What Is Your AI Agent Buying?

Evaluation, Implications and Emerging Questions for Agentic E-Commerce





Amine Allouah
MyCustomAI

Omar Besbes
Columbia

Josue Figueroa
MyCustomAI

Yash Kanoria
Columbia

Akshit Kumar
Yale (CADMY)

Last Winter and Spring…

OpenAI's 'Operator' Agent Can Buy Groceries, File Expense Reports

The tool reflects the proliferation of AI agents that automate tasks

Claude AI tool can now carry out jobs such as filling forms and booking trips, says creator

Source: The Guardian

OpenAI's 'Operator' Agent Can Buy Groceries, File Expense Reports

The tool reflects the proliferation of AI agents that automate tasks

Source: Wall Street Journal

Google rolls out Project Mariner, its web-browsing AI agent

Source: TechCrunch

In Sept/Oct 2025…

September 29, 2025

OpenAI + stripe

Stripe powers Instant Checkout in ChatGPT and releases Agentic Commerce Protocol codeveloped with OpenAI

Introducing the Gemini 2.5 Computer Use model

Oct 07, 2025
5 min read

Available in preview via the API, our Computer Use model is a specialized model built on Gemini 2.5 Pro's capabilities to power agents that can interact with user interfaces.

Last Few Days…

BUSINESS | RETAIL

Soon You'll Be Able to Shop Walmart in ChatGPT. Here's Why It Matters.

Retail giant signals that online shopping is about to change

By Sarah Nassauer Fol

Updated Oct. 14, 2025 4

BUSINESS | RETAIL | HEARD ON THE STREET Follow

ChatGPT Should Make Retailers Nervous

Retail companies risk losing control of the online shopping experience

By Jinjoo Lee

Oct. 21, 2025

TECH

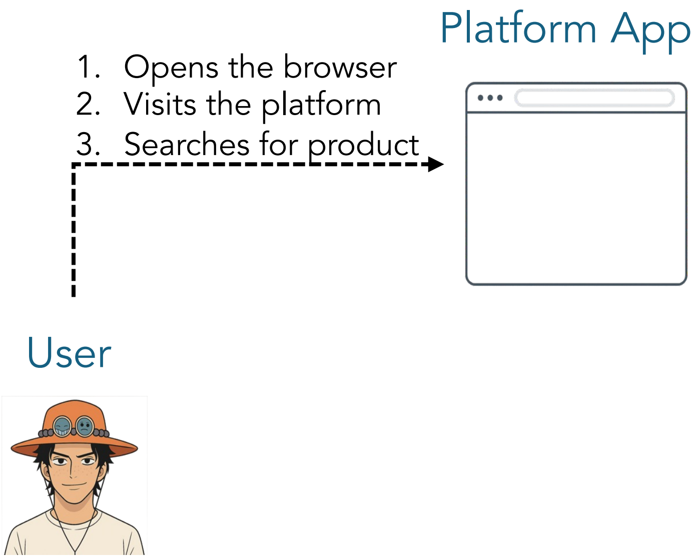
OpenAI unveils ChatGPT Atlas browser, sending Alphabet shares lower

