹

61.7%

reported seeing advertising
or sponsored content

111 of 180 responses

71.2%

rated the ads as at least as useful
as conventional Google Search ads

79 of 111 responses

48.6%

said the ads helped them at least a little

54 of 111 responses

35.1%

influenced or expanded
what they considered buying

39 of 111 responses

A separate sponsorship-clarity question received **75 responses**. About half, **49.3%**, called the distinction very or somewhat clear; **30.7%** said it was not very clear or not clear at all. The remaining **20.0%** said they had not noticed sponsored or advertising content.¹¹

49.3%

called the distinction
very or somewhat clear

37 of 75 responses

30.7%

said it was not very
clear or not clear at all

23 of 75 responses

20.0%

said they had not
noticed sponsored or
advertising content

15 of 75 responses

These are task-level responses collected repeatedly from the same participants, not from separate people. Treat them as a temperature check on how participants perceived advertising during the study, not as a population estimate or an ad-effectiveness test.

**What this means
for brands and
retailers.**

Relevance and disclosure both matter. An ad placement that breaks intent can capture attention and still damage consideration. Check whether the product matches the request and whether the sponsorship is clear.

Trust does not end inside the AI interface

09

An external visit lasting more than two seconds was recorded in **104 of 224 tasks**. In **88 of those (84.6%)**, the final product choice was associated with a product card.¹⁰ This identifies the final-choice interface in tasks that included a visit, not the route of the outbound click.

85.0%

of tasks recorded an external visit in the two categories where we asked participants to visit a seller

68 of 80 tasks

In the two categories where we asked participants to visit a seller, an external visit was recorded in **85.0%** of tasks. In the other four, the share was **25.0%**. Among the 104 tasks with an external visit, **87** recorded exactly one external click lasting more than two seconds.¹²

25.0%

in the other four

36 of 144 tasks

The recordings suggest that some visits served as confirmation or continuation after an in-interface choice. The click count alone does not establish that shoppers stopped browsing or completed a purchase. Retailer and brand or direct-to-consumer sites were the most commonly recorded destinations.

Retailer identity

Retailer names worked as a shortcut to quality for some shoppers:



By Nordstrom, that is another factor that I am considering. I know Nordstrom usually only carries brands that are reputable and good quality.

STUDY PARTICIPANT

Another shopper used unfamiliarity as a disqualifier:



I know Dick's Sporting Goods. I haven't heard of these other online sellers ... they're an established company, I'm going to go with them.

STUDY PARTICIPANT

And the same merchant that reassures one shopper can rule out the sale for another. One shopper found the lowest price available and refused it:



Looks like Walmart's the best one. Okay. We don't want to go to Walmart.

STUDY PARTICIPANT

Retailer preference is personal. If you sell through partners, which partner appears in the offer list matters to the shopper.

What shoppers wanted at the offer stage

Once shoppers reached a seller, a different set of attributes came into play: delivery time, return policy, and whether a store nearby would let them see the product in person.



I would likely choose Target just because I could go in and try them on compared to a website where it ships.

STUDY PARTICIPANT

These accounts illustrate seller considerations without measuring how widespread each was.

Confidence a product card cannot supply

Shampoo shoppers talked about smell; shoe shoppers talked about trying things on. Others sought reassurance beyond the product card. One went looking for balance:



I always like to try to at least look at one negative review just to try to keep it balanced.

STUDY PARTICIPANT

A product card can supply product details, but it cannot resolve every sensory question or replace every source of reassurance.

What this means for brands and retailers.

Your product or merchant page needs to support confirmation and further evaluation. The price, specifications, images, availability, and trust signals a shopper saw inside the AI need to match what they find on arrival. A discrepancy at that point can undermine confidence in the choice.

Questionnaire

10

Participants came into the study as experienced AI users: **85%** said they used AI chatbots at least three times a week, including **65%** who used them daily. After the tasks, **87.5%** rated ChatGPT as helpful or very helpful compared with other chatbots. **Eighty percent** said its answers were always or often accurate, and **85%** rated their confidence in the information it provided at 4 or 5 out of 5.¹⁵ These describe a panel already comfortable with AI. Stated attitudes are kept separate from observed behavior throughout, because the two sometimes diverged.

