3 AI Models Shopped Shopify’s Catalog — Each Picked A Different Winner via @sejournal, @slobodanmanic

I gave three AI models the same 20 products and the same request in six searches. They never agreed on which one to buy. When a shopper asks an AI agent for nothing more than the product, the choice between the products on offer comes from the model’s preference, and I tested exactly that, on purpose: no budget, no taste, no shopper history, only the search words. For coffee, I searched Shopify’s catalog for “coffee beans” once, through Shop MCP, the connection Shopify opened to AI agents on Oct. 13, and saved the top 20 results. Then I gave three Claude models that same saved list, the same instruction, and the same request, 50 times each. Haiku 4.5 and Sonnet 5 favored the dark roast with the most reviews. Opus 5 picked the medium roast 48 times out of 50, and told me why: “With no stated preferences, a medium roast is the safest crowd-pleaser…”

Shopify Opened A Keyless Connection To Its Catalog For AI Agents On Oct. 13

On Oct. 13, 2026, Shopify announced Shop MCP, which, in Shopify’s words, “lets your agent or app search Shopify Catalog, build carts, and complete Shop Pay checkout through a single MCP connection.” It covers over a billion products, and searching needs no API key. Any company building a shopping agent can now search Shopify’s catalog without an API key and get back a list of products for a shopper’s request. It is one

of several routes into agentic commerce, the agents that find, compare, and buy for a shopper.

A list is not a purchase. Something has to pick one product off it, and when an agent runs on a language model, the model is what picks. The agent hands the model the products and the request and asks it to choose. The company that builds the agent writes its instructions on top, and those matter. They are also private, and they differ from one agent to the next.

So I tested the part underneath, the model by itself. Same products, same short instruction, same request, and the only thing that changed was the model. If three models given exactly the same input pick different products, then which store gets a customer's order depends partly on which model the customer's agent happens to run on.

I have covered the step before this one. In October, I found that Shopify's catalog ranks products with more reviews higher. This test starts where that one stopped: once the catalog has returned its list, what does the model choose from it?

I Ran Each Search Once And Gave 3 Claude Models The Same 20 Products Almost 2,000 Times

I ran each search once through Shop MCP, as an anonymous U.S. buyer, and saved the top 20 results. Every call to a model got that same saved list, in the catalog's order (except in the test where I reversed it), so the products never changed between models or between runs. Eight skincare searches were fetched on 15 October. Six more, backpack, coffee beans, running shoes, water bottle, wireless earbuds and yoga mat, were fetched on Oct. 24. I also fetched "scented candle" and "dog food" and dropped them, because one store filled all 20

candle results and three stores filled the dog food list.

For each product the model saw the title, the store, the price, the star rating and review count, the catalog's one-line summary and its first three listed features, numbered 1 to 20. It saw nothing else. The instruction, word for word:

> *You are a shopping agent. A person has asked you to buy something for them. You will see the products the Shopify catalog returned for their request. Choose the one product you would buy for them. Reply with JSON only, no other text: {"pick": <the product number>, "reason": "<one sentence>"}*

Then came the shopper's request, which was only the search words: "coffee beans", "water bottle", "mineral sunscreen SPF 50". In one test I added a budget and a skin type.

The models were Claude Haiku 4.5, Sonnet 5 and Opus 5, run through Claude Code with default settings, no tools and no web access. Every pick was a fresh call with no memory of the others.

There were three sets of runs, 1,980 picks in all:

- **Six product searches** with only the search words as the request: 50 picks per model per search, 900 picks.
- **One sunscreen search, asked two ways**, first "mineral sunscreen SPF 50", then "mineral sunscreen SPF 50, under $25, for oily skin": 100 picks per model each way, 600 picks.
- **Eight skincare searches** with the list shown in the catalog's order and then reversed: 10 picks per model each way per search, 480 picks.

The reason for the bare setup is the point of the test. No preferences, no shopping history, no system prompt from an agent company. This is the raw response from the model when it has no other inputs, and how it differs depending on the

model. Every real agent adds more on top, so this is the layer everything else gets built on.

Haiku, Sonnet And Opus Agreed On The Top Product In None Of Six Searches

On the six product searches, the three models’ most-picked product matched in none. Haiku and Sonnet agreed on three (backpack, coffee beans, water bottle), Sonnet and Opus on two (wireless earbuds, yoga mat), and Haiku and Opus on one (running shoes).

Image Credit: Slobodan Manic

Each model was consistent with itself, and the models were not consistent with each other. Sonnet's most-picked product took 94% of its picks on average across the six searches, Opus's took 89%, and Haiku's 67%. On any one model, one product gets nearly every pick. Change the model, and a different product does.

On three of the six searches, the different product also came from a different store. On water bottle, Haiku and Sonnet mostly picked nobl's bottle, and Opus always picked Owala's. On yoga mat, Haiku picked POPFLEX and the other two Liforme. On running shoes, Haiku and Opus picked a Brooks shoe sold by Fit2Run and Sonnet an adidas shoe sold by Commonwealth Running Co. On coffee, backpack, and earbuds both picks came from the same brand.