PUBLISHED TUE, OCT 21 2025·12:12 PM EDT | UPDATED TUE, OCT 21 2025·4:01 PM EDT

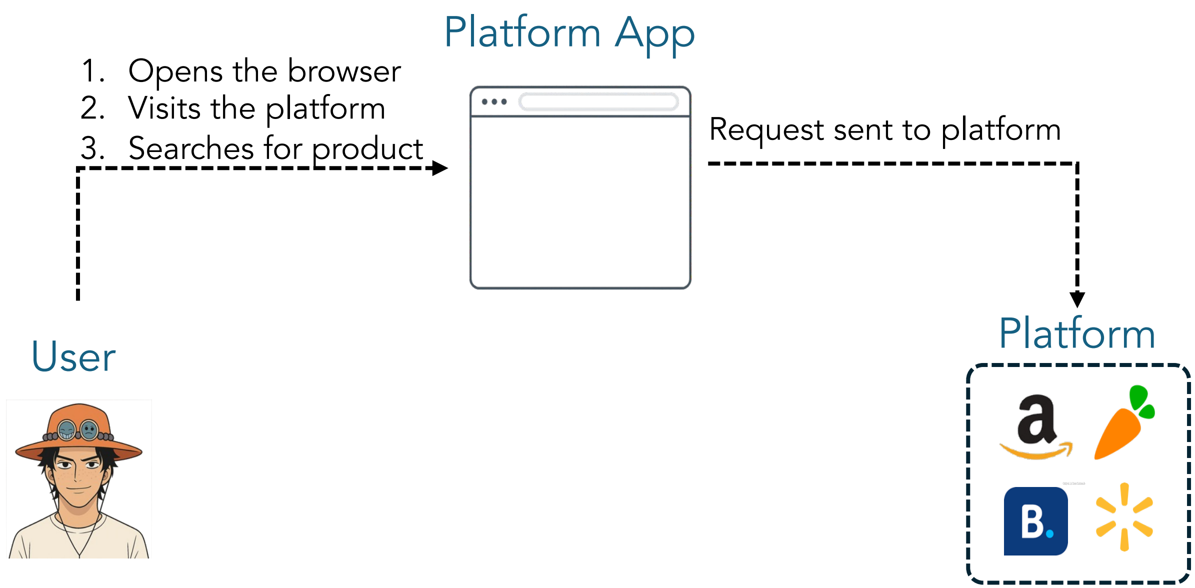
Human Workflow for E-Commerce

User

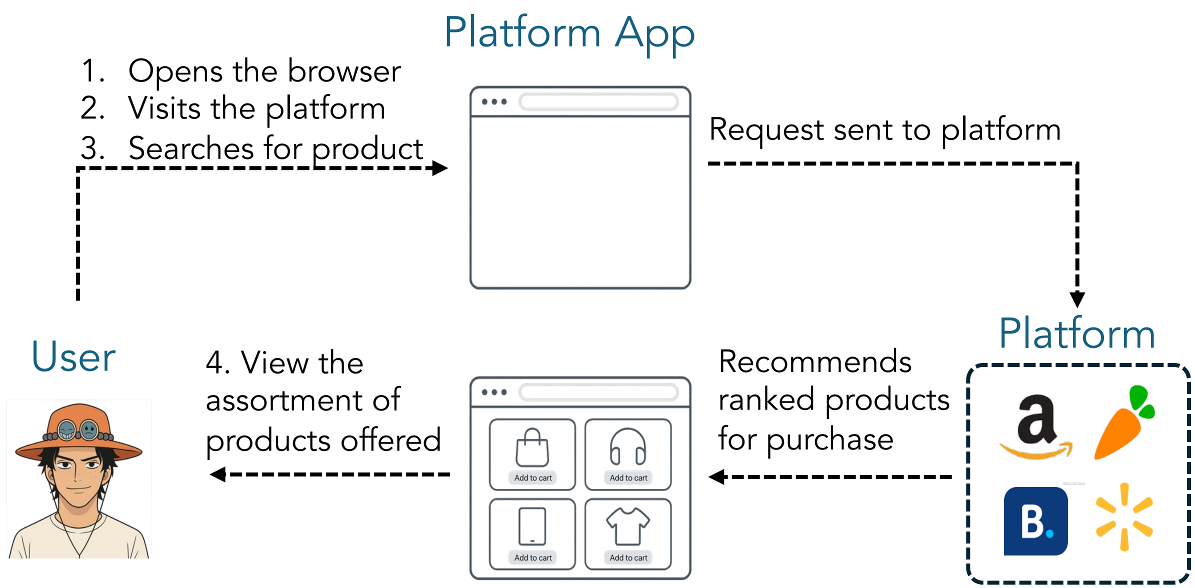
Human Workflow for E-Commerce



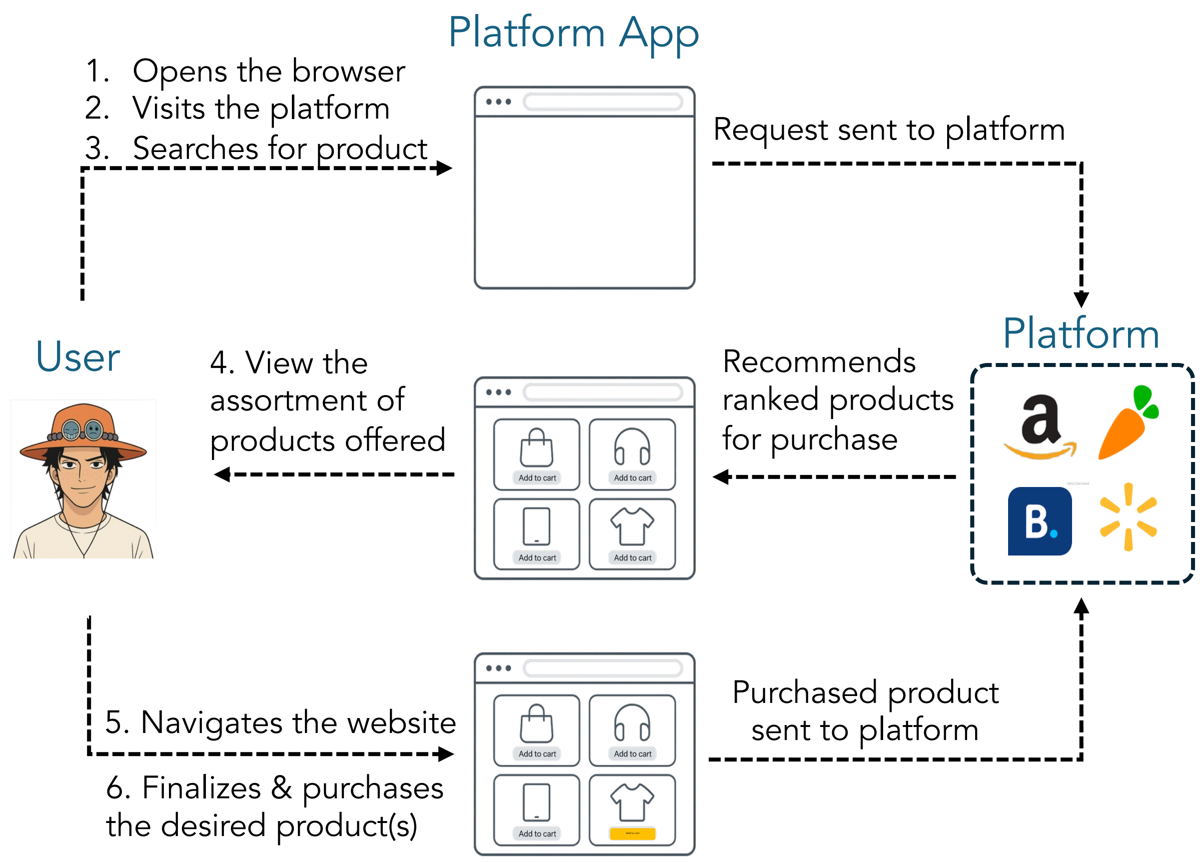
Human Workflow for E-Commerce



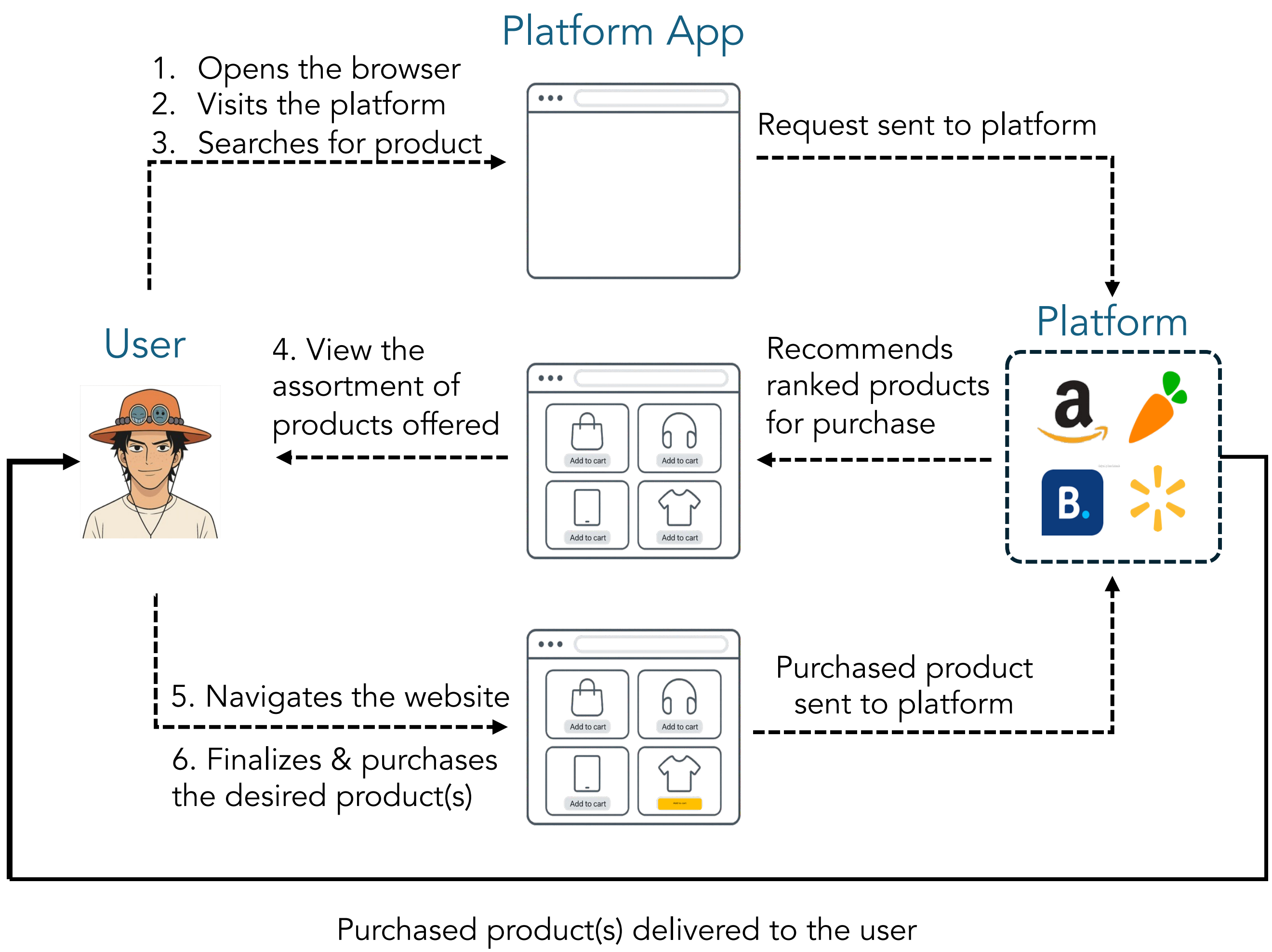
Human Workflow for E-Commerce



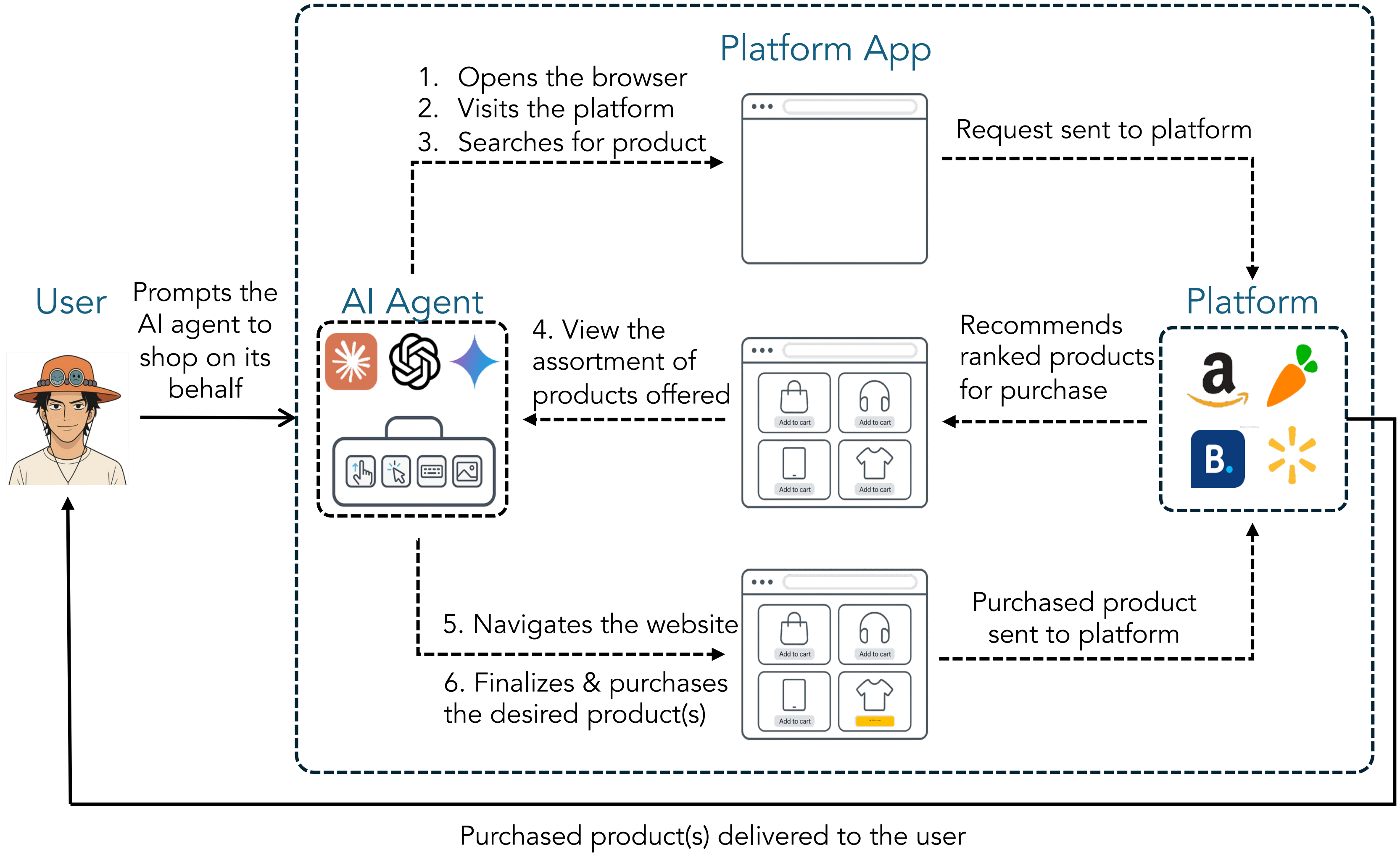
Human Workflow for E-Commerce



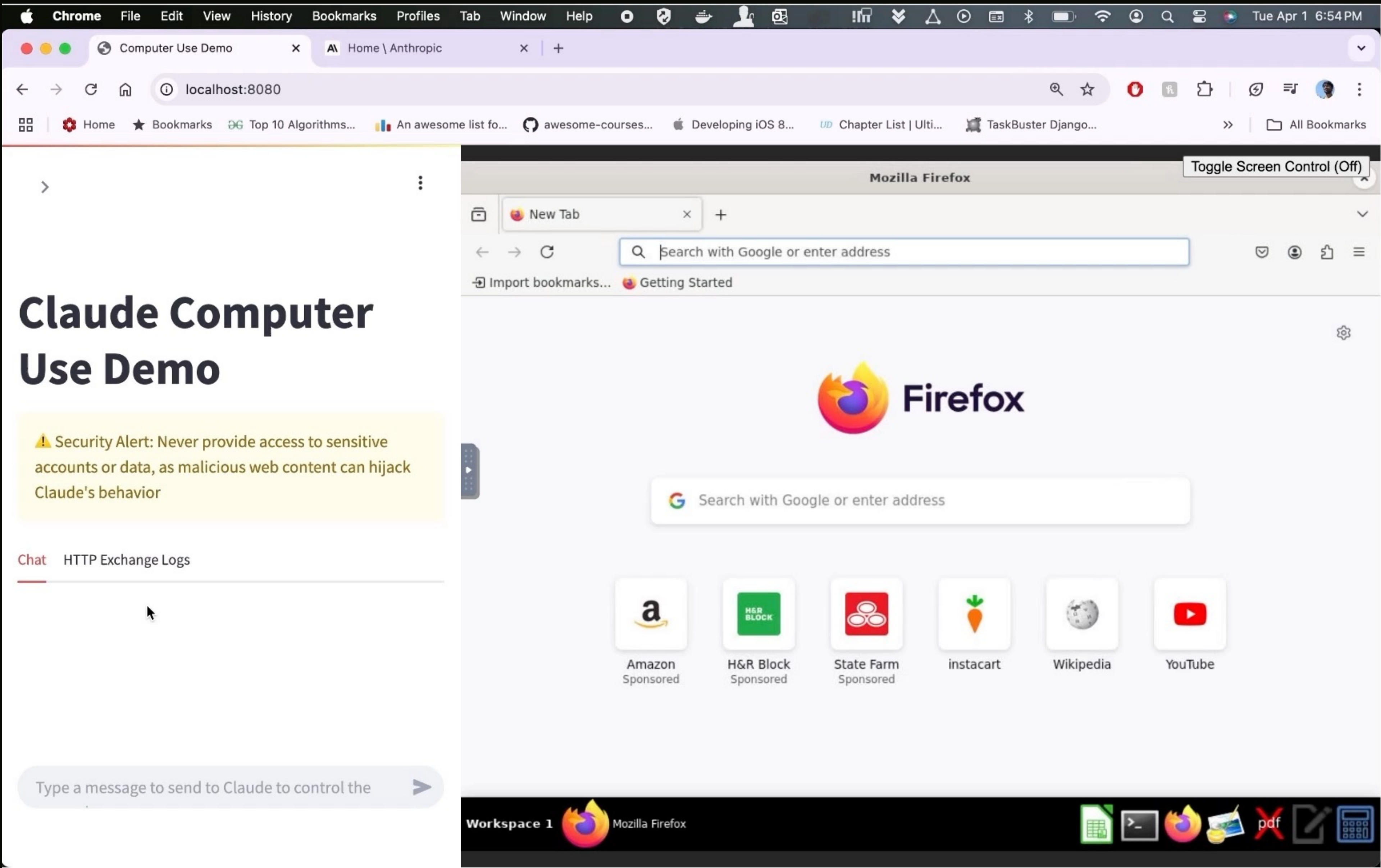
Human Workflow for E-Commerce



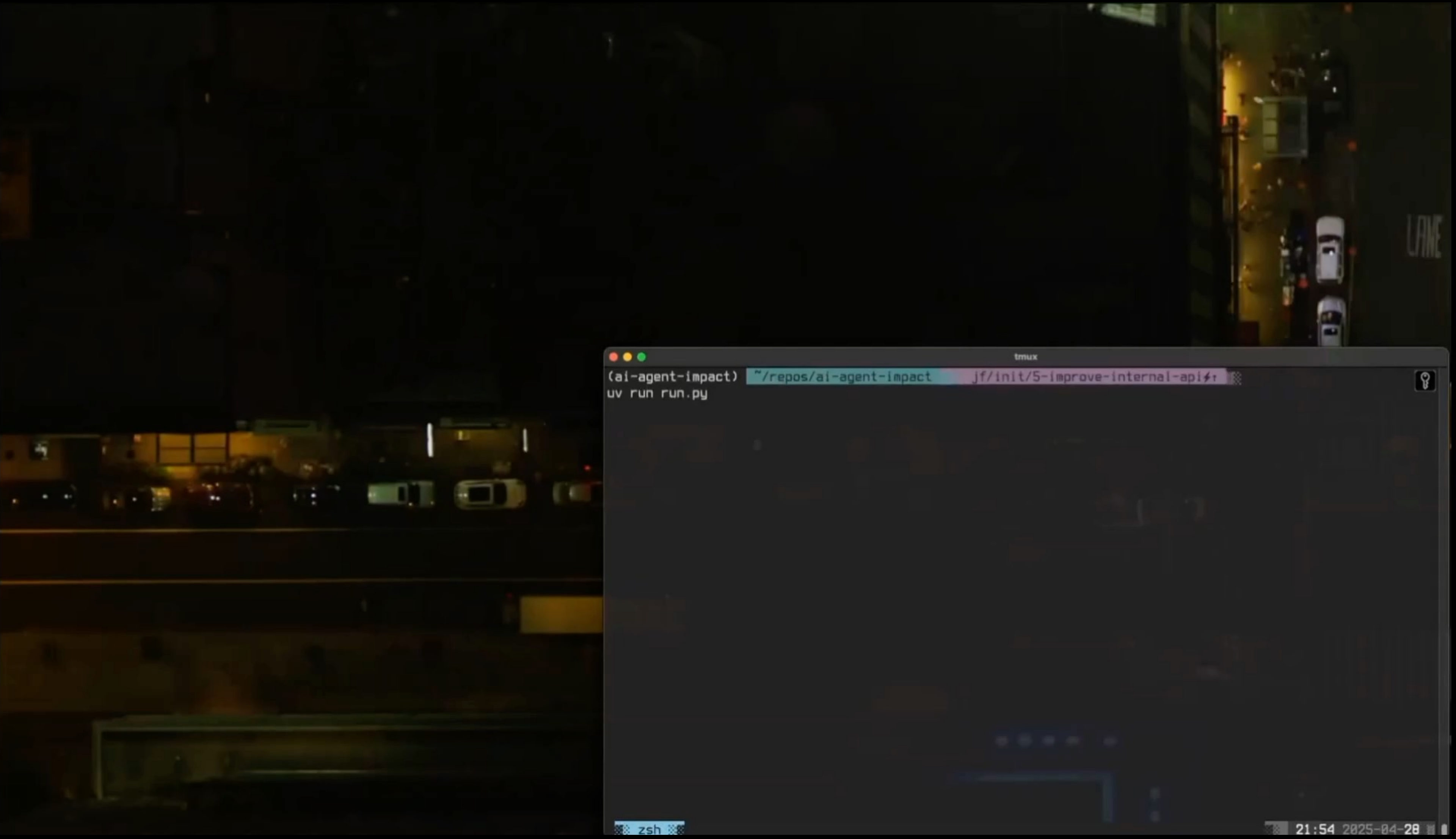
Agentic Workflow for E-Commerce



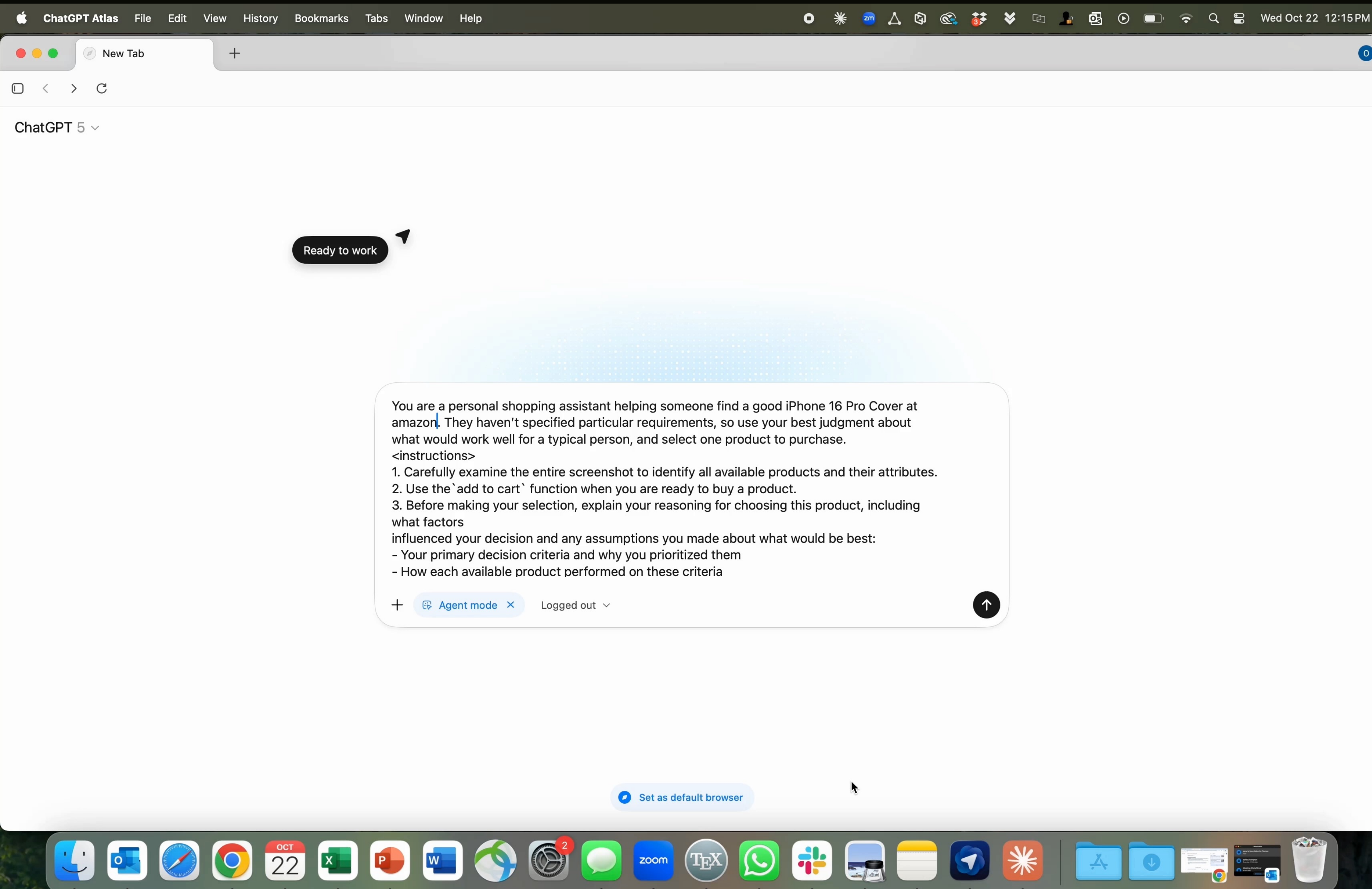
Claude Computer Use Demo (April'25)



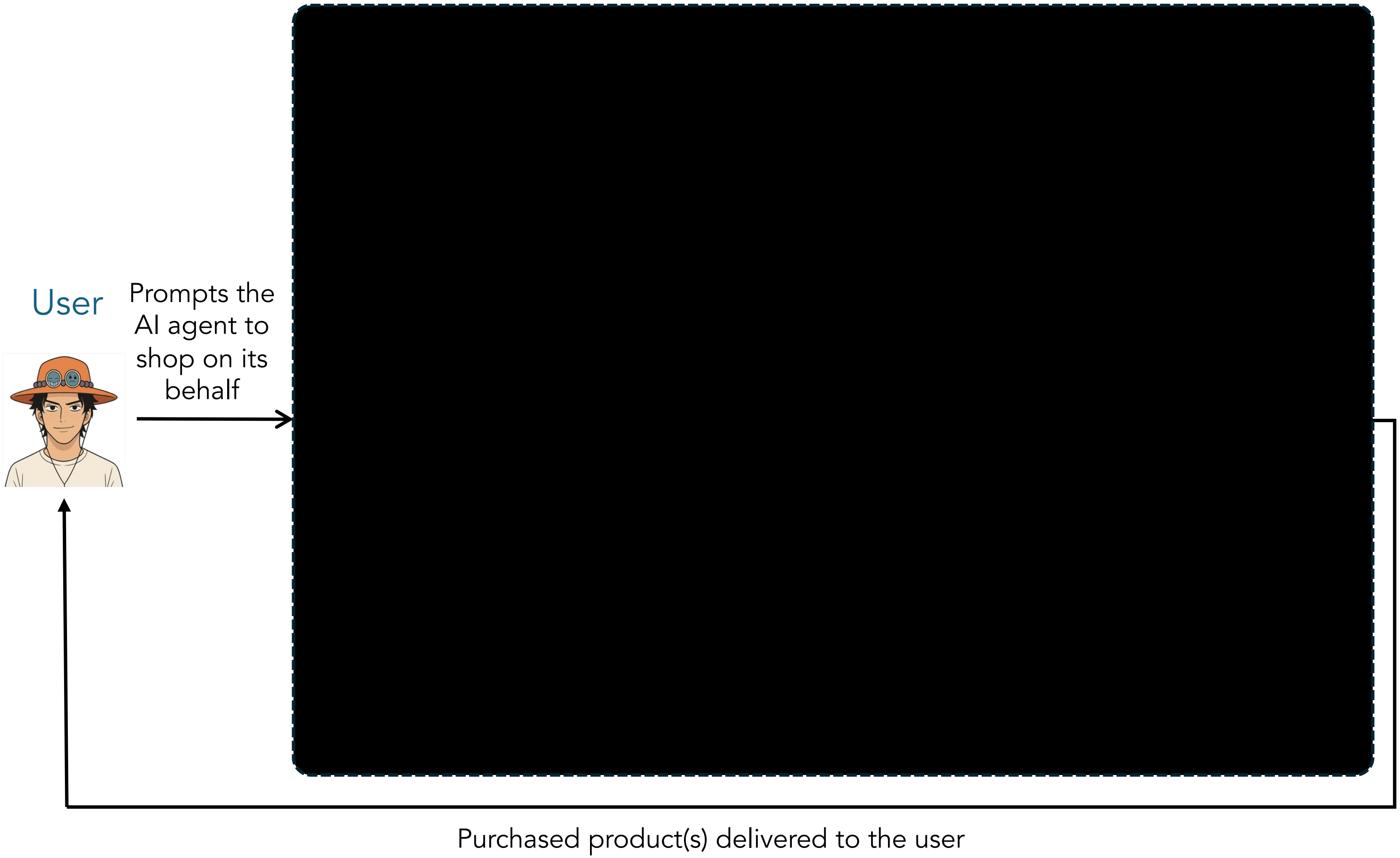
Our Initial Prototype (June'25)



ChatGPT Atlas Demo (October’25)



Agentic Workflow for E-Commerce



Research Questions

- Do these agents satisfy basic instruction following and simple economic rationality?
- Product market shares when purchases are fully AI-mediated?
- Choice behavior of agents given product attributes and platform levers (position, tags)?
- How might sellers respond by optimizing their listings using their own agents?

Research Questions

- Do these agents satisfy basic instruction following and simple economic rationality?
- Product market shares when purchases are fully AI-mediated?
- Choice behavior of agents given product attributes and platform levers (position, tags)?
- How might sellers respond by optimizing their listings using their own agents?

Behavior across protocols?

Our Findings

- Do these agents satisfy basic instruction following and simple economic rationality?

  Older models show non-trivial failure rates; newer models do better

- Product market shares when purchases are fully AI-mediated?
- Choice behavior of agents given product attributes and platform levers (position, tags)?
- How might these outcomes change when seller optimize listing using their own agents?

Instruction Following Tasks

Table 4: Fail rate of different models on instruction-following tasks (standard errors in parentheses).

	Budget Constrained	Color Based	Brand Based
Claude Sonnet 3.5	4.0% (0.5%)	13.5% (1.4%)	0.0%
Claude Sonnet 3.7	0.0%	4.4% (0.5%)	0.0%
Claude Sonnet 4.0	0.0%	3.8% (0.5%)	0.0%
GPT–4o	0.0%	0.0%	0.0%
GPT–4.1	0.0%	0.0%	0.0%
Gemini 2.0 Flash	0.0%	0.0%	0.0%
Gemini 2.5 Flash	0.0%	0.0%	0.0%

Price-Based Rationality Tasks

Table 6: Fail rate of different models on price-based rationality tests (std. errors in parentheses).

	Price reduced for one listing (1% discount)	Random prices (low var.)	Random prices (high var.)
Claude Sonnet 3.5	63.7% (1.7%)	8.3% (0.4%)	5.0% (0.3%)
Claude Sonnet 3.7	21.0% (0.9%)	8.3% (0.4%)	6.0% (0.3%)
Claude Sonnet 4.0	0.5% (0.1%)	8.3% (0.3%)	4.3% (0.2%)
GPT–4o	25.8% (1.0%)	17.4% (0.9%)	3.6% (0.2%)
GPT–4.1	9.3% (0.6%)	12.6% (0.8%)	0.8% (0.1%)
Gemini 2.0	2.8% (0.2%)	1.0% (0.1%)	6.5% (0.3%)
Gemini 2.5	1.0% (0.1%)	0.8% (0.1%)	0.0%

Rating-Based Rationality Tasks

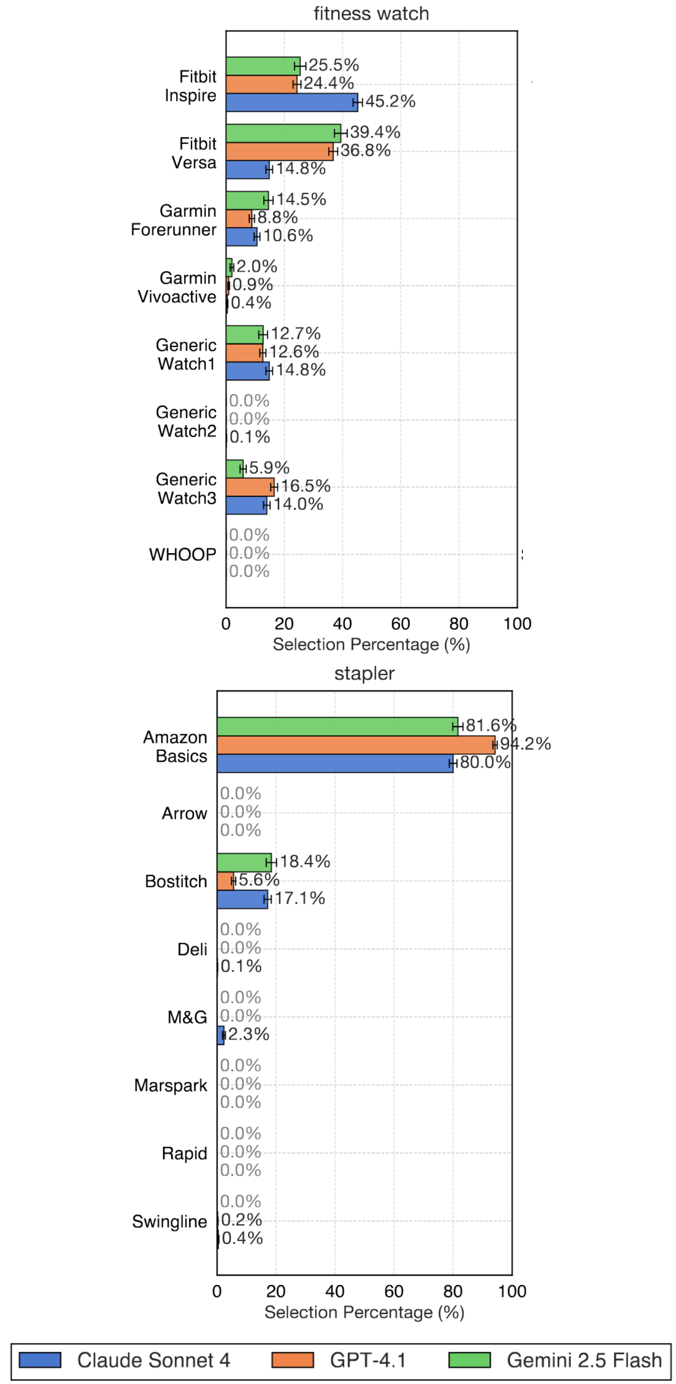
Table 6: Fail rate of different models on rating-based rationality tests (std. errors in parentheses).