85%



said they used AI chatbots
at least three times a week

34 of 40 participants

65%



used them daily

26 of 40 participants

87.5%



rated ChatGPT as helpful
or very helpful compared
with other chatbots

35 of 40 participants

80%



said its answers were
always or often accurate

32 of 40 participants

85%



rated their confidence
in the information it
provided at 4 or 5 out of 5

34 of 40 participants

Note. Participant-level questionnaire, base 40 participants. Self-reports, not observed behavior. See note 15.

What to do now

The study gives Agentic Commerce Optimization three concrete jobs: help products get found, give shoppers reasons to choose them, and help sellers win the offer comparison. The eight actions below are ReFiBuy's practical recommendations; the study did not test whether changing product data improves placement or sales.

1

Win eligibility before you chase clicks.

A brand mention in the written answer is not a product card. For your priority shopping requests, record which products and variants appear as product cards, on which engines, and whether the product card shows the right item. Check titles, identifiers, category, price, availability, and variant details together. A product missing from the product card set is absent from that comparison.

2

Treat position 1 as a priority.

It drew about 1.5× its equal-chance share among product cards. Record the request, competing products, and what each product card showed alongside its position. Position was not randomized, and a rank number alone cannot explain the choice.

3

Optimize seller offers separately.

In the multi-offer subset, **75.9%** of seller selections went to position 1. Check merchant identity, current price, availability, fulfillment, returns, and the destination for the selected variant. These are shopper considerations and operational checks; the study did not identify the engines' ranking factors.

4

Make specs, price and brand easy to compare.

Together they accounted for **82.1%** of primary trust reasons across tasks. Provide specific values, consistent units, and clear brand identification. Make fit, materials, audience, intended use, and compatibility explicit, then test requests that reflect real shopping needs.

5

Treat imagery as product data, not decoration.

One shopper chose a product without a useful image, then disliked it when she finally saw it. Supply angles, interiors, variants, and images that show scale or fit, then check what the AI interface actually renders.

6

Audit price and variant accuracy across product cards, offer cards, and your page.

Item, size, quantity, price, availability, and imagery need to agree wherever the shopper meets them. Trace discrepancies across the feed, product page, and displayed product and offer cards without assuming their source. Recheck product and offer cards after correcting the underlying data.

7

Make sponsored placements relevant before you make them prominent.

Shoppers tolerated and sometimes welcomed sponsorship that matched their intent. Check whether the product matches the request and whether the sponsorship is clear. Relevance can earn consideration; it does not guarantee selection.

8

Measure the stages before the click.

In **36.2%** of card-present tasks with usable click data, shoppers hovered over product cards and clicked none. Combine referral analytics with checks of product card presence, position, representation, and seller-offer rank. Use behavioral research to understand the evaluation those measures miss.

The errors shoppers caught were live at the moment they were deciding. Product data and AI interfaces can change, so this work needs repeated checks rather than a one-time cleanup.