The catalog's ranking mattered less than I expected. Across the six searches, Haiku picked the product Shopify's catalog ranked first 8% of the time, Sonnet never, and Opus 20%.

Haiku And Sonnet Picked The Product With The Most Reviews, Even A Count The Store's Page Does Not Show

On backpack, coffee beans, and water bottle, the product Claude Haiku 4.5 and Sonnet 5 picked most was the one with the most reviews in the list of 20, and on yoga mat, Haiku's was too. Their reasons say so in the first words: "EASE Dark Roast has the highest review count (24,398 reviews) with a 4.9/5 rating…" (Haiku), "The Roll Top 20L has the highest review count by far (63,165)…" (Sonnet).

The water bottle is where that preference shows its cost. Shopify's catalog lists nobl's Stainless Steel Water Bottle at 5.0 stars from 447,221 reviews, and that is what the models saw. Haiku picked it 45 times out of 50 and quoted the 447,221

every time. Sonnet picked it 39 times and cited the figure in 33 of them, for example, “Highest rating by far (5.0 from 447,221 reviews) at a reasonable $29 price point…”.

Image Credit: Slobodan Manic

The same bottle has three different review numbers, and the two models picked it on the biggest one. nobl’s product page said 5.0 from 6,537 when I checked on 24 October. The catalog’s record for the Blush color it listed says 4.2 from 1,674.

I do not know what 447,221 counts. When I ran the search again later on 24 October, the catalog showed 447,227, six more, so the number is live. It does not appear on nobl’s product page or home page, and it is about 68 times the count nobl’s product page shows. Opus 5 never picked this bottle. It picked

the Owala FreeSip, 4.6 from 34,394 reviews, all 50 times, calling it “the safest default for a generic water bottle request…”.

The water bottle result was weird. Neither Haiku nor Sonnet checked the 447,221 figure, and neither could have from what it was given. Whatever the catalog says about a product is what the model reasons from.

Opus Called Its Choice Safe In 222 Of 300 Picks

In 222 of its 300 picks on the six product searches, Claude Opus 5’s reason called its choice “safe” or “safest”. Haiku did that once in 300, and Sonnet 11 times.

- “With no stated preferences, a medium roast is the safest crowd-pleaser…” (coffee beans)
- “With no stated specifics, the Everyday Backpack 20L is the safest all-purpose pick…” (backpack)
- “…the safest default for a generic water bottle request…” (water bottle)

Image Credit: Slobodan Manic

Given nothing to go on, Opus builds a default shopper the request never described and buys for that person. Its safe pick is the general-purpose version of the product: the everyday backpack over the roll-top, the medium roast over the dark, and on none of the six searches the one with the most reviews. On a mineral sunscreen search, it picked the most-reviewed product, a tone-adapting one with 15,414 reviews, once in 100 tries, while 17 of its 100 reasons ruled out a “tint, powder, or stick” format the shopper never asked for.

I was very surprised by how consistently Opus reached for “safe.” It is tempting to ask whether playing it safe is the better habit. I think that is the wrong question. Every model has a default. Opus says its default out loud. Haiku and

Sonnet have one too, and about half of their picks went to the product with the most reviews.

Moving A Product From First To Last Stopped The Models Picking It On 2 Of 8 Searches

On a mineral sunscreen search, moving the catalog’s top product from first to last in a list of 20 made all three Claude models almost stop picking it; Haiku and Opus did the same on lip balm, and no model’s reason mentioned the position. On eight skincare searches, I showed each model the same 20 products in two versions: In the catalog’s order, with the catalog’s top product first, and the same 20 reversed, with that product last. Each model got each version 10 times per search, and I counted how often it picked that one product.

Image Credit: Slobodan Manic

Each cell is the number of picks with the product listed first, then listed last, out of 10. Across all eight searches, out of 80 per version, Haiku went from 28 to 11, Sonnet from 15 to 11, and Opus from 24 to 10.

On mineral sunscreen, all three models almost stopped picking the catalog’s top product once it was last on the list. On lip balm, Haiku and Opus did the same, and Sonnet went the other way. On body lotion, Haiku and Opus picked the top product in both positions, and on the other five searches it was rarely picked in either. So the effect is real, strong on some searches and absent on most.

In 480 reasons, not one mentioned where a product sat in the list, and that bothers me. They cited reviews, ingredients,

and price. A shopper reading the agent’s explanation would never know the position played a part. Moving the first product to the last place should not change what gets picked. It shouldn’t be like that, in my opinion.

A Budget And A Skin Type In The Request Moved Every Model Off The $48 Top Product

When the request added “under $25, for oily skin”, Claude Haiku 4.5, Sonnet 5 and Opus 5 stopped picking the catalog’s $48 top-ranked sunscreen and picked one under $25 in 298 of their 300 picks. On a list of 20 sunscreens, I ran the same test twice, 100 picks per model each time, changing only the request: first “mineral sunscreen SPF 50”, then “mineral sunscreen SPF 50, under $25, for oily skin”.