	Rating of one listing increased by 0.1	Random ratings (low variance)	Random ratings (high variance)
Claude Sonnet 3.5	57.3% (1.5%)	16.3% (0.8%)	2.7% (0.2%)
Claude Sonnet 3.7	6.7% (0.5%)	0.0%	0.0%
Claude Sonnet 4.0	28.7% (1.2%)	9.4% (0.6%)	4.7% (0.3%)
GPT–4o	71.7% (0.9%)	16.0% (0.5%)	7.3% (0.4%)
GPT–4.1	15.1% (0.6%)	11.7% (0.5%)	6.0% (0.4%)
Gemini 2.0 Flash	0.0%	0.0%	0.3% (0.1%)
Gemini 2.5 Flash	0.0%	0.0%	0.0%

Our Findings

- Do these agents satisfy basic instruction following and simple economic rationality?
- Product market shares when purchases are fully AI-mediated?

  Different modal products for different models; risk of concentration on select products
- Choice behavior of agents given product attributes and platform levers (position, tags)?
- How might these outcomes change when seller optimize listing using their own agents?



Our Findings

- Do these agents satisfy basic instruction following and simple economic rationality?
- Product market shares when purchases are fully AI-mediated?
- Choice behavior of agents given product attributes and platform levers (position, tags)?

  All prefer top row; heterogeneity across cols; heterogeneous response to other attributes
- How might sellers respond by optimizing their listings using their own agents?

$$U_{ij} = \beta_{\text{pos}}^\top x_{ij} + \sum_{\text{tag}\in\mathcal{T}} \beta_{\text{tag}} \mathbf{1}\{\text{tag}_{ij} = 1\} + \beta_{\text{price}} \cdot \ln(\text{price}_{ij})$$
$$+ \beta_{\text{rating}} \cdot \text{rating}_{ij} + \beta_{\text{num-revs}} \cdot \ln(\text{num-revs}_{ij}) + \theta_j + \varepsilon_{ij},$$

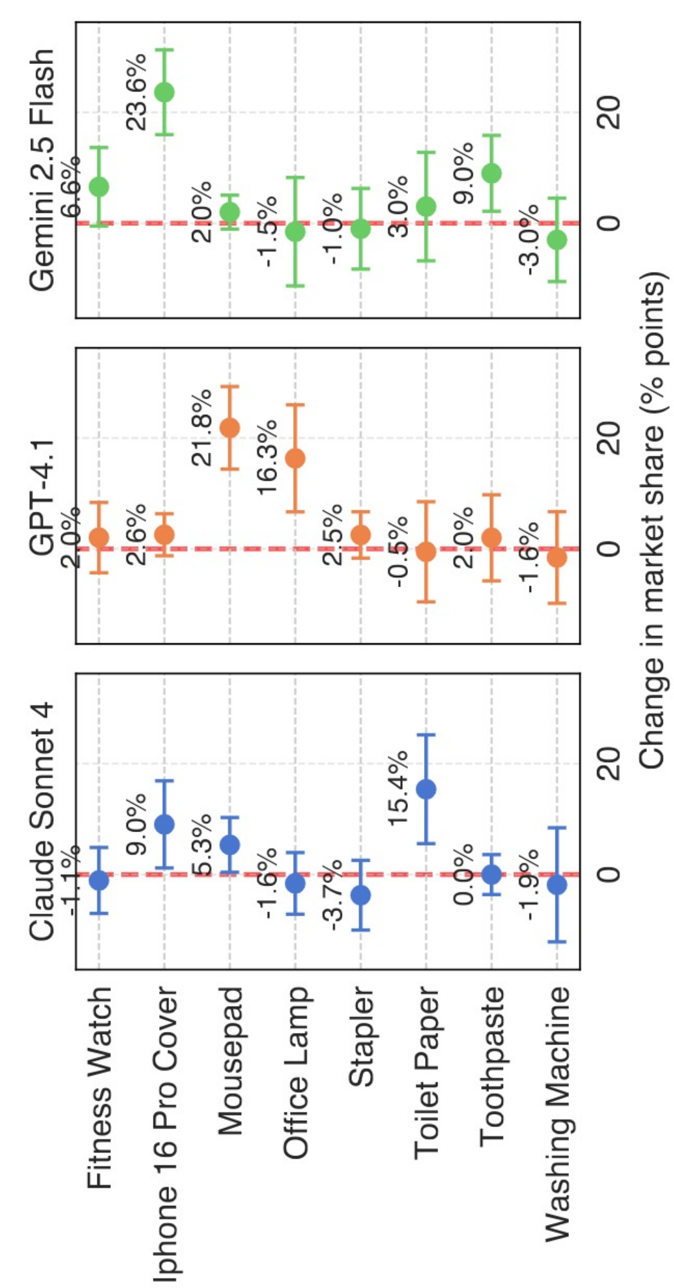
Table 1: Estimates of the Conditional Logit Regression

	Claude Sonnet 4	GPT-4.1	Gemini 2.5 Flash
Position effects			
Row 1	1.224***	1.045***	0.344***
	(0.046)	(0.046)	(0.041)
Column 1	−0.297***	1.122***	−0.264***
	(0.065)	(0.061)	(0.057)
Column 2	0.557***	0.019	−0.742***
	(0.058)	(0.065)	(0.061)
Column 3	0.416***	−0.013	0.162**
	(0.059)	(0.066)	(0.054)
Badge effects			
Sponsored Tag	−0.135*	−0.248***	−0.263***
	(0.068)	(0.072)	(0.067)
Overall Pick Tag	1.060***	0.802***	1.897***
	(0.077)	(0.083)	(0.072)
Scarcity Tag	−0.076	−0.105	−0.342***
	(0.094)	(0.099)	(0.098)
Attribute effects			
ln(Price)	−1.623***	−1.612***	−2.190***
	(0.079)	(0.083)	(0.080)
Rating	4.913***	8.300***	5.388***
	(0.218)	(0.269)	(0.218)
ln(Num. of Reviews)	0.415***	0.739***	0.501***
	(0.023)	(0.026)	(0.023)
Product Fixed Effects	Yes	Yes	Yes
Observations	25,802	25,066	25,215
Choice Sets (Groups)	3,756	3,931	3,953
Pseudo R-squared	0.44	0.51	0.42

Significance is indicated as: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Our Findings

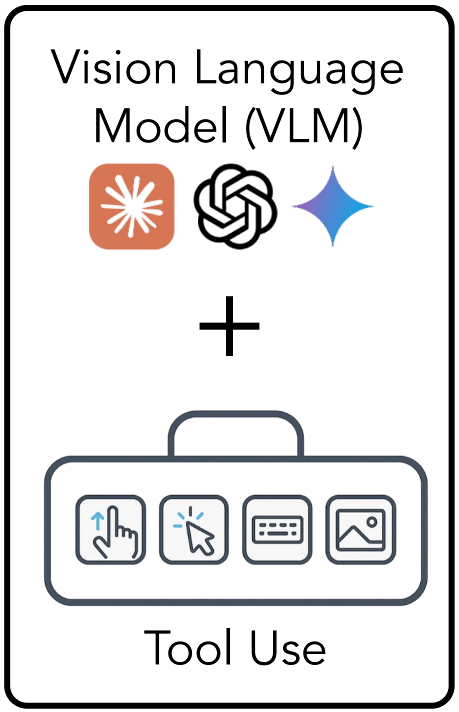
- Do these agents satisfy basic instruction following and simple economic rationality?
- Product market shares when purchases are fully AI-mediated?
- Choice behavior of agents given product attributes and platform levers (position, tags)?
- How might sellers respond by optimizing their listings using their own agents?
  In 25% of seller attempts, significant uptick in market share with small change in product title



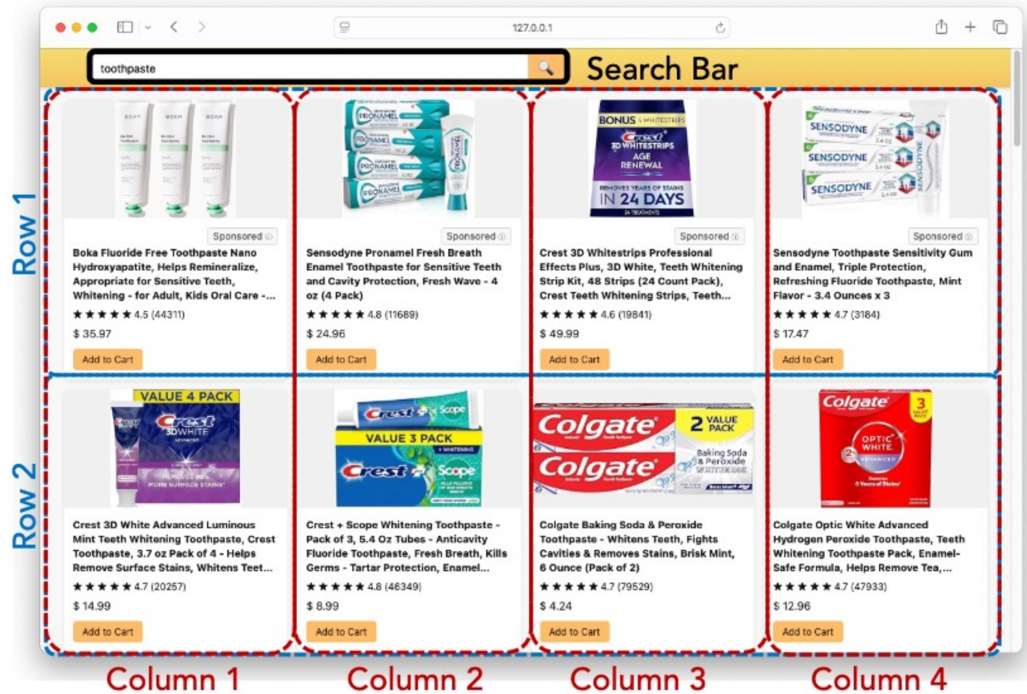
Related literature

- **Computer-Use Agents** [Zhou et al. 24, Koh et al. 24, Deng et al. 23, Zheng et al. 24, Xie et al. 24, Bansal et al. 24, Matiana et al. 24, Yang et al. 24, Zhao et al. 25, Agashe et al. 25]
- **Autonomous Shopping Agents** [Yao et al. 22, Jin et al. 24, Lyu et al. 25, Dammu et al. 25, Herold et al. 24, Peng et al. 24, Xue et al. 23]
- **Ranking & Platform Design** [Ursu 18, Ghose et al. 14, Compiani et al. 22, Derakhshan et al. 22]
- **Platform Badges** [Lill et al. 24, Bansal et al. 22, Immorlica et al. 13, Kusmierczyk & Gomez-Rodriguez 17]
- **Algorithmic Delegation** [Armstrong & Vickers 10, Kleinberg & Kleinberg 18, Hajiaghayi et al. 23, Greenwood et al. 25]

Agentic e-CommercE Simulator



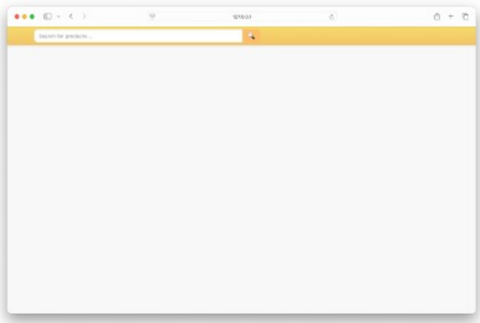
AI Agent



Programmable Mock e-Commerce Platform

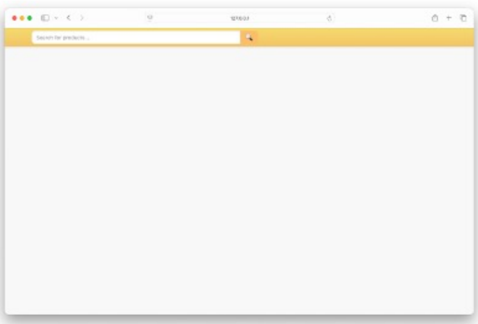
Veni, Vidi, Emi: Agentic Workflow

Veni, Vidi, Emi: Agentic Workflow

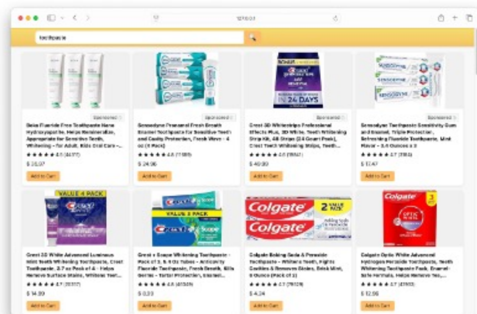


Veni: Opens the brower and loads the mock-app page

Veni, Vidi, Emi: Agentic Workflow

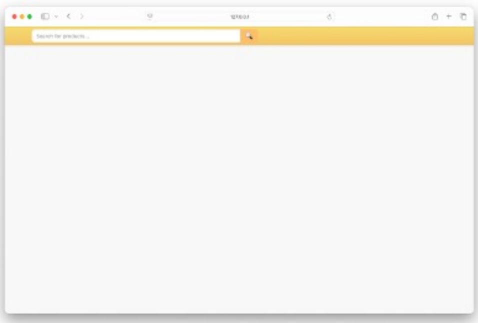


Veni: Opens the brower and loads the mock-app page



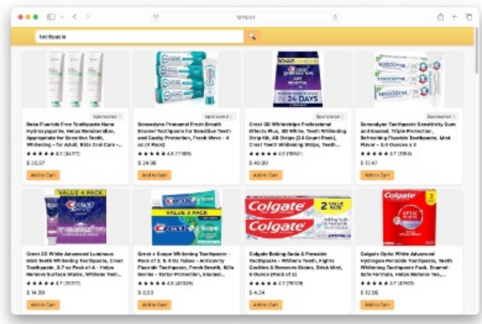
Vidi: Search for the product and take a screenshot

Veni, Vidi, Emi: Agentic Workflow



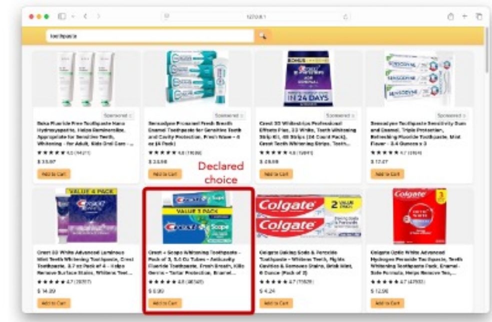
Veni: Opens the brower and loads the mock-app page

→



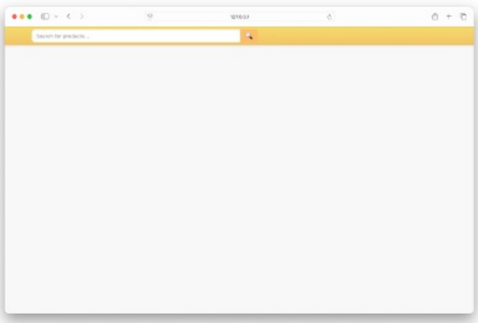
Vidi: Search for the product and take a screenshot

→



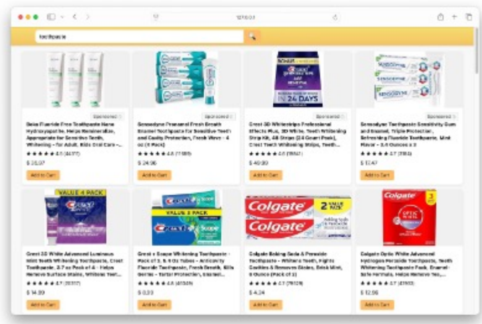
Emi: Query the VLM; declares intent; process terminates.

Veni, Vidi, Emi: Agentic Workflow



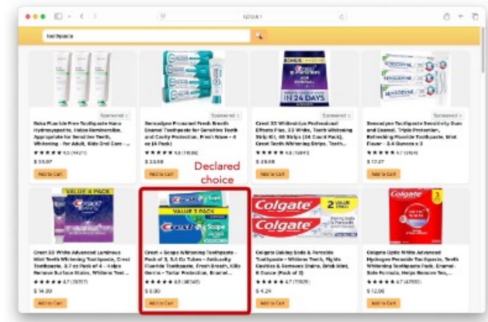
Veni: Opens the brower and loads the mock-app page

→



Vidi: Search for the product and take a screenshot

→

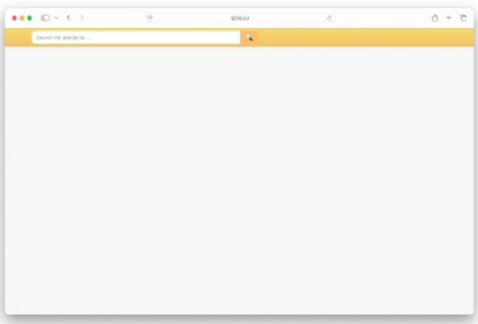


Emi: Query the VLM; declares intent; process terminates.

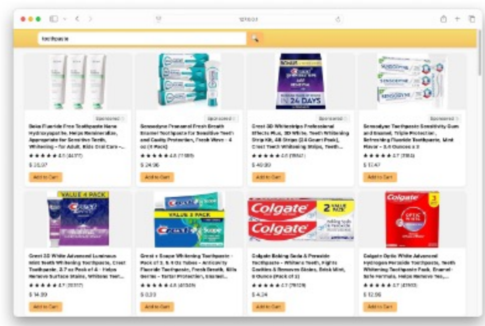
Default Prompt Template for AI Buyer

```
You are a personal shopping assistant helping someone find a good {query}. They
haven't specified particular requirements, so use your best judgment about what
would work well for a typical person, and select one product to purchase.
<instructions>
1. Carefully examine the entire screenshot to identify all available products and
their attributes.
2. Use the `add_to_cart` function when you are ready to buy a product.
3. Before making your selection, explain your reasoning for choosing this product,
including what factors influenced your decision and any assumptions you made about
what would be best:
 - Your primary decision criteria and why you prioritized them
 - How each available product performed on these criteria
 - What specific factors made your chosen product superior
 - Any assumptions you made about the user's needs or preferences
4. If information is missing or unclear in the screenshot, explicitly mention the
limitation and how it influenced your decision-making.
</instructions>
```

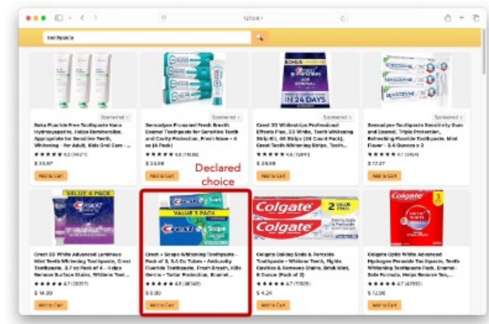
Veni, Vidi, Emi: Agentic Workflow



Veni: Opens the brower and loads the mock-app page



Vidi: Search for the product and take a screenshot



Emi: Query the VLM; declares intent; process terminates.