Agentic commerce is reshaping where products compete

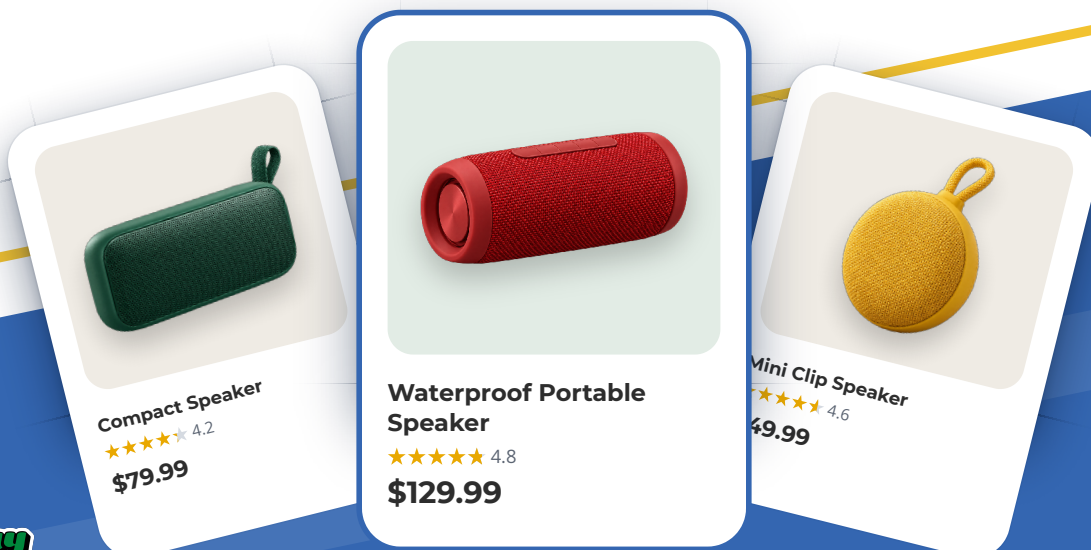
Watched hover by hover, shoppers compared, rejected, caught pricing errors, and refused retailers they disliked. They worked hard inside these interfaces, and much of that work is absent from referral analytics.

What has changed is who builds the room they work in. In agentic commerce, the AI shopping agent assembles the consideration set, orders it, and decides what evidence appears beside each option. Active choice still happens inside that environment. Brands and retailers need to supply accurate product data and check how the agent represents it.

Getting recommended is where the competition starts. Treat the product card as your storefront, and give shoppers the evidence to take the next step.

ReFiBuy finds and fixes product data gaps before shoppers do.

[Book a demo](#)



What held up, and what to treat as directional

The research plan framed five working hypotheses. The table distinguishes the observed findings from questions the study did not resolve.

WE SET OUT TO TEST	WHAT THIS ANALYSIS ESTABLISHES
First visible product cards and first visible offers receive disproportionate attention	First position led at both stages, with different strength. First product card: 43.4% of choices in multi-card groups, about 1.5× the 29.2% equal-chance benchmark. First offer: 75.9% in multi-offer lists, about 2× the 37.8% equal-chance benchmark. These are selected-position results; position was not randomized, so they describe concentration rather than a causal effect.
Product and offer clicks understate engagement	Established descriptively for product cards: in 36.2% of card-present tasks with usable click counts, participants hovered over product cards and clicked none. Hovers exceeded clicks in every category and on both platforms. The exploratory task-level test is described in note 13. An offer-specific estimate was not produced.
Brand, features, and price dominate early comparison; retailer trust, shipping, and returns matter more near the offer	The first half is well supported by the prompt, comparison, and trust coding. Seller considerations appear in individual accounts at the offer stage, but five-minute tasks did not allow a measured early-to-late shift; stage was defined by interface rather than time.
AI-defined shopping criteria increase acceptance of the product set	Not resolved by this analysis. Product card consideration (198 of 219) and the absence of follow-ups (59.0%) are consistent with acceptance but do not measure who supplied a criterion or whether shoppers adopted it. A direct coding of criteria ownership is the next step.
Lower-knowledge shoppers rely more on AI framing and product card cues	Partly supported. Using zero follow-ups as the study's measure of first-answer acceptance, the rate was 81.2% for low-knowledge tasks, 47.6% for medium, and 64.4% for high. This supports the low-knowledge acceptance finding, but not a straight-line relationship between less knowledge and greater reliance. Rates formed a U shape (differences between the three groups were not tested). Acceptance here is an operational measure, not proof that shoppers skipped other evaluation. ¹⁶

Appendix: study methods and context

Participants and tasks.

Forty U.S. participants were recruited through [Prolific](#). All were familiar with ChatGPT and had used a Free or Go plan for at least one month. They were not told before recruitment that the work would involve shopping. Sessions were remote, unmoderated, and on desktop; mobile shopping was not studied. In each task participants picked a subcategory they would genuinely consider buying within six months, within a set price band. The analysis covers 224 annotated tasks: 24 in watches and 40 in each of consumer electronics, bags, shoes, pants, and personal care. Incomplete tasks and recordings with technical problems were excluded.

Design and its limits.