The catalog’s top-ranked sunscreen cost $48. With the plain request, Haiku picked it 100 times out of 100, Sonnet 90, and Opus 33. With the budget and skin type in the request, all three picked it zero times. Sonnet and Opus picked a sunscreen under $25 every time, and Haiku did 98 times, twice choosing one priced at exactly $25.00.

Image Credit: Slobodan Manic

Opus’s reasons put the choice between the cheap ones down to the wording of the listings. Six of the 20 cost under $25. Sonnet picked ATTITUDE’s Mineral Sunscreen ($22.95, 4.7 from 292 reviews) all 100 times. Opus split its picks between ATTITUDE (67) and Sun Bum’s Mineral SPF 50 Sunscreen Lotion (33).

That $17.49 Sun Bum lotion had the lowest rating of the six under $25, 4.2 from 105 reviews, and its listing was the only one of the six that said: “without clogging pores.” In 31 of Opus’s 33 reasons for picking it, the reason pointed to that. Badger’s Daily Mineral Sunscreen had the best rating of the six, 4.9, and Opus never picked it. Its listing mentions “organic plant oils.”

Image Credit: Slobodan Manic

The models also give a store’s claims as their reasons. On the eight skincare searches, “clinically proven” appeared in 63 reasons, from Sonnet (39) and Haiku (24), and in 61 of them it came straight from the picked product’s listing. All 22 reasons that said a retinal product works “11x faster” than retinol were Haiku’s, and all 22 came from listings: 19 word for word, and 3 from an eye cream listing that says “11 times faster”. Haiku asked for an eye cream: “next-generation retinal technology that delivers results 11x faster than retinol”. The model has no way to check the claim, and the shopper hears it from the agent, not from the store.

The Decision Layer Between The Shopper And The Store Is A Black Box Nobody Can Monitor

There is always going to be a decision-making layer when AI agents shop: the agent asking its language model what to choose out of a list, based on a system prompt or whatever else it is given. That layer is hidden from us, hidden from the shopper and hidden from the store, and we will only ever see it as a black box. It matters to know what influences it, and that it may have ways of choosing that nobody is aware of.

Google's rankings have always been a black box too. Nobody outside Google knows exactly how they work, but a store can see them. It can monitor where it ranks and watch that change over time. Google personalizes results by location and search history, and the results are still there to check. With agents, everyone gets their own version of the decision, inside an app the store never sees, and there is nothing to monitor.

This test is the simplest version of that layer, with no shopper preferences, no history, and no agent company's instructions. It can only get more complicated from here, and the outcome will vary for everyone. The first step is to be aware of it. Trying to control it comes later, and so does trying to understand it.

4 Things A Store In Shopify's Catalog Can Do About The Decision Layer

A store cannot choose the model a customer's agent runs on, and it cannot see the decision. Four things are within reach.

1. **Check what the catalog shows an agent about your products.** Search for your product the way an agent would, through Shop MCP or Shopify's catalog, and read the rating, the review count, the one-line summary, and the listed features. The models reason from exactly that, and the water bottle shows it can differ from your product page by a factor of 68.
2. **Write listings in the words customers use when they say what they need.** When the request named a budget and a skin type, the listing that said "without clogging pores" got a third of Opus's picks with the lowest rating of the six. A need the customer names is a phrase your listing either matches or does not.
3. **Keep product claims true.** The models repeated "clinically proven" and "11x faster" to the shopper as their reasons, without any way to check them. A claim in a listing is now a claim an agent makes on your behalf.
4. **Test your category.** Run your top search through a few models with your competitors' products in the same list, and see which product each one picks. Awareness comes first, and the method above is cheap enough to repeat.

This Test Covers Three Claude Models, A Bare Instruction And Nothing Bought

- **Three Claude models only.** Haiku 4.5, Sonnet 5, and Opus 5 are all made by Anthropic. Other companies' models may pick differently, and this test does not show how.
- **Not real shopping apps.** Shopping features in apps such as ChatGPT, Gemini, or Claude Duet run their own setup around the model, which this test leaves out. This test isolates the model.
- **One instruction, no shopper context.** No preferences, no history, and in all but one test, no stated need.

- **Fixed lists from two days.** Each list is what Shopify's catalog returned to an anonymous US buyer on 15 or 24 October 2026. A signed-in shopper may see different products.
- **"Safe" was counted by searching for each reason** for the words "safe" or "safest." Every prompt and every reply is saved.
- **Nothing was bought.** The models picked. No order was placed.

A store used to compete for a shopper's attention. With agents, it also competes for a model's preference, and there is no ranking to check.

More Resources:

This post was originally published on No Hacks.

Featured Image: Koupei Studio/Shutterstock

https://www.searchenginejournal.com/3-ai-models-shopped-shopifys-catalog-each-picked-a-different-winner/594298/