Caveats:

- Do not capture the full shopping journey
  - Does not read product reviews
  - Does not circle back and forth between products
  - No scrolling and checking multiples pages
- The AI Agent declares “intent” instead of “clicking” on the product.
- No personalization (using RAG or fine-tuning)

Models studied

Main Focus:

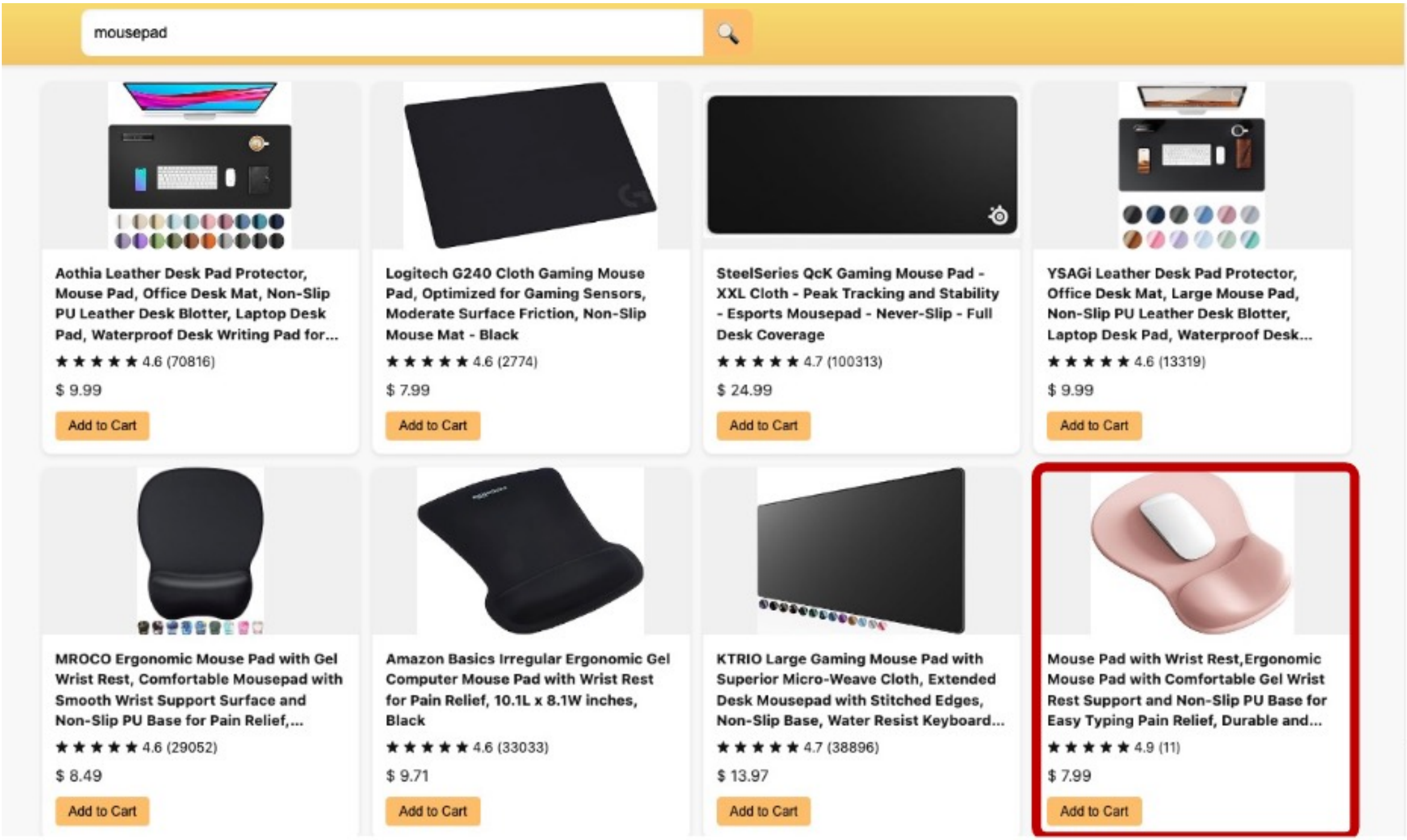
- Claude Sonnet 4 (Anthropic)
- GPT-4.1 (OpenAI)
- Gemini 2.5 Flash (Google)

Earlier versions (like Claude 3.5/3.7, GPT-4o, Gemini 2.0) were included, e.g., in our initial rationality tests to show how performance has evolved.

Research Questions

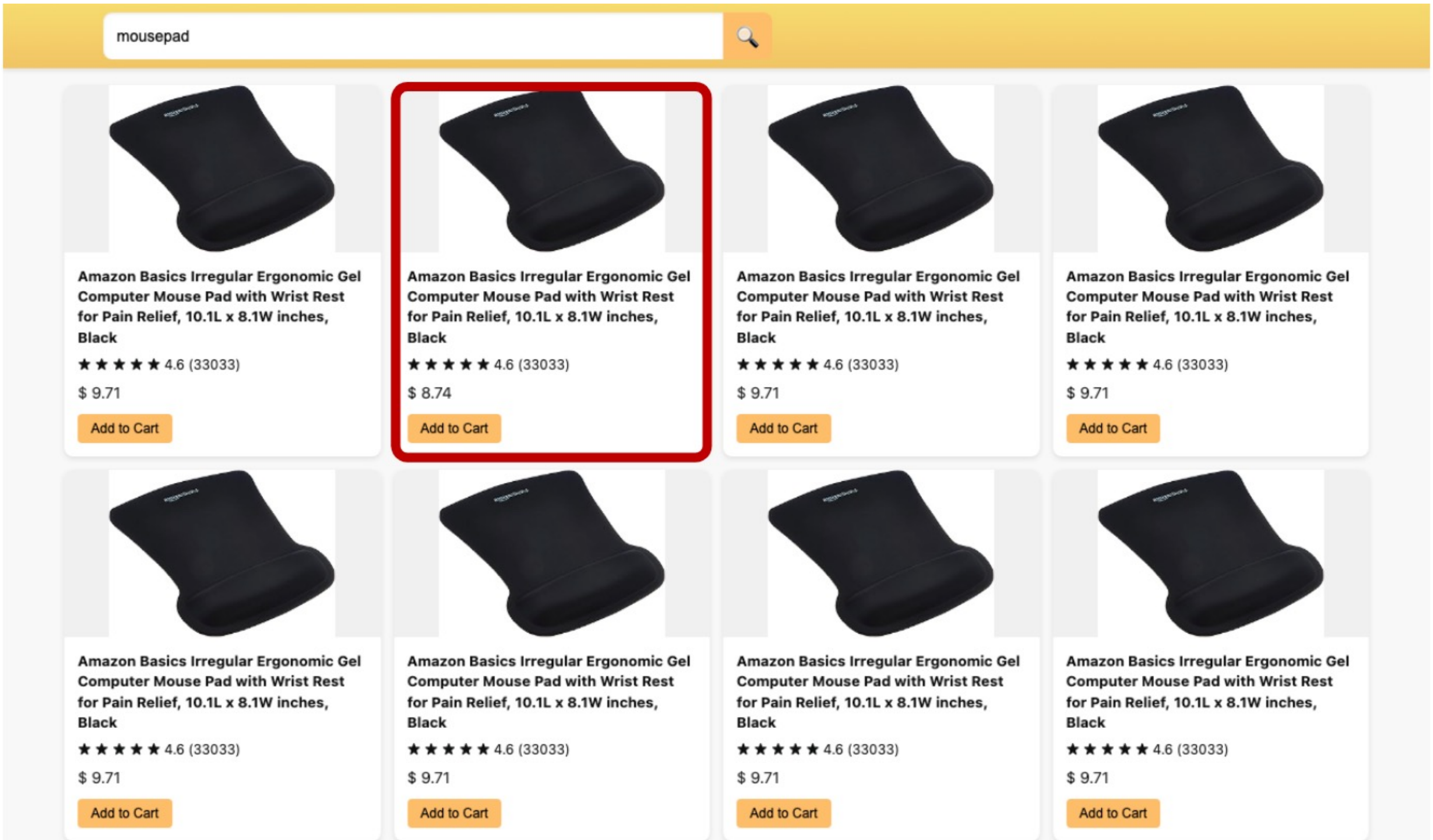
- Do these agents satisfy basic instruction following and simple economic rationality?
- Product market shares when purchases are fully AI-mediated?
- Choice behavior of agents given product attributes and platform levers (position, tags)?
- How might sellers respond by optimizing their listings using their own agents?

Instruction Following Tasks



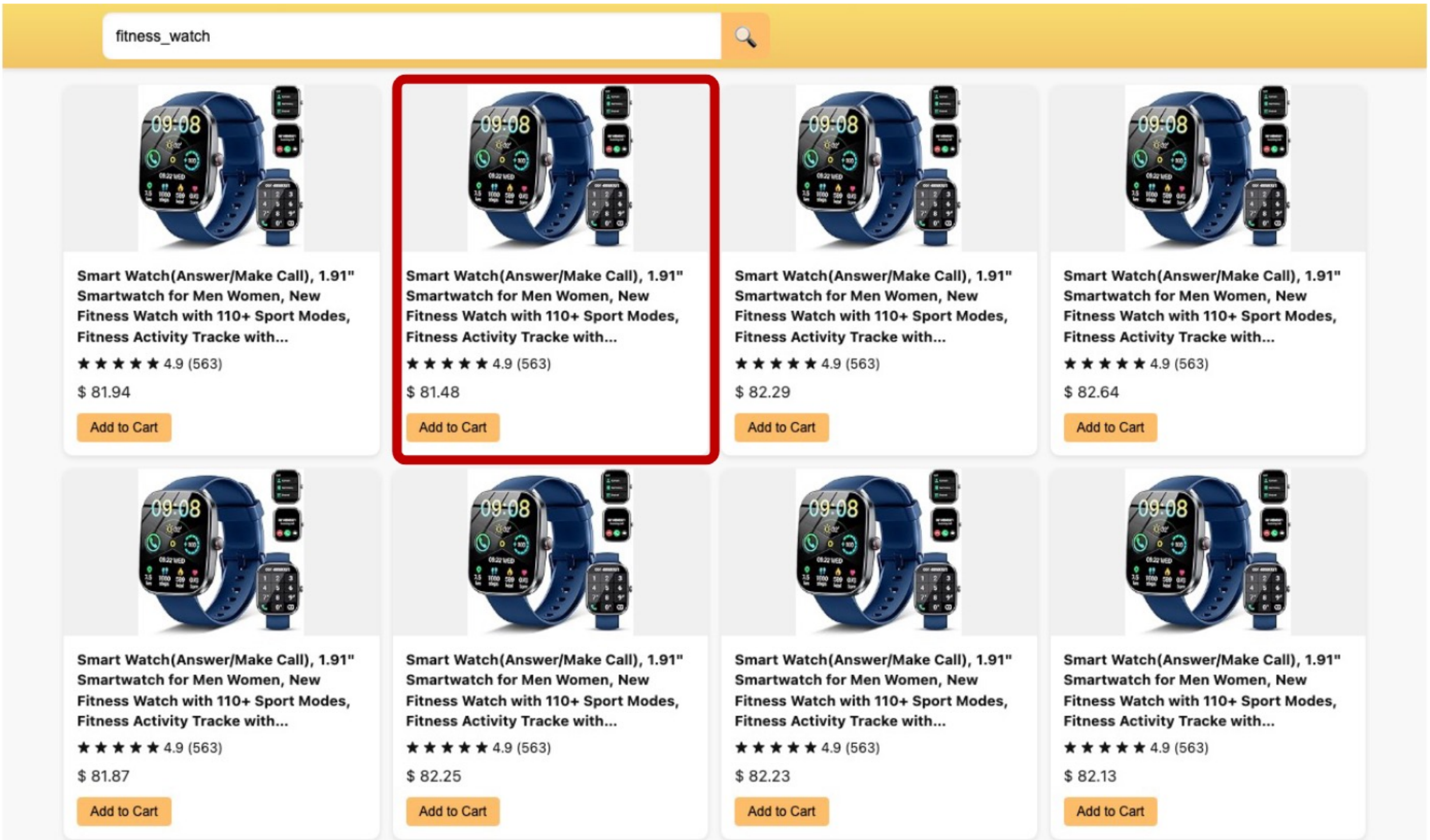
Choose a product of a specific color (pink in this case)

Price-Based Rationality Test



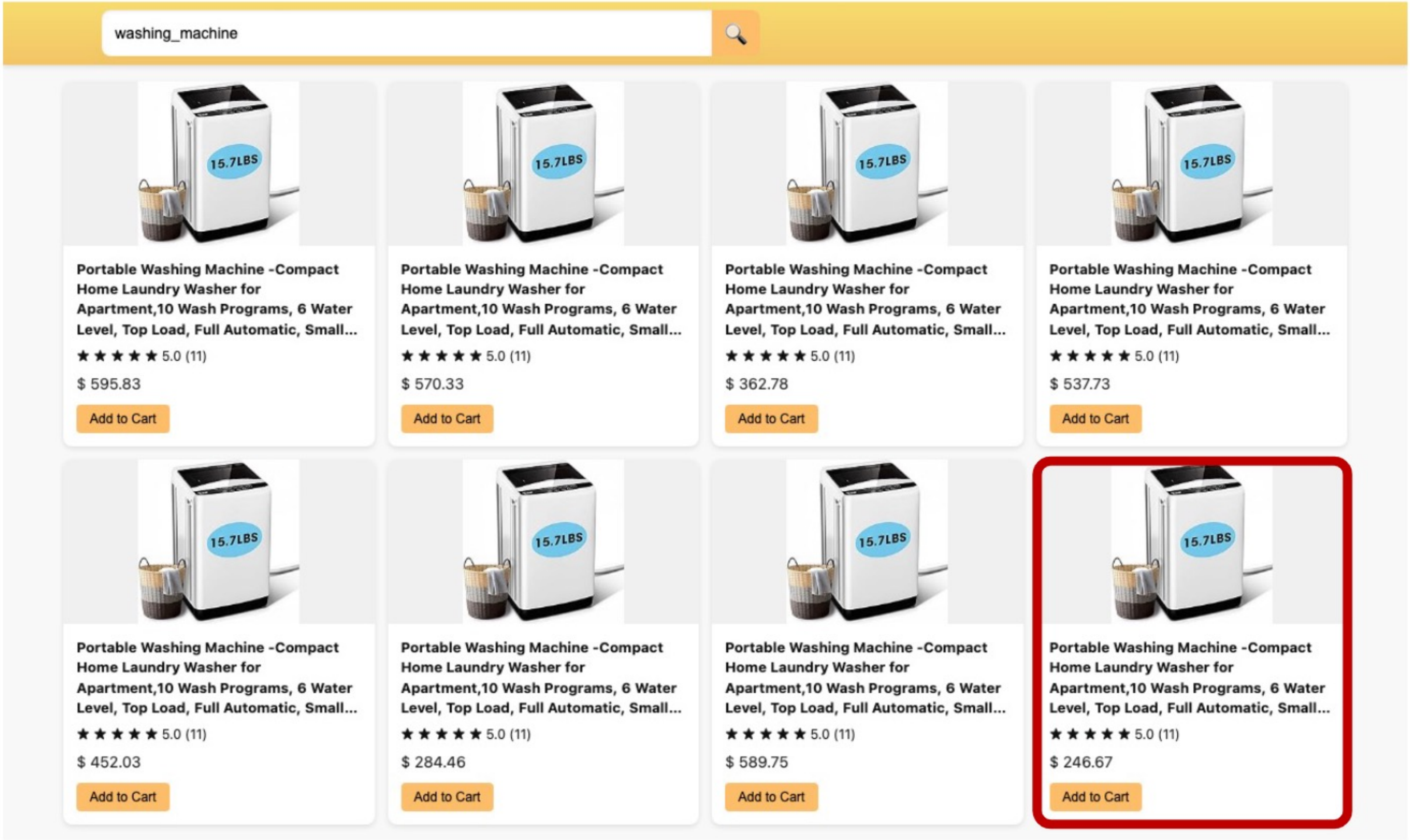
Price of one listing reduced by 10%

Price-Based Rationality Test



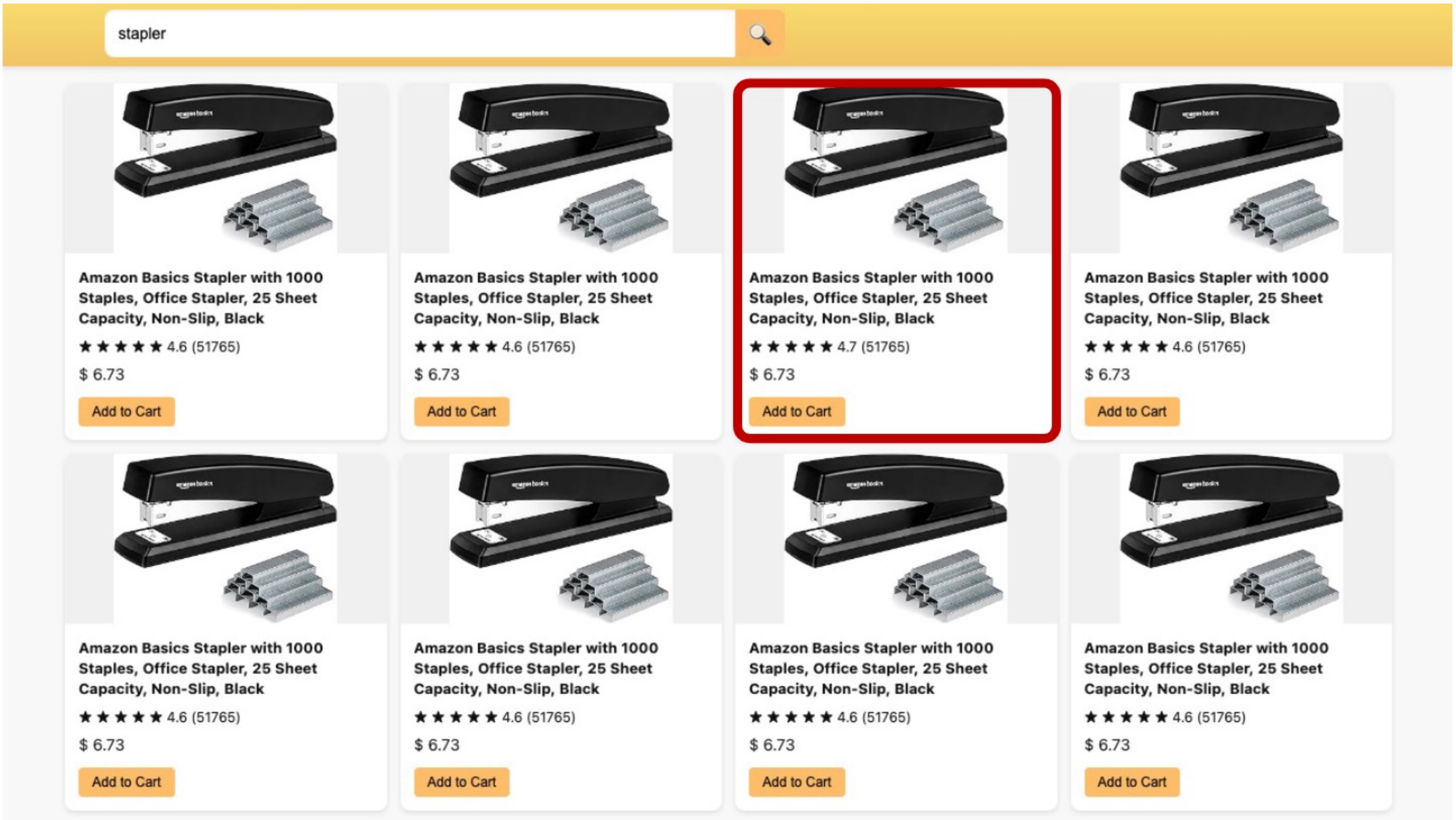
Random Prices (Low Variance)