There were 158 ChatGPT tasks and 66 Google AI Mode tasks. Category, platform, task order, and instructions varied together, as figure 2 shows: the three categories without a card-choice instruction were all completed in ChatGPT, and personal care was almost entirely Google AI Mode. Pooled results therefore should not be read as a balanced comparison of platforms or categories. The protocol changed during fieldwork: later cohorts completed five categories after watches were removed, for reasons that included lengthy-recording technical failures, fatigue, and weak product card surfacing in that category. The 24 completed watch tasks remain in the analysis.¹⁴

Collection and coding.

Screen and voice were captured through UXtweak after consent. Participants narrated as they shopped. Trained annotators coded each session across 32 fields covering prompts, engagement, product card presence and format, product and offer selection, rank and group size, trust reasoning, rejection reasons, external visits, and final-choice rationale. Hover was defined as a mouse pause of at least two seconds over a product card. Hovering over elements of interest was instructed throughout.

Who participated.

Participants ranged from 18 to 57 years old, with a median age of 32. Twenty-one were female, 17 male, one preferred not to say, and one had no recorded age. Roughly four in five held at least a bachelor's degree. Sixty-five percent said they used AI chatbots daily and 90% at least weekly, although frequent AI use did not mean they were accustomed to shopping this way; two of the most sophisticated participants said they do not normally shop with AI.

Caveats.

Task instructions shaped card choices, hovering, and external visits. Multiple tasks came from each participant and should not be treated as independent observations for inference. Category, platform, cohort, and instruction differences limit comparisons. Rank was not randomly assigned, so position cannot be separated from relevance, price, or other aspects of the options shown. The study measured selections, not transactions, and did not test the commercial effect of changing product data. Qualitative accounts illustrate behavior; they do not estimate its prevalence.

Notes on the findings

Task counts are drawn from the Clickstream Solutions task annotations, restricted to the 224 tasks with a recorded platform. Questionnaire findings and statistical summaries use the sources and bases specified below. Percentages describe tasks or the stated subset unless identified as participant-level.

- 1** Product cards present: 220 of 224. Final-choice interface coded Shopping cards: 191 of 224. Categories without the card-choice instruction (watches, electronics, bags): 87 of 104, all ChatGPT tasks. Task-level interface classifications, not purchases or evidence that the product card alone caused the choice.
- 2** First element of the ordered engagement sequence: product cards 165, written answer 47, comparison table 7, no meaningful engagement 5. Base 219. As a share of all 224 tasks, product cards first in 73.7%.
- 3** Product cards hovered (two-second threshold) and product cards clicked. At least one hover in 210 of 220 card-present tasks. Two card-present tasks lack a click count, leaving 218. At least one click in 133; hover with zero card clicks in 79; neither in 6. Card clicks only; an external visit may still have occurred. Consideration code (yes 156 plus partially 42) over 219 tasks with a usable code.
- 4** Prompt attributes (base 224): features/specs 142, price 179, brand 29. Comparison factors (base 224): features/specs 158, price 126, brand 88; 41 blank and 3 NONE records are counted in the base as no mention. Final-choice rationale (base 221 with an identifiable rationale): features/specs/style/fit 165, price 119, brand 103. Each category counted once per task.
- 5** Primary trust reason, one code per task: product specs 78, brand familiarity 65, price/value 41 (combined 184 of 224, 82.1%); AI explanation 15; none clear 13; reviews/ratings 5 (2.2%); retailer familiarity 4; multiple-source agreement 2; expert or review source 1. Questionnaire: reviews/testimonials selected by 30 of 40 participants among the most important factors in completing the tasks; multiple selections allowed. Participant-level, not task-level.
- 6** Prompt style: PROMPT 172, KEYWORD 39, MIXED 13, base 224. A task-level style code; assigned categories and budgets shaped the requests.
- 7** Number of follow-ups: usable in 222 tasks; 131 recorded zero (ChatGPT 89 of 158; Google AI Mode 42 of 64). Maximum recorded follow-ups: 9 in ChatGPT, 4 in Google AI Mode. H5 uses zero follow-ups as an operational measure of first-answer acceptance (note 16); it does not establish that shoppers skipped other verification or made a purchase.
- 8** Final-choice interface, product card rank, and group size. Of 191 Shopping cards final choices, 187 have a numeric rank. Excluded: 51 whose recorded group contained one product, and 7 with unknown group totals. Five lower-bound totals (3+, 5+) kept at their visible count. Base 129; first position 56. Group minima: 18 groups of 2, 42 of 3, 30 of 4, 21 of 5, 16 of 6, one of 7, one of 9. Equal-chance benchmark: the mean of 1 divided by recorded group size across the 129 groups, 29.16%, rounded to 29.2%; using minima for the five lower-bound groups makes it an upper bound. Observed 43.41% divided by the benchmark is 1.49. A descriptive comparison with a uniform-choice model, not a causal or significance test. Hover split: zero or one product card hovered, 20 of 44; two or more, 36 of 85. Rank refers to the chosen product card's recorded group, not necessarily its position across the whole answer or in the initial response.
- 9** Offer rank and list size. Exact lists of two or more offers: 58 tasks; first offer 44, second 8, third 6. Equal-chance benchmark: the mean of 1 divided by list size across the same 58 lists, 37.8%. Including 14 single-offer lists: 58 of 72 (80.6%). Two further first-offer records had unknown list totals and are excluded from both. Platform split on the 72 base: ChatGPT 35 of 43 (81.4%), Google AI Mode 23 of 29 (79.3%). Of the 58 multi-offer records, 39 came from pants or personal care, which instructed a seller-site visit. Neither rate establishes that participants stopped looking after the first offer.
- 10** External clicks over two seconds: 104 tasks with at least one. Of those, the final-choice interface was coded Shopping cards in 88 (84.6%). Pants and personal care instructed the visit. Final-choice interface and outbound-click route are different measures.

