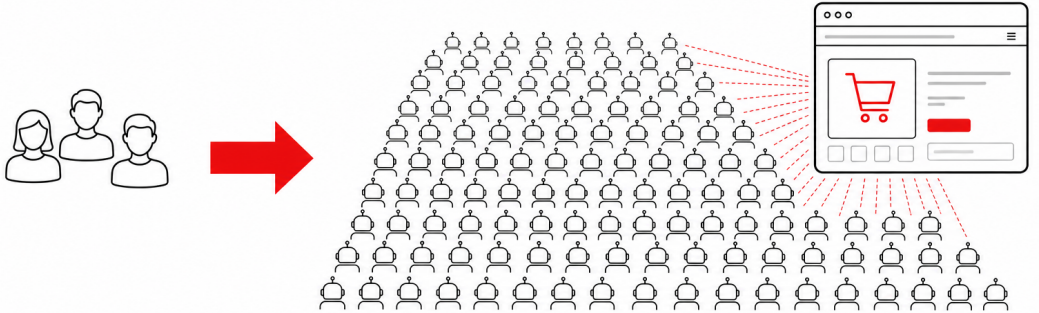
UXAgent: LLM agents simulate users at *small scale* to help UX researchers.

UXAgent: LLM agents simulate users at *small scale* to help UX researchers.

But real design decisions ship through *large-scale A/B testing*.

Can simulated users scale up?



Three Bottlenecks in A/B Testing

Northeastern University

No lightweight piloting

Scarce user traffic

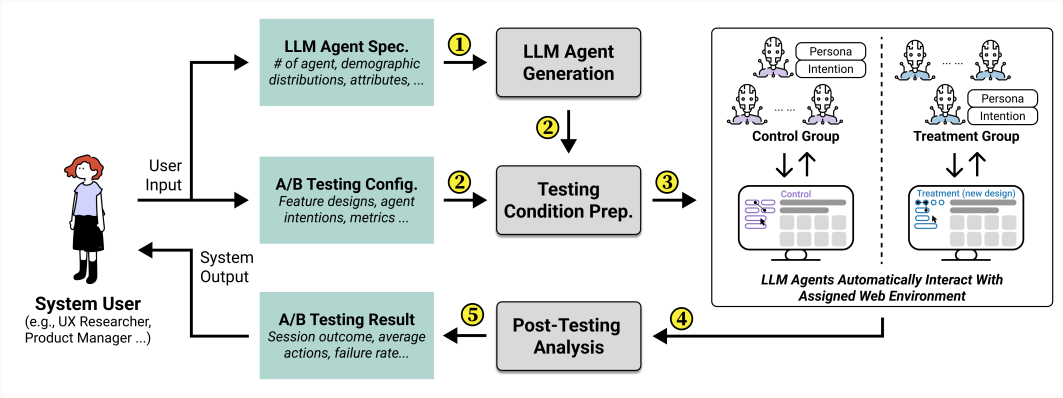
Slow feedback, weeks

From a formative study with 6 A/B-testing practitioners — many promising ideas never get piloted before launch.

Can we deploy thousands of persona-driven LLM agents on the *live* website

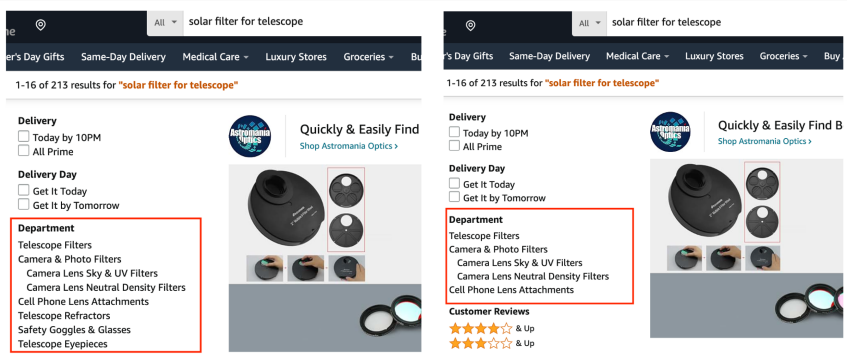
— to pilot an A/B test before spending real user traffic?

Our Solution: Agent A/B



Plug-and-play with existing agent stacks (Claude computer-use, ReAct).

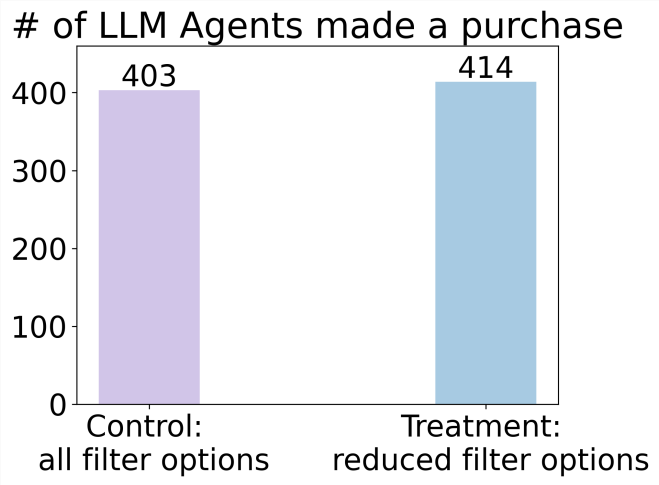
Case Study: Amazon.com



Control: full filter panel vs. **Treatment**: reduced panel (similarity ranking)
1,000 agents (500/condition) — ≈10% of a real A/B test.

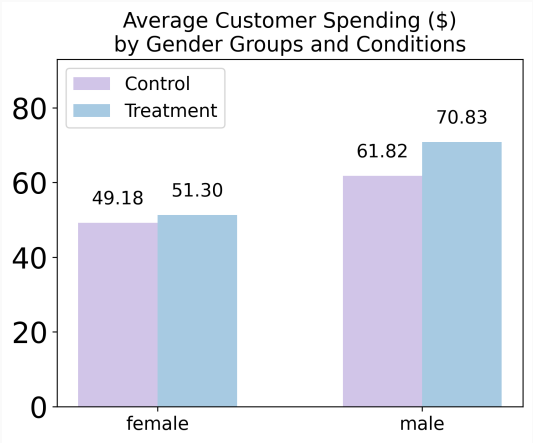
Results: Reduced Filter → More Purchases

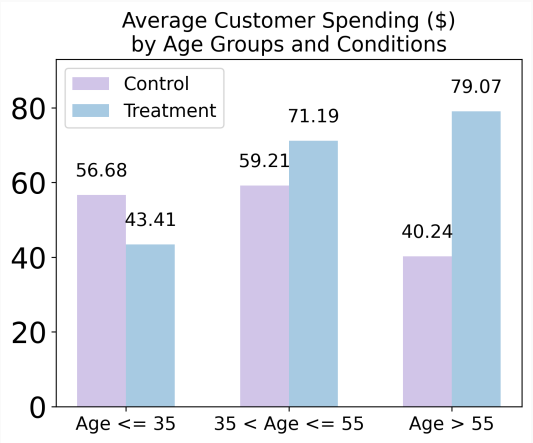
Northeastern University



414 vs. 403 purchases — a modest but statistically reliable increase.

Subgroup Effects





Effects are stronger for **male** and **older** personas.

Agent outcomes align *directionally* with a parallel large-scale human A/B experiment.

Simulated users can scale —
and point in the right direction.

Simulated users can scale —
and point in the right direction.

But “directionally aligned” $\neq$ “accurate”.

How faithful are these agents, really?