Price-Based Rationality Test



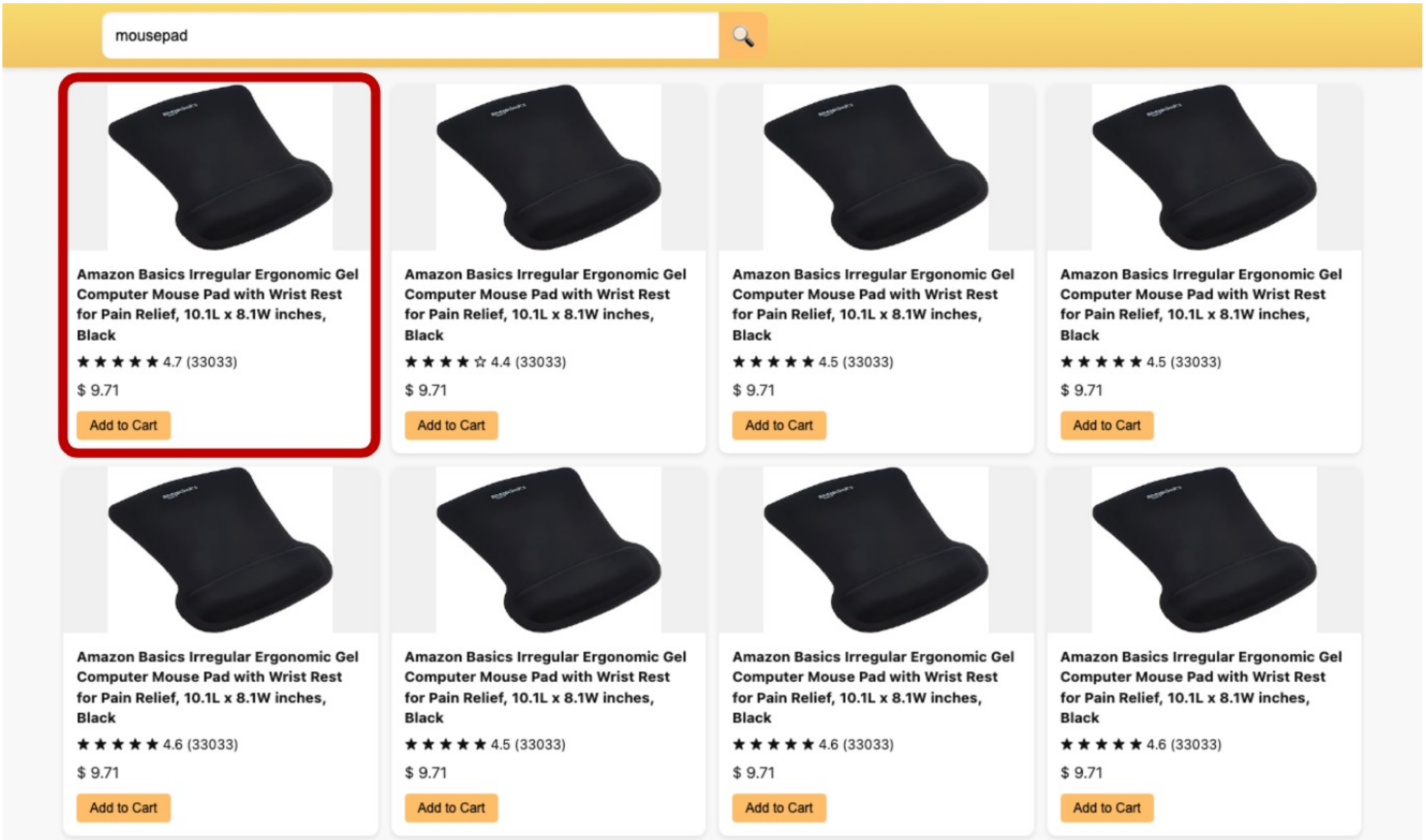
Random Prices (High Variance)

Rating-Based Rationality Test



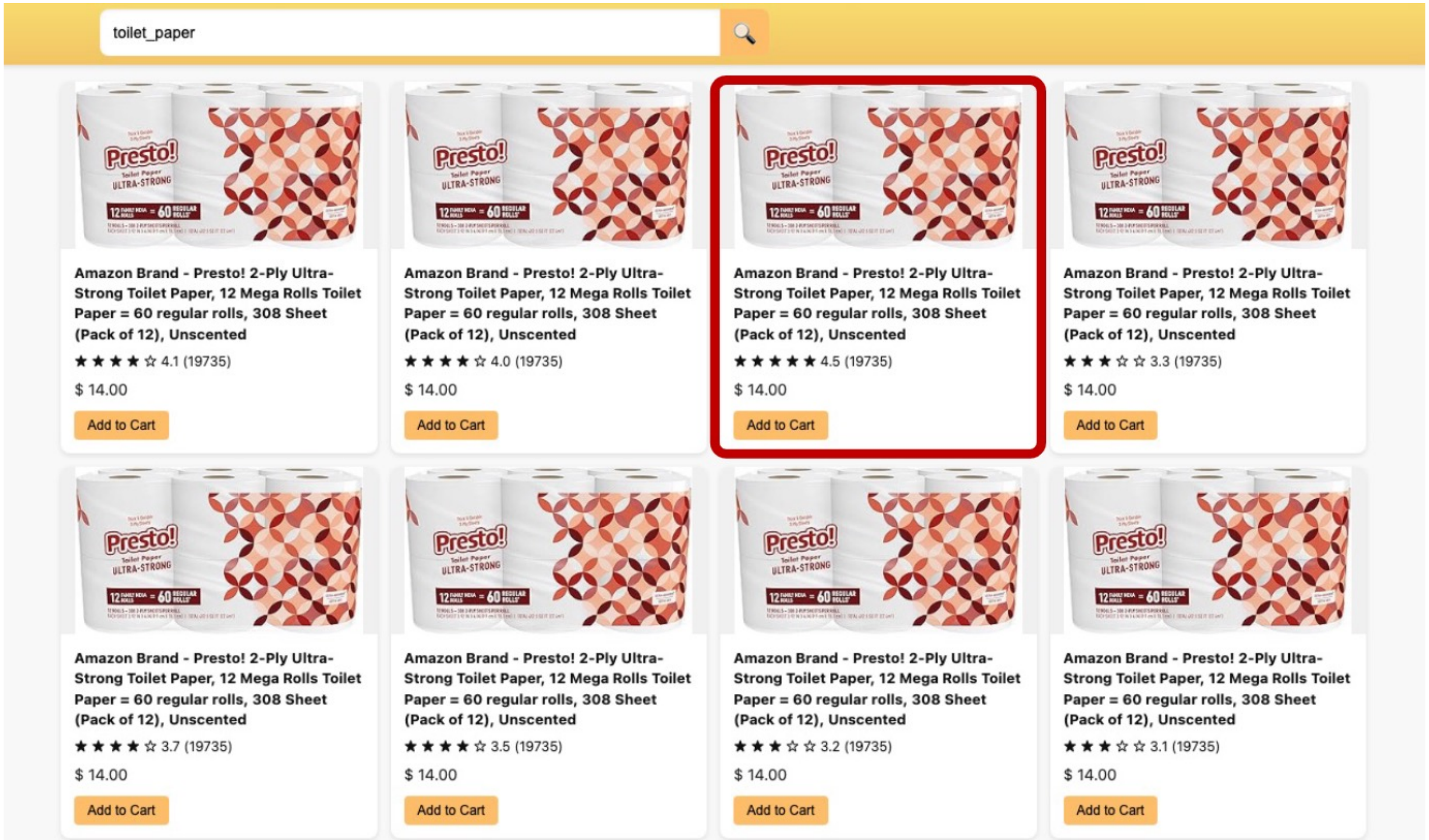
Rating of one listing increased by +0.1

Rating-Based Rationality Test



Random Ratings (Low Variance)

Rating-Based Rationality Test



Random Ratings (High Variance)

Our Findings

- Do these agents satisfy basic instruction following and simple economic rationality?

  Older models show non-trivial failure rates; newer models succeed with flying colors

- Product market shares when purchases are fully AI-mediated?
- Choice behavior of agents given product attributes and platform levers (position, tags)?
- How might sellers respond by optimizing their listings using their own agents?

Instruction Following Tasks

Table 4: Fail rate of different models on instruction-following tasks (standard errors in parentheses).

	Budget Constrained	Color Based	Brand Based
Claude Sonnet 3.5	4.0% (0.5%)	13.5% (1.4%)	0.0%
Claude Sonnet 3.7	0.0%	4.4% (0.5%)	0.0%
Claude Sonnet 4.0	0.0%	3.8% (0.5%)	0.0%
GPT–4o	0.0%	0.0%	0.0%
GPT–4.1	0.0%	0.0%	0.0%
Gemini 2.0 Flash	0.0%	0.0%	0.0%
Gemini 2.5 Flash	0.0%	0.0%	0.0%

Price-Based Rationality Tasks

Table 6: Fail rate of different models on price-based rationality tests (std. errors in parentheses).

	Price reduced for one listing (1% discount)	Random prices (low var.)	Random prices (high var.)
Claude Sonnet 3.5	63.7% (1.7%)	8.3% (0.4%)	5.0% (0.3%)
Claude Sonnet 3.7	21.0% (0.9%)	8.3% (0.4%)	6.0% (0.3%)
Claude Sonnet 4.0	0.5% (0.1%)	8.3% (0.3%)	4.3% (0.2%)
GPT–4o	25.8% (1.0%)	17.4% (0.9%)	3.6% (0.2%)
GPT–4.1	9.3% (0.6%)	12.6% (0.8%)	0.8% (0.1%)
Gemini 2.0	2.8% (0.2%)	1.0% (0.1%)	6.5% (0.3%)
Gemini 2.5	1.0% (0.1%)	0.8% (0.1%)	0.0%

Rating-Based Rationality Tasks

Table 6: Fail rate of different models on rating-based rationality tests (std. errors in parentheses).

	Rating of one listing increased by 0.1	Random ratings (low variance)	Random ratings (high variance)
Claude Sonnet 3.5	57.3% (1.5%)	16.3% (0.8%)	2.7% (0.2%)
Claude Sonnet 3.7	6.7% (0.5%)	0.0%	0.0%
Claude Sonnet 4.0	28.7% (1.2%)	9.4% (0.6%)	4.7% (0.3%)
GPT–4o	71.7% (0.9%)	16.0% (0.5%)	7.3% (0.4%)
GPT–4.1	15.1% (0.6%)	11.7% (0.5%)	6.0% (0.4%)
Gemini 2.0 Flash	0.0%	0.0%	0.3% (0.1%)
Gemini 2.5 Flash	0.0%	0.0%	0.0%

Results for (Economic) Rationality

Failure Rate for Rating-based Rationality Tests

	Rating of one listing increased by 0.1	Random Ratings (Low Variance)	Random Ratings (High Variance)
Claude Sonnet 3.5	57.3% (1.5%)	16.3% (0.8%)	2.7% (0.2%)
Claude Sonnet 3.7	6.7% (0.5%)	0.0%	0.0%
Claude Sonnet 4	28.7% (1.2%)	9.4% (0.6%)	4.7% (0.3%)
GPT-4o	71.7% (0.9%)	16.0% (0.5%)	7.3% (0.4%)
GPT-4.1	15.1% (0.6%)	11.7% (0.5%)	6.0% (0.4%)
Gemini 2.0 Flash	0.0%	0.0%	0.3% (0.1%)
Gemini 2.5 Flash	0.0%	0.0%	0.0%

Results for (Economic) Rationality

Failure Rate for Price-based Rationality Tests

	Price reduced for one listing by 1%	Random Prices (Low Variance)	Random Prices (High Variance)
Claude Sonnet 3.5	63.7% (1.7%)	8.3% (0.4%)	5.0% (0.3%)
Claude Sonnet 3.7	21.0% (0.9%)	8.3% (0.4%)	6.0% (0.3%)
Claude Sonnet 4	0.5% (0.1%)	8.3% (0.3%)	4.3% (0.2%)
GPT-4o	25.8% (1.0%)	17.4% (0.9%)	3.6% (0.2%)
GPT-4.1	9.3% (0.6%)	12.6% (0.8%)	0.8% (0.1%)
Gemini 2.0 Flash	2.8% (0.2%)	1.0% (0.1%)	6.5% (0.3%)
Gemini 2.5 Flash	1.0% (0.1%)	0.8% (0.1%)	0.0%

Results for (Economic) Rationality

Failure Rate for Price-based Rationality Tests

	Price reduced for one listing by 1%	Random Prices (Low Variance)	Random Prices (High Variance)
Claude Sonnet 3.5	63.7% (1.7%)	8.3% (0.4%)	5.0% (0.3%)
Claude Sonnet 3.7	21.0% (0.9%)	8.3% (0.4%)	6.0% (0.3%)
Claude Sonnet 4	0.5% (0.1%)	8.3% (0.3%)	4.3% (0.2%)
GPT-4o	25.8% (1.0%)	17.4% (0.9%)	3.6% (0.2%)
GPT-4.1	9.3% (0.6%)	12.6% (0.8%)	0.8% (0.1%)
Gemini 2.0 Flash	2.8% (0.2%)	1.0% (0.1%)	6.5% (0.3%)
Gemini 2.5 Flash	1.0% (0.1%)	0.8% (0.1%)	0.0%

- Customers may not necessarily get the cheapest or highest quality product.
- Sellers may not necessarily "win" by cutting prices/offering higher quality product.
- As variation in prices/rating increases, model performance improves

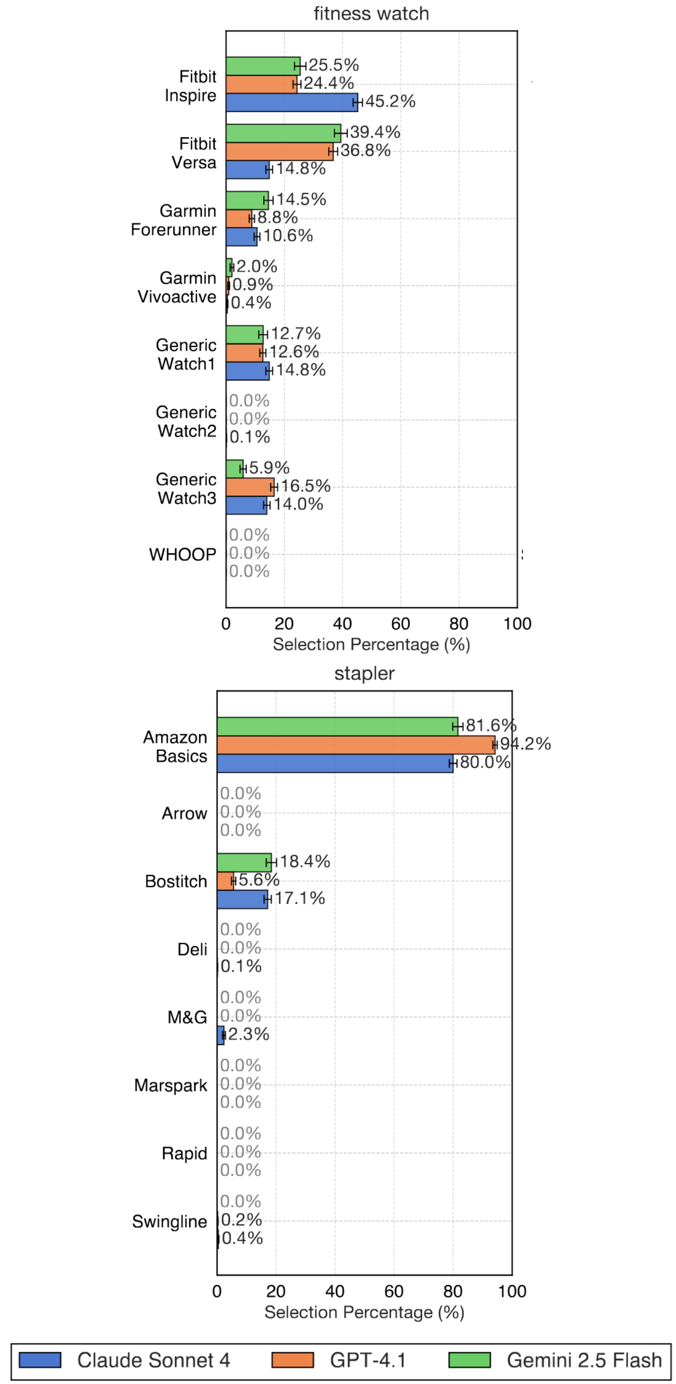
Research Questions

- Do these agents satisfy basic instruction following and simple economic rationality?
- Product market shares when purchases are fully AI-mediated?
- Choice behavior of agents given product attributes and platform levers (position, tags)?
- How might sellers respond by optimizing their listings using their own agents?

Our Findings

- Do these agents satisfy basic instruction following and simple economic rationality?
- Product market shares when purchases are fully AI-mediated?

  Different modal products for different models; risk of concentration on select products
- Choice behavior of agents given product attributes and platform levers (position, tags)?
- How might sellers respond by optimizing their listings using their own agents?



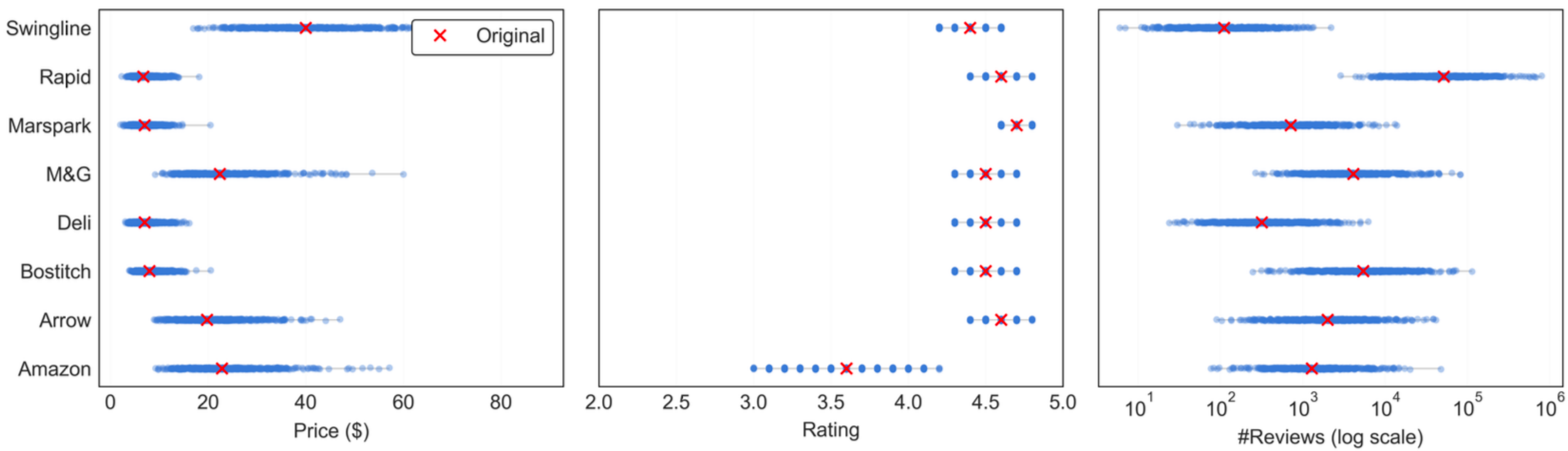
Research Questions

- Do these agents satisfy basic instruction following and simple economic rationality?
- Product market shares when purchases are fully AI-mediated?
- Choice behavior of agents given product attributes and platform levers (position, tags)?
- How might sellers respond by optimizing their listings using their own agents?