- 11 After-task questionnaire, task-level responses. Ads noticed: 111 of 180 (61.7%). Among those 111: at least as useful as Google Search ads, 79 (71.2%); helped a little or a lot, 54 (48.6%); no effect, 46 (41.4%); hurt a little or a lot, 10 (9.0%). One response instead said no ads were noticed, accounting for the remaining 0.9%. Shortlist influence or additional consideration: 39 (35.1%). Sponsorship clarity, separate 75-response pool: 37 very or somewhat clear (49.3%), 23 not very or not at all clear (30.7%), 15 did not notice sponsored content (20.0%). Repeated responses from the same 40 participants. The transcript sentiment tally of 20 reactions across 11 sessions is separate from these questionnaire pools.
- 12 External clicks over two seconds. Instructed categories (pants, personal care): 68 of 80 tasks with an external visit (85.0%). Other four categories: 36 of 144 (25.0%). Exactly one external click among visiting tasks: 87 of 104 (83.7%). The summary of external sites visited uses a separate base of 105 non-empty records containing 121 website mentions: retailer 56 (53.3% of records), brand or direct-to-consumer 47 (44.8%), marketplace 8 (7.6%), other 10 (9.5%). These are website-mention counts divided by records, not shares of 105 destinations; multiple mentions per record make the percentages sum to more than 100%. This base differs from the 104 tasks meeting the over-two-second click definition.
- 13 The Clickstream Solutions H4 analysis reports a paired Wilcoxon signed-rank test across 218 task observations: hovers exceeded clicks in 133 tasks, were equal in 76, and clicks exceeded hovers in 9 ($V = 9646$, $p < .001$). Reported category results: $p < .001$ in watches, electronics, bags, pants, and personal care; $p = .002$ in shoes. Both platform results: $p < .001$. Each participant completed several tasks, so these tests count tasks rather than people. Read the p-values as a strong first indication rather than a final measure. The same pattern appeared in every category and on both platforms, which makes it unlikely to come from a few participants. The descriptive hover/click counts do not depend on that inferential claim.
- 14 Task annotations for category and platform counts; Clickstream Solutions study protocol, sections 2 to 4 and the August 10 amendment, for desktop scope, instructions, and fieldwork changes. The amendment records the reduction from six task categories to five and the removal of watches from later cohorts. Results are observations under the protocol's task conditions, not representative market rates.
- 15 Participant-level questionnaire, base 40: AI chatbot use at least three times weekly 34, daily 26; ChatGPT helpful or very helpful compared with other chatbots 35; always or often accurate 32; confidence 4 or 5 of 5 on information about features and options 34. Self-reports, not observed behavior. Source: Clickstream Solutions questionnaire analysis.
- 16 The Clickstream Solutions H5 analysis defines first-answer acceptance as zero follow-ups. Low knowledge: 26 of 32 tasks (81.2% as reported); medium: 49 of 103 (47.6%); high: 56 of 87 (64.4%); base 222. Within-group binomial tests against 50% were reported significant for low and high, not medium; these are not pairwise tests between groups. Kendall's tau-b was -0.006 ($p = .9306$), a monotonic test that does not assess the U-shaped pattern. Low-knowledge acceptance is supported under this definition; a monotonic less-knowledge/more-reliance gradient is not. Repeated task observations make inferential results exploratory.
- 17 Clickstream Solutions correlation analysis: prior knowledge versus product card hovers, Spearman $r = +0.259$ ($n = 220$, reported adjusted $p < .001$); versus external clicks over two seconds, $r = -0.215$ ($n = 224$, $p = .009$). Without a seller-visit instruction: hovers $r = +0.267$ ($n = 143$, $p = .009$); external clicks $r = -0.240$ ($n = 144$, $p = .022$). Platform-specific significance was not uniform, so the text describes the pooled and uninstructed-task associations. These tests count tasks rather than people, since each participant completed several. The associations held both across all tasks and in tasks without a seller-visit instruction. Participant-level rank-1 selection correlations ($n = 40$): knowledge -0.010 , prompt style $+0.193$, follow-ups -0.117 , external-click propensity $+0.050$; none statistically significant. These use the original rank analysis, not the 129-task product card subset.