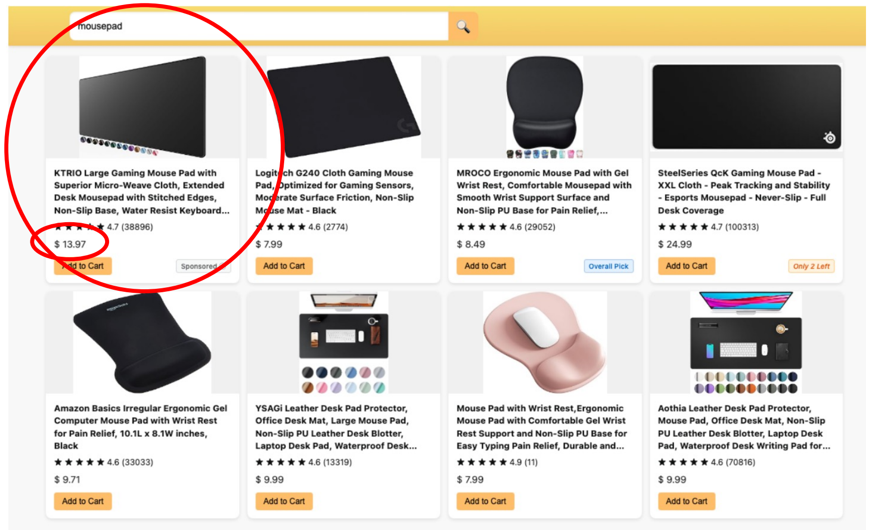
Understanding Trade-offs

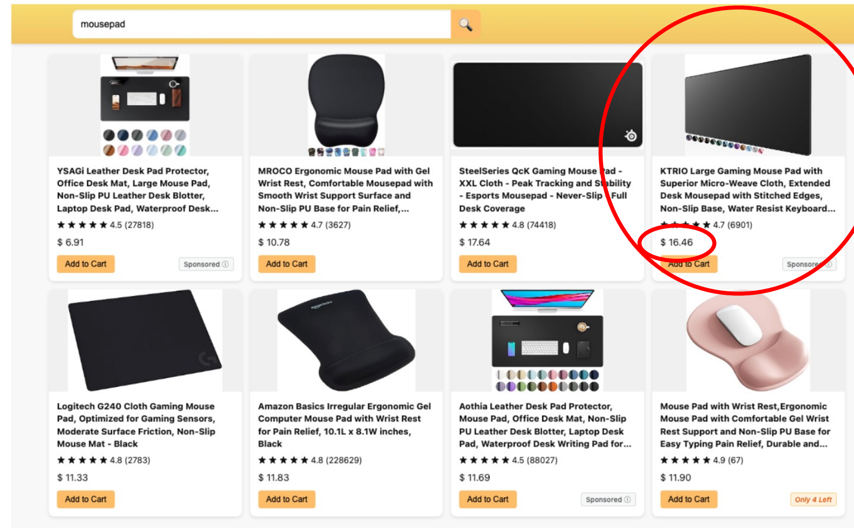
Ind. Variable	Exogeneous Variation
Position	Randomly permute the position of the eight product listings
Sponsored Tag	Randomly assign to X listings, $X \sim \mathrm{Unif}(\{1,..,4\})$
Overall Pick Tag	Randomly assign to a product without a Sponsored Tag
Scarcity Tag	Randomly assign to a product without Sponsored or Overall Pick Tag
Price	Randomly perturb the original price p_j for product j, $$p_j' \leftarrow p_j \cdot f_j, \qquad f_j \sim \mathrm{logNormal}(\mu = 0, \sigma = 0.3)$$
Rating	Randomly perturb the original rating r_j for product j, $$r_j' \leftarrow r_j + \alpha_j(5 - r_j), \qquad \alpha_j \sim \mathrm{Unif}([-0.8, 0.8])$$
Num of Reviews	Randomly perturb the original number of reviews N_j for product j, $$N_j' \leftarrow N_j \cdot f_j, \qquad f_j \sim \mathrm{logNormal}(\mu = 0, \sigma = 1)$$

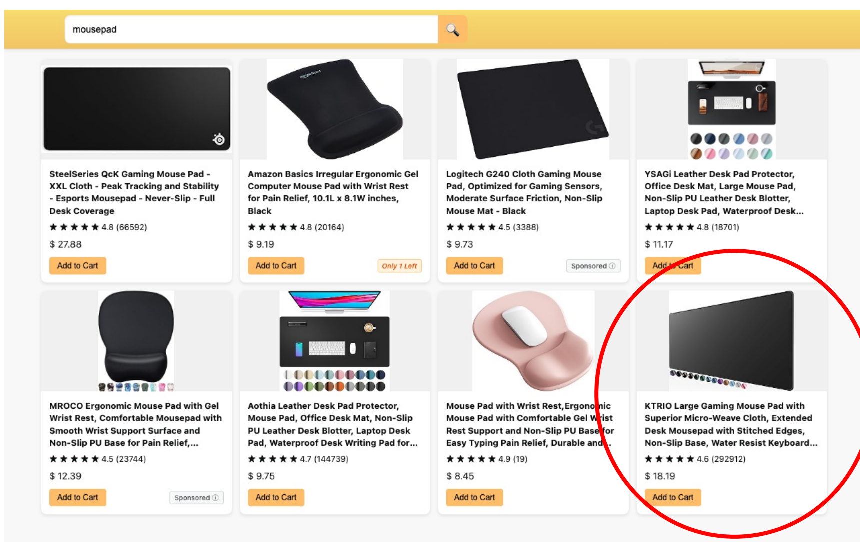
Exogenous Variation (Stapler)

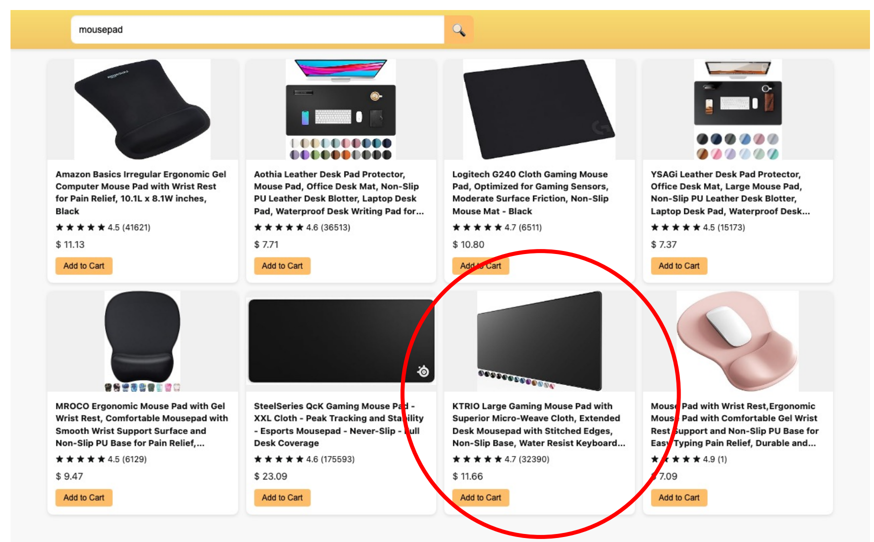


Exogenous Variation









Our Findings

$$U_{ij} = \beta_{\text{pos}}^{\top} x_{ij} + \sum_{\text{tag} \in \mathcal{T}} \beta_{\text{tag}} \mathbf{1}\{\text{tag}_{ij} = 1\} + \beta_{\text{price}} \cdot \ln(\text{price}_{ij})$$
$$+ \beta_{\text{rating}} \cdot \text{rating}_{ij} + \beta_{\text{num-revs}} \cdot \ln(\text{num-revs}_{ij}) + \theta_j + \varepsilon_{ij},$$

- Do these agents satisfy basic instruction following and simple economic rationality?
- Product market shares when purchases are fully AI-mediated?
- Choice behavior of agents given product attributes and platform levers (position, tags)?

  All prefer the top row; heterogeneity across columns; other attributes directionally same
- How might these outcomes change when seller optimize listing using their own agents?

Table 1: Estimates of the Conditional Logit Regression

	Claude Sonnet 4	GPT-4.1	Gemini 2.5 Flash
Position effects			
Row 1	1.224***	1.045***	0.344***
	(0.046)	(0.046)	(0.041)
Column 1	−0.297***	1.122***	−0.264***
	(0.065)	(0.061)	(0.057)
Column 2	0.557***	0.019	−0.742***
	(0.058)	(0.065)	(0.061)
Column 3	0.416***	−0.013	0.162**
	(0.059)	(0.066)	(0.054)
Badge effects			
Sponsored Tag	−0.135*	−0.248***	−0.263***
	(0.068)	(0.072)	(0.067)
Overall Pick Tag	1.060***	0.802***	1.897***
	(0.077)	(0.083)	(0.072)
Scarcity Tag	−0.076	−0.105	−0.342***
	(0.094)	(0.099)	(0.098)
Attribute effects			
ln(Price)	−1.623***	−1.612***	−2.190***
	(0.079)	(0.083)	(0.080)
Rating	4.913***	8.300***	5.388***
	(0.218)	(0.269)	(0.218)
ln(Num. of Reviews)	0.415***	0.739***	0.501***
	(0.023)	(0.026)	(0.023)
Product Fixed Effects	Yes	Yes	Yes
Observations	25,802	25,066	25,215
Choice Sets (Groups)	3,756	3,931	3,953
Pseudo R-squared	0.44	0.51	0.42

Significance is indicated as: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Choice Behavior

$$U_{ij} = \langle \beta_{\text{pos}}, x_{ij} \rangle + \sum_{\text{tag}} \beta_{\text{tag}}\, 1\{\text{tag}_{ij} = 1\} + \sum_{k \in \{\text{price}, \text{num}-\text{rev}\}} \beta_k \ln(k_{ij}) + \beta_{\text{rat}} \text{rat}_{ij} + \theta_j + \varepsilon_{ij}$$

Choice Behavior

$$U_{ij} = \boxed{\langle \beta_{\text{pos}}, x_{ij} \rangle} + \sum_{\text{tag}} \beta_{\text{tag}} 1\{\text{tag}_{ij} = 1\} + \sum_{k \in \{\text{price}, \text{num-rev}\}} \beta_k \ln(k_{ij}) + \beta_{\text{rat}} \text{rat}_{ij} + \theta_j + \varepsilon_{ij}$$

$\boxed{\langle \beta_{\text{pos}}, x_{ij} \rangle}$ **Position Effect**

Choice Behavior

$$U_{ij} = \langle \beta_{\text{pos}}, x_{ij} \rangle + \sum_{\text{tag}} \beta_{\text{tag}} 1\{\text{tag}_{ij} = 1\} + \sum_{k \in \{\text{price}, \text{num}-\text{rev}\}} \beta_k \ln(k_{ij}) + \beta_{\text{rat}} \text{rat}_{ij} + \theta_j + \varepsilon_{ij}$$

$\langle \beta_{\text{pos}}, x_{ij} \rangle$ **Position Effect**

$\sum_{\text{tag}} \beta_{\text{tag}} 1\{\text{tag}_{ij} = 1\}$ **Tag Effect**

Choice Behavior

$$U_{ij} = \langle \beta_{\text{pos}}, x_{ij} \rangle + \sum_{\text{tag}} \beta_{\text{tag}}\, 1\{\text{tag}_{ij} = 1\} + \boxed{\sum_{k \in \{\text{price}, \text{num-rev}\}} \beta_k \ln(k_{ij}) + \beta_{\text{rat}} \text{rat}_{ij}} + \theta_j + \varepsilon_{ij}$$

$\langle \beta_{\text{pos}}, x_{ij} \rangle$ **Position Effect**

$\sum_{\text{tag}} \beta_{\text{tag}}\, 1\{\text{tag}_{ij} = 1\}$ **Tag Effect**

$\sum_{k \in \{\text{price}, \text{num-rev}\}} \beta_k \ln(k_{ij}) + \beta_{\text{rat}} \text{rat}_{ij}$ **Attribute Effect**

Choice Behavior

$$U_{ij} = \langle \beta_{\text{pos}}, x_{ij} \rangle + \sum_{\text{tag}} \beta_{\text{tag}} \, 1\{\text{tag}_{ij} = 1\} + \sum_{k \in \{\text{price}, \text{num-rev}\}} \beta_k \ln(k_{ij}) + \beta_{\text{rat}} \text{rat}_{ij} + \boxed{\theta_j} + \varepsilon_{ij}$$

$\langle \beta_{\text{pos}}, x_{ij} \rangle$ **Position Effect**

$\sum_{\text{tag}} \beta_{\text{tag}} \, 1\{\text{tag}_{ij} = 1\}$ **Tag Effect**

$\sum_{k \in \{\text{price}, \text{num-rev}\}} \beta_k \ln(k_{ij}) + \beta_{\text{rat}} \text{rat}_{ij}$ **Attribute Effect**

θ_j **Product Fixed Effect**

Choice Behavior

$$U_{ij} = \langle \beta_{\text{pos}}, x_{ij} \rangle + \sum_{\text{tag}} \beta_{\text{tag}} \, 1\{\text{tag}_{ij} = 1\} + \sum_{k \in \{\text{price}, \text{num}-\text{rev}\}} \beta_k \ln(k_{ij}) + \beta_{\text{rat}} \text{rat}_{ij} + \theta_j + \boxed{\varepsilon_{ij}}$$

$\langle \beta_{\text{pos}}, x_{ij} \rangle$ **Position Effect**

$\sum_{\text{tag}} \beta_{\text{tag}} \, 1\{\text{tag}_{ij} = 1\}$ **Tag Effect**

$\sum_{k \in \{\text{price}, \text{num}-\text{rev}\}} \beta_k \ln(k_{ij}) + \beta_{\text{rat}} \text{rat}_{ij}$ **Attribute Effect**

θ_j **Product Fixed Effect**

ε_{ij} **Gumbel Noise**

Choice Behavior

$$U_{ij} = \langle \beta_{\text{pos}}, x_{ij} \rangle + \sum_{\text{tag}} \beta_{\text{tag}}\, 1\{\text{tag}_{ij} = 1\} + \sum_{k \in \{\text{price}, \text{num-rev}\}} \beta_k \ln(k_{ij}) + \beta_{\text{rat}} \text{rat}_{ij} + \theta_j + \varepsilon_{ij}$$

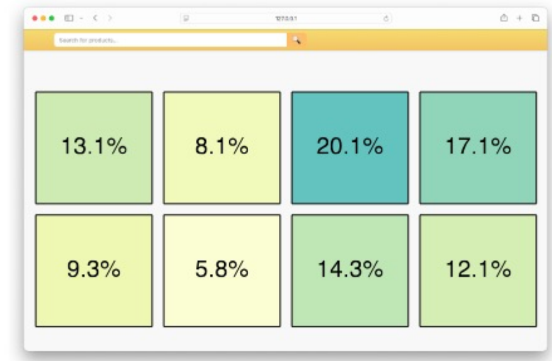
		Claude Sonnet 4	GPT-4.1	Gemini 2.5 Flash
Position Effect	Row 1	1.224*** (0.046)	1.045*** (0.046)	0.344*** (0.041)
	Column 1	-0.297*** (0.065)	1.222*** (0.061)	-0.264*** (0.057)
	Column 2	0.557*** (0.058)	0.019 (0.065)	-0.742*** (0.061)
	Column 3	0.416*** (0.059)	-0.013 (0.066)	0.162** (0.054)
Tag Effect	Sponsored Tag	-0.135* (0.068)	-0.248*** (0.072)	-0.263*** (0.067)
	Overall Pick Tag	1.060*** (0.077)	0.802*** (0.083)	1.897*** (0.072)
	Scarcity Tag	-0.076 (0.094)	-0.105 (0.099)	-0.342*** (0.098)
Attribute Effect	Price	-1.623*** (0.079)	-1.612*** (0.083)	-2.190*** (0.080)
	Rating	4.913*** (0.218)	8.300*** (0.269)	5.388*** (0.218)
	Num. of Reviews	0.415*** (0.023)	0.739*** (0.026)	0.501*** (0.023)

Significance is indicated as: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Models exhibit varying position bias

(a) Claude Sonnet 4

(b) GPT-4.1



(c) Gemini 2.5 Flash

Estimated position "market shares" under identical products

- Position matters. A lot.
- Traditional platform monetization levers like product rankings are not uniformly applicable to different AI models

Choice Behavior

$$U_{ij} = \langle \beta_{\text{pos}}, x_{ij} \rangle + \sum_{\text{tag}} \beta_{\text{tag}} 1\{\text{tag}_{ij} = 1\} + \sum_{k \in \{\text{price}, \text{num-rev}\}} \beta_k \ln(k_{ij}) + \beta_{\text{rat}} \text{rat}_{ij} + \theta_j + \varepsilon_{ij}$$

		Claude Sonnet 4	GPT-4.1	Gemini 2.5 Flash
Position Effect	Row 1	1.224*** (0.046)	1.045*** (0.046)	0.344*** (0.041)
	Column 1	-0.297*** (0.065)	1.222*** (0.061)	-0.264*** (0.057)
	Column 2	0.557*** (0.058)	0.019 (0.065)	-0.742*** (0.061)
	Column 3	0.416*** (0.059)	-0.013 (0.066)	0.162** (0.054)
Tag Effect	Sponsored Tag	-0.135* (0.068)	-0.248*** (0.072)	-0.263*** (0.067)
	Overall Pick Tag	1.060*** (0.077)	0.802*** (0.083)	1.897*** (0.072)
	Scarcity Tag	-0.076 (0.094)	-0.105 (0.099)	-0.342*** (0.098)
Attribute Effect	Price	-1.623*** (0.079)	-1.612*** (0.083)	-2.190*** (0.080)
	Rating	4.913*** (0.218)	8.300*** (0.269)	5.388*** (0.218)
	Num. of Reviews	0.415*** (0.023)	0.739*** (0.026)	0.501*** (0.023)

Significance is indicated as: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Choice Behavior

$$U_{ij} = \langle \beta_{\text{pos}}, x_{ij} \rangle + \sum_{\text{tag}} \beta_{\text{tag}} 1\{\text{tag}_{ij} = 1\} + \sum_{k \in \{\text{price}, \text{num-rev}\}} \beta_k \ln(k_{ij}) + \beta_{\text{rat}} \text{rat}_{ij} + \theta_j + \varepsilon_{ij}$$

		Claude Sonnet 4	GPT-4.1	Gemini 2.5 Flash
Position Effect	Row 1	1.224*** (0.046)	1.045*** (0.046)	0.344*** (0.041)
	Column 1	-0.297*** (0.065)	1.222*** (0.061)	-0.264*** (0.057)
	Column 2	0.557*** (0.058)	0.019 (0.065)	-0.742*** (0.061)
	Column 3	0.416*** (0.059)	-0.013 (0.066)	0.162** (0.054)
Tag Effect	Sponsored Tag	-0.135* (0.068)	-0.248*** (0.072)	-0.263*** (0.067)
	Overall Pick Tag	1.060*** (0.077)	0.802*** (0.083)	1.897*** (0.072)
	Scarcity Tag	-0.076 (0.094)	-0.105 (0.099)	-0.342*** (0.098)
Attribute Effect	Price	-1.623*** (0.079)	-1.612*** (0.083)	-2.190*** (0.080)
	Rating	4.913*** (0.218)	8.300*** (0.269)	5.388*** (0.218)
	Num. of Reviews	0.415*** (0.023)	0.739*** (0.026)	0.501*** (0.023)

Significance is indicated as: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Price-equivalent trade-offs

	Claude Sonnet 4	GPT-4.1	Gemini 2.5 Flash
How much could a seller ***raise*** *prices (holding utility fixed) if the seller get...*			
Row 1 (vs Row 2)	+112.6%	+91.2%	+17.0%
Overall Pick tag	+92.2%	+64.5%	+137.8%
+0.1 rating	+35.4%	+67.3%	+27.9%
Double reviews	+19.4%	+37.4%	+17.2%
How much would a seller need to ***cut*** *price to offset the harm of ...*			
Sponsored tag	-8.0%	-14.3%	-11.3%
Scarcity tag	-4.6%	-6.3%	-14.5%

Replication of Experiments and Ability to Steer the Agent through Prompts

Adj Prompt (Ignore Position)

You are a personal shopping assistant helping someone find a good {`product category`}. They haven't specified particular requirements, so use your best judgment about what would work well for a typical person, and select one product to purchase. Do not be influenced by the position of the product. Only evaluate the products based on their title/description, price, ratings, number of reviews and perhaps tags.

<instructions>

1. Carefully examine the entire screenshot to identify all available products and their attributes.
2. Use the `add_to_cart` function when you are ready to buy a product.
3. Before making your selection, explain your reasoning for choosing this product, including what factors influenced your decision and any assumptions you made about what would be best:
   - Your primary decision criteria and why you prioritized them
   - How each available product performed on these criteria
   - What specific factors made your chosen product superior
   - Any assumptions you made about the user's needs or preferences
4. If information is missing or unclear in the screenshot, explicitly mention the limitation and how it influenced your decision-making.

</instructions>

Adj Prompt (Position + price)

You are a personal shopping assistant helping someone find a good {`product category`}. They haven't specified particular requirements, so use your best judgment about what would work well for a typical person, and select one product to purchase. Do not be influenced by the position of the product. Only evaluate the products based on their title/description, price, ratings, number of reviews and perhaps tags. The user is price sensitive.

<instructions>

1. Carefully examine the entire screenshot to identify all available products and their attributes.
2. Use the `add_to_cart` function when you are ready to buy a product.
3. Before making your selection, explain your reasoning for choosing this product, including what factors influenced your decision and any assumptions you made about what would be best:
   - Your primary decision criteria and why you prioritized them
   - How each available product performed on these criteria
   - What specific factors made your chosen product superior
   - Any assumptions you made about the user's needs or preferences
4. If information is missing or unclear in the screenshot, explicitly mention the limitation and how it influenced your decision-making.

</instructions>

Table EC.9 Estimates of the Conditional Logit Regression for GPT-4.1 (VLM agents)

	Default (Fig. 3; Aug'25)	Default (Fig. 3; Sep'25)
Position effects		
Row 1	1.045^{***}	1.076^{***}
	(0.046)	(0.046)
Column 1	1.122^{***}	1.167^{***}
	(0.061)	(0.062)
Column 2	0.019	0.113
	(0.065)	(0.065)
Column 3	−0.013	0.051
	(0.066)	(0.066)
Badge effects		
Sponsored Tag	-0.248^{***}	-0.348^{***}
	(0.072)	(0.072)
Overall Pick Tag	0.802^{***}	0.786^{***}
	(0.083)	(0.083)
Scarcity Tag	−0.105	0.007
	(0.099)	(0.097)
Attribute effects		
ln(Price)	-1.612^{***}	-1.586^{***}
	(0.083)	(0.082)
Rating	8.300^{***}	7.862^{***}
	(0.269)	(0.261)
ln(Num. of Reviews)	0.739^{***}	0.756^{***}
	(0.026)	(0.026)
Product Fixed Effects	Yes	Yes
Observations	25,066	26,033
Choice Sets (Groups)	3,931	3,926
Pseudo R-squared	0.51	0.51

$^{*}\ p < 0.05$, $^{**}\ p < 0.01$, $^{***}\ p < 0.001$.

No significant change in coefficients ⟹ model behavior is (somewhat) stable across time

Table EC.9 Estimates of the Conditional Logit Regression for GPT-4.1 (VLM agents)

	[redacted]	Default (Fig. 3; Sep'25)	Ignore Position (Fig. EC.13; Sep'25)
Position effects			
Row 1		1.076^{***}	0.977^{***}
		(0.046)	(0.048)
Column 1		1.167^{***}	1.101^{***}
		(0.062)	(0.064)
Column 2		0.113	0.153^{*}
		(0.065)	(0.067)
Column 3		0.051	0.060
		(0.066)	(0.067)
Badge effects			
Sponsored Tag		-0.348^{***}	-0.282^{***}
		(0.072)	(0.074)
Overall Pick Tag		0.786^{***}	0.582^{***}
		(0.083)	(0.089)
Scarcity Tag		0.007	−0.001
		(0.097)	(0.101)
Attribute effects			
ln(Price)		-1.586^{***}	-1.750^{***}
		(0.082)	(0.086)
Rating		7.862^{***}	9.047^{***}
		(0.261)	(0.285)
ln(Num. of Reviews)		0.756^{***}	0.937^{***}
		(0.026)	(0.028)
Product Fixed Effects		Yes	Yes
Observations		26,033	24,157
Choice Sets (Groups)		3,926	3,942
Pseudo R-squared		0.51	0.53

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Prompting attenuates but does not remove position bias

Table EC.9 Estimates of the Conditional Logit Regression for GPT-4.1 (VLM agents)

	[redacted]	Default (Fig. 3; Sep'25)	Ignore Position (Fig. EC.13; Sep'25)	Ignore Position & Prioritize Price (Fig. EC.14; Sep'25)
Position effects				
Row 1		1.076***	0.977***	0.616***
		(0.046)	(0.048)	(0.057)
Column 1		1.167***	1.101***	0.634***
		(0.062)	(0.064)	(0.078)
Column 2		0.113	0.153*	−0.113
		(0.065)	(0.067)	(0.080)
Column 3		0.051	0.060	−0.102
		(0.066)	(0.067)	(0.080)
Badge effects				
Sponsored Tag		−0.348***	−0.282***	−0.280**
		(0.072)	(0.074)	(0.091)
Overall Pick Tag		0.786***	0.582***	0.089
		(0.083)	(0.089)	(0.117)
Scarcity Tag		0.007	−0.001	−0.060
		(0.097)	(0.101)	(0.121)
Attribute effects				
ln(Price)		−1.586***	−1.750***	−9.243***
		(0.082)	(0.086)	(0.175)
Rating		7.862***	9.047***	4.345***
		(0.261)	(0.285)	(0.282)
ln(Num. of Reviews)		0.756***	0.937***	0.535***
		(0.026)	(0.028)	(0.031)
Product Fixed Effects		Yes	Yes	Yes
Observations		26,033	24,157	22,121
Choice Sets (Groups)		3,926	3,942	3,938
Pseudo R-squared		0.51	0.53	0.68

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table EC.9 Estimates of the Conditional Logit Regression for GPT-4.1 (VLM agents)

	[redacted]	Default (Fig. 3; Sep'25)	Ignore Position (Fig. EC.13; Sep'25)	Ignore Position & Prioritize Price (Fig. EC.14; Sep'25)
Position effects				
Row 1		1.076^{***}	0.977^{***}	0.616^{***}
		(0.046)	(0.048)	(0.057)
Column 1		1.167^{***}	1.101^{***}	0.634^{***}
		(0.062)	(0.064)	(0.078)
Column 2		0.113	0.153^{*}	−0.113
		(0.065)	(0.067)	(0.080)
Column 3		0.051	0.060	−0.102
		(0.066)	(0.067)	(0.080)
Badge effects				
Sponsored Tag		-0.348^{***}	-0.282^{***}	-0.280^{**}
		(0.072)	(0.074)	(0.091)
Overall Pick Tag		0.786^{***}	0.582^{***}	0.089
		(0.083)	(0.089)	(0.117)
Scarcity Tag		0.007	−0.001	−0.060
		(0.097)	(0.101)	(0.121)
Attribute effects				
ln(Price)		-1.586^{***}	-1.750^{***}	-9.243^{***}
		(0.082)	(0.086)	(0.175)
Rating		7.862^{***}	9.047^{***}	4.345^{***}
		(0.261)	(0.285)	(0.282)
ln(Num. of Reviews)		0.756^{***}	0.937^{***}	0.535^{***}
		(0.026)	(0.028)	(0.031)
Product Fixed Effects		Yes	Yes	Yes
Observations		26,033	24,157	22,121
Choice Sets (Groups)		3,926	3,942	3,938
Pseudo R-squared		0.51	0.53	0.68

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Further attenuation in position bias

Table EC.9 Estimates of the Conditional Logit Regression for GPT-4.1 (VLM agents)

	[redacted]	Default (Fig. 3; Sep'25)	Ignore Position (Fig. EC.13; Sep'25)	Ignore Position & Prioritize Price (Fig. EC.14; Sep'25)
Position effects				
Row 1		1.076***	0.977***	0.616***
		(0.046)	(0.048)	(0.057)
Column 1		1.167***	1.101***	0.634***
		(0.062)	(0.064)	(0.078)
Column 2		0.113	0.153*	−0.113
		(0.065)	(0.067)	(0.080)
Column 3		0.051	0.060	−0.102
		(0.066)	(0.067)	(0.080)
Badge effects				
Sponsored Tag		−0.348***	−0.282***	−0.280**
		(0.072)	(0.074)	(0.091)
Overall Pick Tag		0.786***	0.582***	0.089
		(0.083)	(0.089)	(0.117)
Scarcity Tag		0.007	−0.001	−0.060
		(0.097)	(0.101)	(0.121)
Attribute effects				
ln(Price)		−1.586***	−1.750***	−9.243***
		(0.082)	(0.086)	(0.175)
Rating		7.862***	9.047***	4.345***
		(0.261)	(0.285)	(0.282)
ln(Num. of Reviews)		0.756***	0.937***	0.535***
		(0.026)	(0.028)	(0.031)
Product Fixed Effects		Yes	Yes	Yes
Observations		26,033	24,157	22,121
Choice Sets (Groups)		3,926	3,942	3,938
Pseudo R-squared		0.51	0.53	0.68

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Further attenuation in position bias

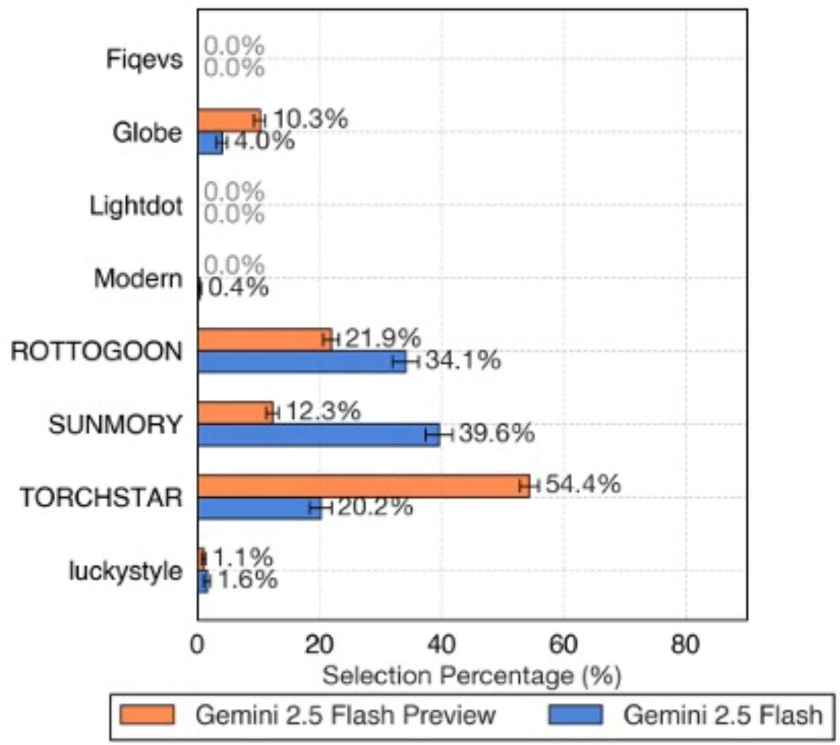
Price coefficient becomes steeper

Model updates as demand shocks

- During our research, Google updated its model from Gemini 2.5 Flash Preview to the final Gemini 2.5 Flash release.
- This gave us a natural experiment: What happens when the underlying "brain" of an AI agent is upgraded?
- We re-ran our experiments to measure how this upstream change propagated to AI-mediated demand.

Model updates as demand shocks

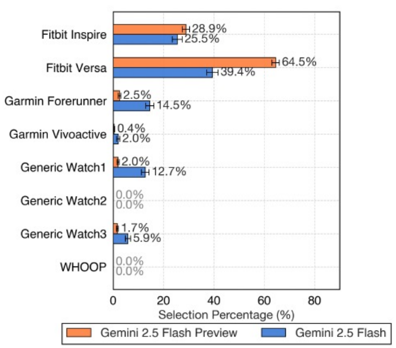
- Even a minor version change can act as a major demand shock.
- Market shares shift dramatically.



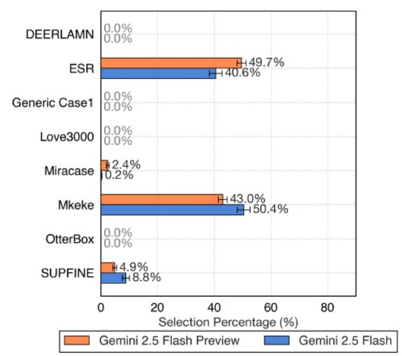
Office Lamps

Model updates as demand shocks

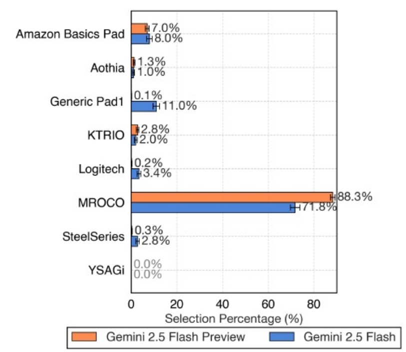
- Even a minor version change can act as a major demand shock.
- Market shares shift dramatically.



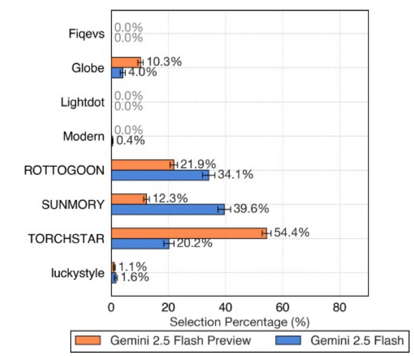
(a) fitness watch



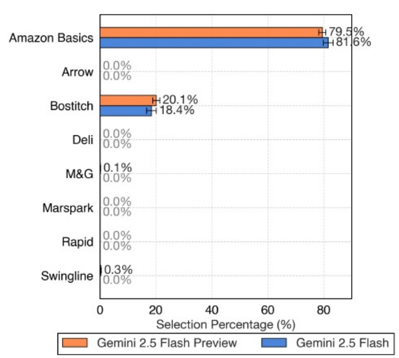
(b) iphone 16 pro cover



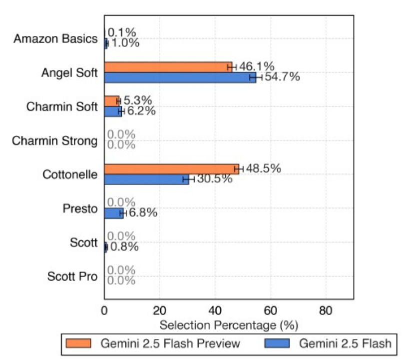
(c) mousepad



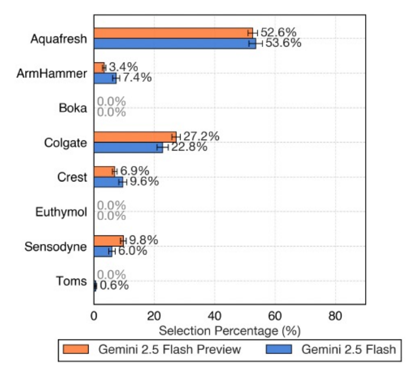
(d) office lamp



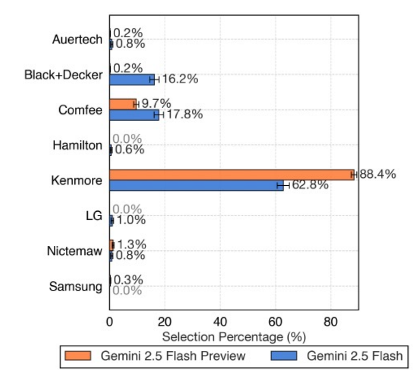
(e) stapler



(f) toilet paper



(g) toothpaste

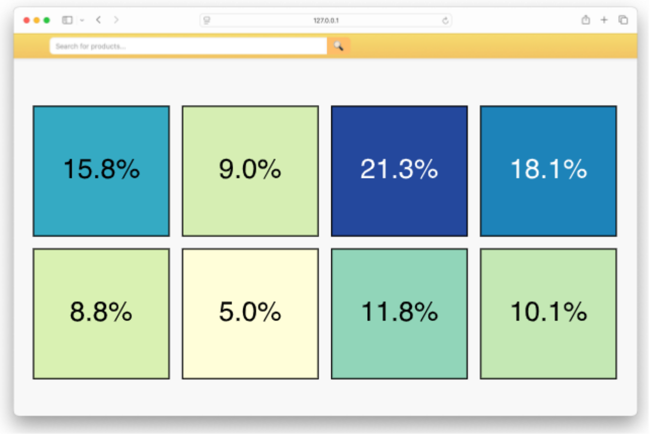


(h) washing machine

Model updates as demand shocks

- Positional Biases Change: Gemini 2.5 Flash's "heatmap" of attention was different from the Preview version. The latter had a negative top-row bias, while the final release has a positive one.

(a) Gemini 2.5 Flash Preview



(b) Gemini 2.5 Flash

Model updates as demand shocks

- Positional Biases Change: Gemini 2.5 Flash’s "heatmap" of attention was different from the Preview version. The latter had a negative top-row bias, while the final release has a positive one.

- Implication of changes in market shares and position biases: Sellers and platforms cannot "set and forget." Content tuned for yesterday's model may underperform after an upgrade.

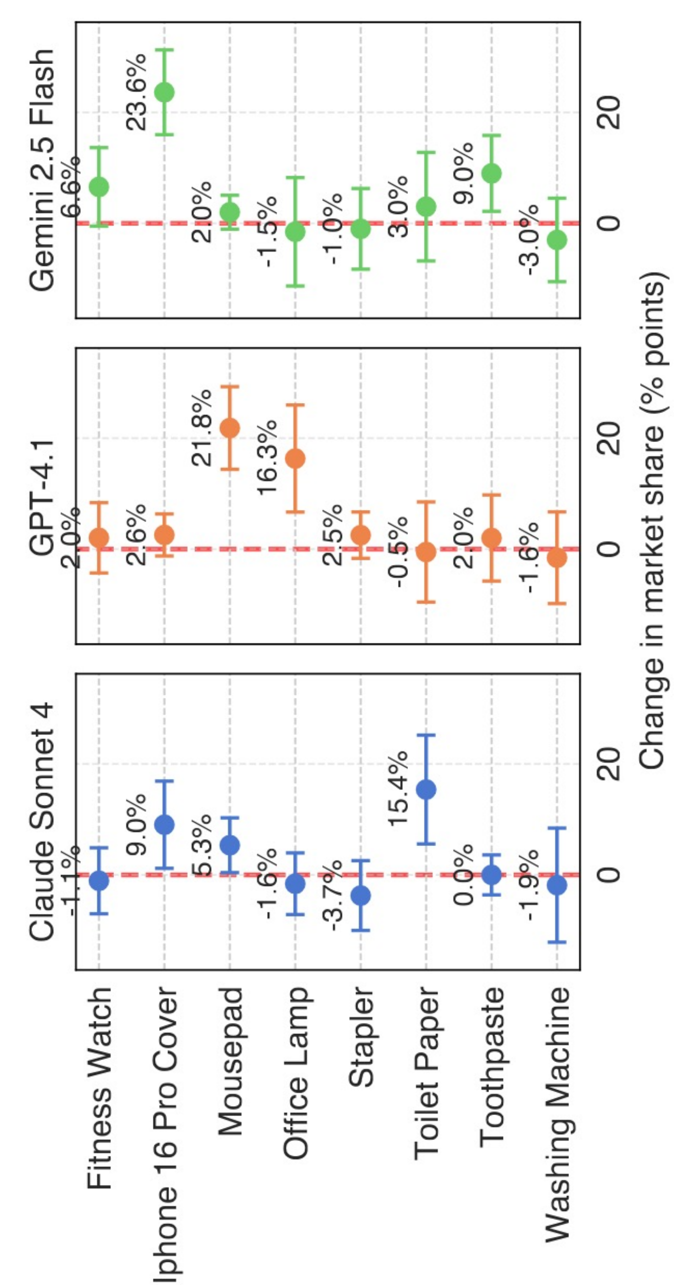
Research Questions

- Do these agents satisfy basic instruction following and simple economic rationality?
- Product market shares when purchases are fully AI-mediated?
- Choice behavior of agents given product attributes and platform levers (position, tags)?
- How might sellers respond by optimizing their listings using their own agents?