About this study



COMMISSIONED AND PUBLISHED BY

ReFiBuy is the Agentic Commerce Optimization (ACO) platform for brands and retailers. It continuously evaluates, enriches, and monitors product data, the information behind the product cards and offer cards this study examined, so AI shopping agents can accurately interpret, include, and recommend products. ReFiBuy commissioned this study and contributed the ACO lens: what the findings mean for brands and retailers, and what to do about them.



CLICKSTREAM

RESEARCH PARTNER

Clickstream Solutions runs studies on how people use Google and AI search. Its private studies help clients discover how AI platforms shape awareness of their products compared with direct competitors. It also partners on published research, pairing its findings with clients' industry expertise. The firm has done SEO for eleven years. For this study, Clickstream Solutions designed the research, coded the sessions, and ran the analysis.

AUTHORS



Eric Van Buskirk

Clickstream Solutions

Research and analysis

Founder of Clickstream Solutions. Eric has directed research studies for some of the best-known names in SEO and AI search, most recently behavioral studies with Kevin Indig and Profound on how people use AI platforms.



Scot Wingo

ReFiBuy

Commerce strategy

Co-founder and CEO of ReFiBuy. Scot is a serial entrepreneur with deep ecommerce experience, founding five startups and taking one public: ChannelAdvisor (now Rithum).



Brian Chapman

ReFiBuy

Editorial

VP of Growth at ReFiBuy. Brian has over a decade of experience building go-to-market engines at B2B SaaS and high-growth startups, including Spiffy and Dude Solutions (now Brightly).

PUBLISHED

October 2026

FIELDWORK

July–August 2026

SAMPLE

40 U.S. participants,
224 tasks

PLATFORMS

ChatGPT and Google
AI Mode

The study measured selections, not completed purchases. Product imagery in illustrations is fictional concept art; real interface captures are labeled “Recorded session.”





In AI Shopping, the Product Card Is Your Storefront

A User Behavior Study by ReFiBuy, with research by Clickstream Solutions. 2026.

CITE THIS REPORT

Van Buskirk, E., Wingo, S., and Chapman, B. (2026). *In AI Shopping, the Product Card Is Your Storefront*. ReFiBuy and Clickstream Solutions.

GET IN TOUCH

refibuy.ai