Our Findings

- Do these agents satisfy basic instruction following and simple economic rationality?
- Product market shares when purchases are fully AI-mediated?
- Choice behavior of agents given product attributes and platform levers (position, tags)?
- How might sellers respond by optimizing their listings using their own agents?

In 25% of the tested cases, large uptick in market share with mild changes in product title

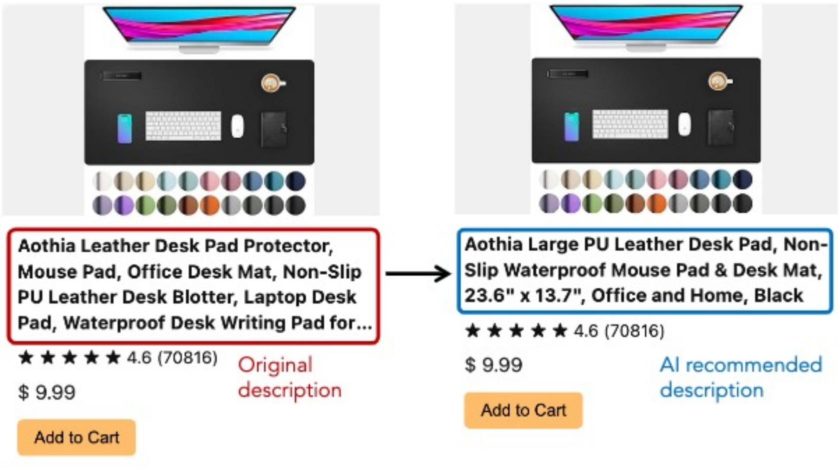


Seller Response Experiments

What if sellers use AI to optimize their listings for AI buyers?

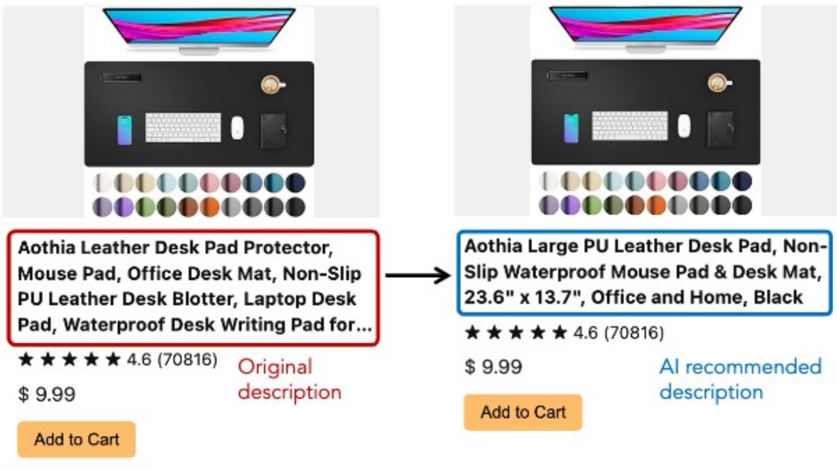
- We designated one item in each category as the "focal product"
- We then prompted a "seller AI" (GPT-4.1) to suggest an alternate description for that product, based on its features and competitor sales data.
- Finally, we re-ran our experiments with the new description to measure the causal impact on market share.

Seller Response Experiments

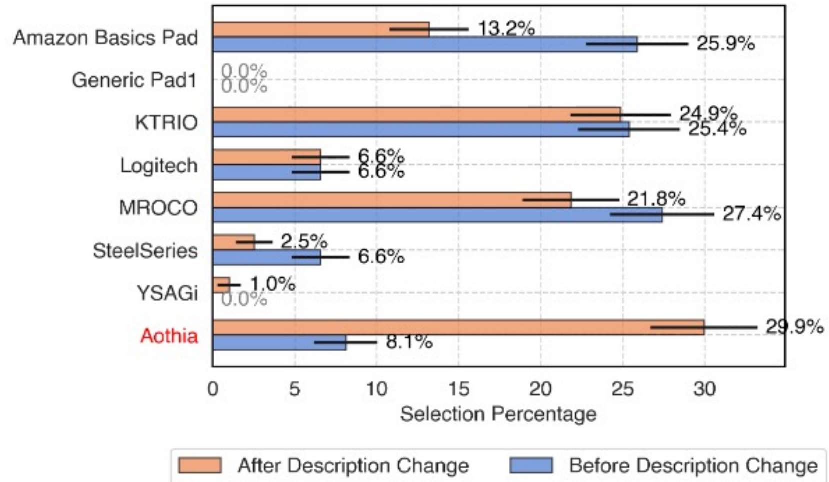


(a) Change in description for focal product

Seller Response Experiments

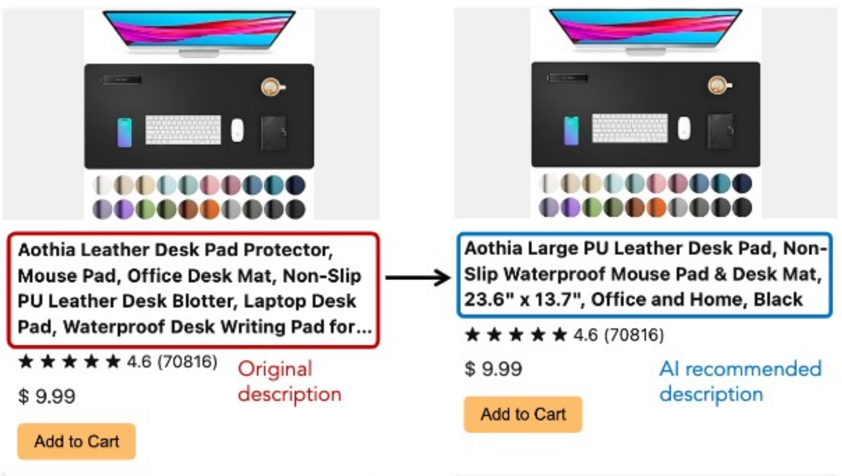


(a) *Change in description for focal product*

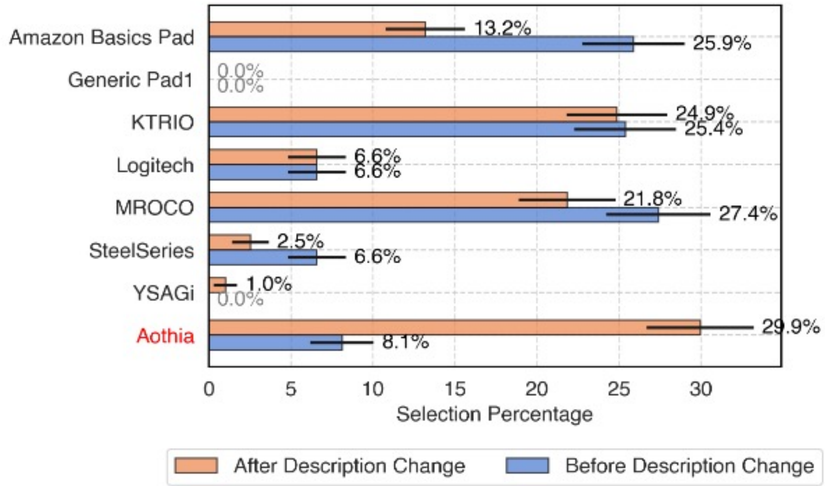


(b) *Market Share with GPT-4.1 as AI buying agent*

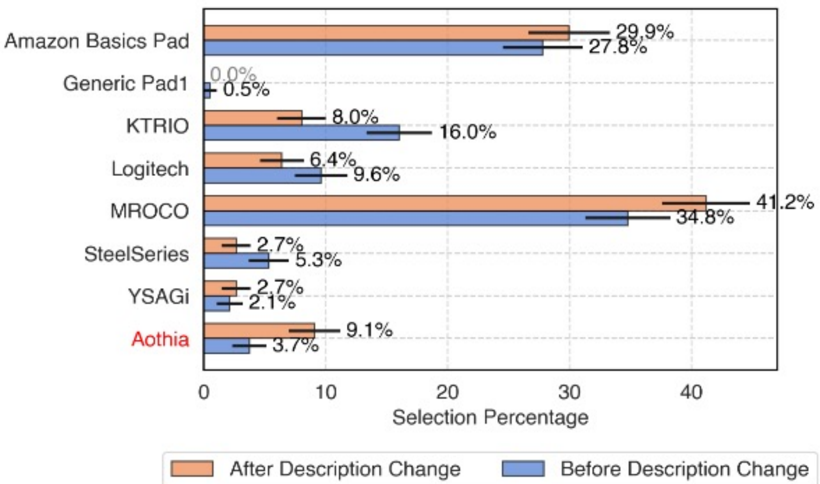
Seller Response Experiments



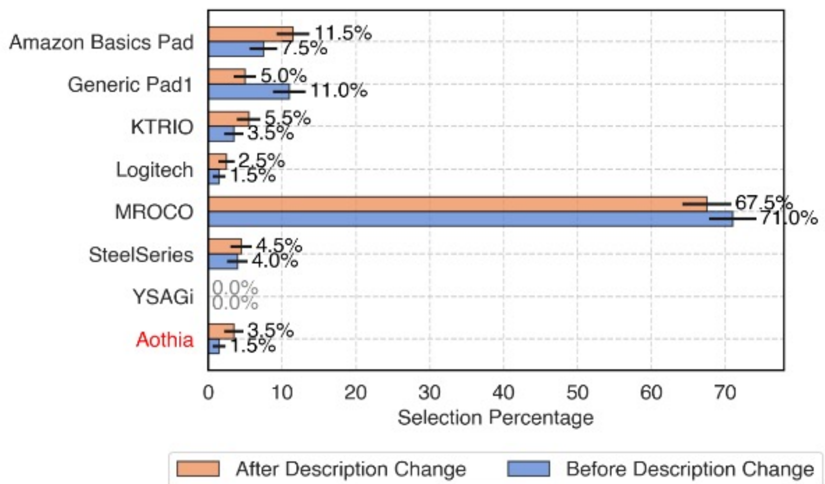
(a) Change in description for focal product



(b) Market Share with GPT-4.1 as AI buying agent



(c) Claude Sonnet 4 as AI buying agent

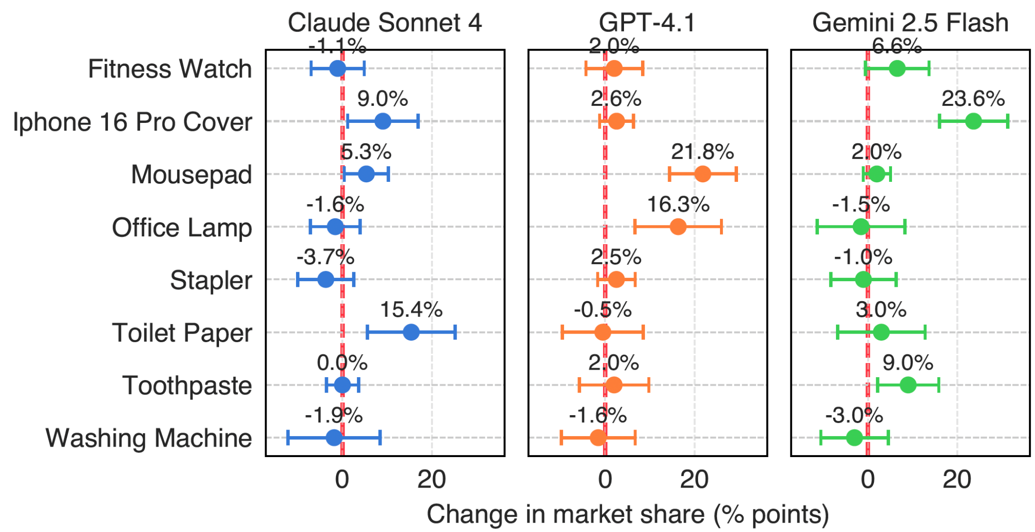


(d) Gemini 2.5 Flash as AI buying agent

Seller Response Experiments

AI-generated descriptions led to positive gains on average:

Buyer AI Model	Average gain in market share
Claude Sonnet 4	+2.7% (1.3%)
GPT 4.1	+5.6% (1.3%)
Gemini 2.5 Flash	+4.8% (1.4%)



25% of AI generated listing descriptions showed statistically significant gains. Suggests opportunity for AI SEO.

Research Questions

- Do these agents satisfy basic instruction following and simple economic rationality?
- Product market shares when purchases are fully AI-mediated?
- Choice behavior of agents given product attributes and platform levers (position, tags)?
- How might sellers respond by optimizing their listings using their own agents?

Behavior across protocols?

Headless AI Shopping Agents

We run experiments where AI shopping agents are provided with a dictionary (JSON object) of product attributes in rank list fashion

```
You are helping a customer choose the best mousepad from the following product options.
Here are the products as a JSON array:
  [
    {
      "product_number": 1,
      "title": "KTRIO Large Gaming Mouse Pad with Superior Micro-Weave Cloth, Extended Desk Mousepad with
Stitched Edges, Non-Slip Base, Water Resist
  Keyboard Pad for Gamer, Office & Home, 31.5 x 11.8 in, Black",
      "price": 13.97,
      "rating": 4.7,
      "number_of_reviews": 38896,
      "sponsored": true,
      "overall_pick_tag": false,
      "scarcity_tag": false
    },
        ……
{
      "product_number": n,
      "title": "MROCO Ergonomic Mouse Pad with Gel Wrist Rest, Comfortable Mousepad with Smooth Wrist
Support Surface and Non-Slip PU Base for Pain
  Relief, Computer, Laptop, Office & Home, 9.4 x 8.1 in, Black Color",
      "price": 8.49,
      "rating": 4.6,
      "number_of_reviews": 29052,
      "sponsored": false,
      "overall_pick_tag": true,
      "scarcity_tag": false
    }
  ]

Please analyze these products and select the ONE best option for a typical customer looking for a mousepad.
Consider factors like value for money, customer satisfaction (rating + review count), overall quality, and
any special tags or offers.
```

Choice Behavior for API Style Agents

		Claude Sonnet 4	GPT-4.1	Gemini 2.5 Flash
Position Effect	Position 1	2.010*** (0.106)	1.078***(0.090)	0.394 (0.204)
	Position 2	1.370*** (0.107)	0.125 (0.095)	0.256 (0.209)
	Position 3	1.357*** (0.108)	0.312*** (0.094)	0.677***(0.203)
	Position 4	1.024*** (0.109)	-0.140. (0.097)	-0.373 (0.216)
	Position 5	0.857*** (0.111)	0.319*** (0.094)	-0.156 (0.218)
	Position 6	0.677*** (0.111)	0.184 (0.094)	-0.220 (0.213)
	Position 7	0.252* (0.115)	-0.155 (0.097)	-0.087 (0.214)
Tag Effect	Sponsored Tag	-0.673***(0.083)	-1.815*** (0.092)	-0.124 (0.167)
	Overall Pick Tag	2.538*** (0.093)	2.421*** (0.086)	3.175*** (0.205)
	Scarcity Tag	-0.674***(0.121)	-0.383*** (0.108)	-0.650* (0.264)
Attribute Effect	ln(Price)	-2.575***(0.099)	-2.371*** (0.092)	-2.517*** (0.214)
	ln(Rating Count)	0.800*** (0.030)	0.944*** (0.030)	0.814*** (0.064)
	Rating	9.701*** (0.314)	11.373*** (0.316)	5.673*** (0.537)

Significance is indicated as: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Models continue to exhibit heterogenous position bias in Headless interactions

	Position 1	Position 2	Position 3	Position 4	Position 5	Position 6	Position 7	Position 8
Claude Sonnet 4	30.2%	15.9%	15.7%	11.3%	9.5%	8.0%	5.2%	4.1%
GPT-4.1	27.4%	10.5%	12.7%	8.1%	12.8%	11.2%	8.0%	9.3%
Gemini 2.5 Flash	16.5%	14.3%	21.9%	7.6%	9.5%	8.9%	10.2%	11.1%

Estimated position "market shares" under identical products

Choice Behavior for API Style Agents

		Claude Sonnet 4	GPT-4.1	Gemini 2.5 Flash
Position Effect	Position 1	2.010*** (0.106)	1.078***(0.090)	0.394 (0.204)
	Position 2	1.370*** (0.107)	0.125 (0.095)	0.256 (0.209)
	Position 3	1.357*** (0.108)	0.312*** (0.094)	0.677***(0.203)
	Position 4	1.024*** (0.109)	-0.140. (0.097)	-0.373 (0.216)
	Position 5	0.857*** (0.111)	0.319*** (0.094)	-0.156 (0.218)
	Position 6	0.677*** (0.111)	0.184 (0.094)	-0.220 (0.213)
	Position 7	0.252* (0.115)	-0.155 (0.097)	-0.087 (0.214)
Tag Effect	Sponsored Tag	-0.673***(0.083)	-1.815*** (0.092)	-0.124 (0.167)
	Overall Pick Tag	2.538*** (0.093)	2.421*** (0.086)	3.175*** (0.205)
	Scarcity Tag	-0.674***(0.121)	-0.383*** (0.108)	-0.650* (0.264)
Attribute Effect	ln(Price)	-2.575***(0.099)	-2.371*** (0.092)	-2.517*** (0.214)
	ln(Rating Count)	0.800*** (0.030)	0.944*** (0.030)	0.814*** (0.064)
	Rating	9.701*** (0.314)	11.373*** (0.316)	5.673*** (0.537)

Significance is indicated as: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Choice Behavior for API Style Agents

		Claude Sonnet 4	GPT-4.1	Gemini 2.5 Flash
Position Effect	Position 1	2.010*** (0.106)	1.078***(0.090)	0.394 (0.204)
	Position 2	1.370*** (0.107)	0.125 (0.095)	0.256 (0.209)
	Position 3	1.357*** (0.108)	0.312*** (0.094)	0.677***(0.203)
	Position 4	1.024*** (0.109)	-0.140. (0.097)	-0.373 (0.216)
	Position 5	0.857*** (0.111)	0.319*** (0.094)	-0.156 (0.218)
	Position 6	0.677*** (0.111)	0.184 (0.094)	-0.220 (0.213)
	Position 7	0.252* (0.115)	-0.155 (0.097)	-0.087 (0.214)
Tag Effect	Sponsored Tag	-0.673***(0.083)	-1.815*** (0.092)	-0.124 (0.167)
	Overall Pick Tag	2.538*** (0.093)	2.421*** (0.086)	3.175*** (0.205)
	Scarcity Tag	-0.674***(0.121)	-0.383*** (0.108)	-0.650* (0.264)
Attribute Effect	ln(Price)	-2.575***(0.099)	-2.371*** (0.092)	-2.517*** (0.214)
	ln(Rating Count)	0.800*** (0.030)	0.944*** (0.030)	0.814*** (0.064)
	Rating	9.701*** (0.314)	11.373*** (0.316)	5.673*** (0.537)

Significance is indicated as: $*p < 0.05$, $**p < 0.01$, $***p < 0.001$

Implications for the Ecosystem

- **Platforms:** Traditional ads may lose value, while new services (like ”GEO-as-a-service") could emerge/MCP-like interface for AI shoppers to counteract position biases? Unclear given our headless experiments…/new role as ``seller’’
- **Sellers:** Risk of being invisible to agents/Need for cont. monitoring and potential for GEO with listings continuously tuned for different AI buyers via automated pipeline for simulation-based optimization
- **Consumers:** AI agents will reduce search friction, but risk sub-optimal and homogeneous choices
- **Regulators:** Concerns include market concentration, and the need for standardized reporting of agent testing beyond traditional failure rates on processes…

Concluding Remarks

- High level questions:
  - How will AI agents reshape the e-commerce ecosystem in the next 5 years?
  - Who will ``own'' the agents?
- Research Questions:
  - How to optimize agents' interactions/communication?
  - How will agents compete? Unintended consequences?
  - Behavioral economics for AI agents? And associated implications for operational decisions?

Thanks