CommerceBench: Measuring Agent Performance on Real Commerce Tasks

lazer

CommerceBench: Measuring Agent Performance on Real Commerce Tasks

Evaluation, Benchmarks & RL Environments for Autonomous Commerce Agents

By Lazer

Executive Summary

AI is becoming a mainstream layer for shopping discovery, research, navigation, and transacting. As that happens, execution becomes a critical next step. Agents need to reliably find products, compare options, handle constraints (price, inventory, variants, shipping), manage carts, and initiate checkout.

The problem is we don't have a credible, repeatable way to measure how well agents actually perform end-to-end commerce workflows, and we don't have a safe, controllable environment to post-train them when they fail.

CommerceBench is our answer:

1. A deterministic evaluation harness and benchmarking layer (CommerceBench-Eval)
2. A high-fidelity synthetic commerce environment that can evolve into a Gymnasium-compatible RL training substrate (CommerceBench-Train)
3. A storefront and task generator (CommerceBench-Generator) that lets us scale scenario diversity without relying on the live internet

Our near-term focus is evaluation and benchmarking via commercebench.ai, where we:

- Quantify whether agentic commerce is actually "good enough" vs humans
- Compare models/labs head-to-head (success, reliability, efficiency, cost)

- Compare performance across commerce platforms (Shopify first, then others such as Salesforce Commerce Cloud, Adobe Commerce, Commercetools, and more)
- Compare interaction methods (Browser, UCP, MCP)

Our long-term ambition is bigger. For CommerceBench to become the default substrate for post-training and evaluating autonomous commerce agents.

We're starting with Shopify first, then expanding to other major commerce platforms after.

What We're Building

Lazer is building an Eval environment (CommerceBench) & running controlled evaluations on real Shopify storefronts and synthetic Shopify-like storefronts to measure agent performance on complete shopping workflows. Our goal is to build a benchmark & corresponding harness environment to allow for evaluating Large Language Models deterministically on realistic commerce tasks.

Our Hypotheses

1. LLM models demonstrate a failure rate on real world commerce tasks that is both consistent and significant.
2. The signature pattern of failure is the same or very similar for real storefronts and synthetic eval storefronts.

If we can demonstrate that the failure signature for our synthetic stores reliably tracks the failure patterns from real stores it would signal that our synthetic environment is a faithful proxy for real-world commerce complexity. That would mean we can safely use it as the primary substrate for training, tuning, and benchmarking autonomous commerce agents without dependence on live changing production storefronts.

This would put us in position to establish a fair deterministic benchmark for the realistic performance of LLMs as commerce agents.

Evidence

We can track success rates, step counts, failure taxonomies, traces, and full replayable sessions that demonstrate where and why agents fail on commerce tasks. By comparing these traces against each other and against human baselines, we can determine whether synthetic environments faithfully capture real-world commerce failures and track model progress over time.

What This Currently Is & Follow-Ons

What this currently is:

CommerceBench-Eval is a reproducible, end-to-end evaluation harness that measures whether agents can complete real commerce tasks and explains failures via traces and state diffs.

It also includes a public-facing benchmarking layer that measures how well agents can execute real commerce/shopping workflows end-to-end and makes results comparable across models, platforms, and interaction methods.

The leaderboard is designed to answer four practical questions:

1. **Are commerce agents actually usable today?** We quantify the gap between agent performance and human baselines on the same tasks.
2. **Which labs/models are winning?** We measure success rate, reliability, efficiency, and cost so models can see where they lag competitors and where regressions occur.
3. **Which commerce platforms are agent-friendly?** By running comparable task suites across platforms (starting with Shopify), we reveal where agents succeed or fail due to platform semantics and UX primitives. It also showcases how commerce platforms stack up against each other.
4. **Which interaction method works best?** We compare browser-based agents with protocol-based methods (e.g., UCP, MCP) to identify which surfaces reduce friction and improve success.

Each run produces auditable artifacts so that improvements are measurable and reproducible, including task specification, deterministic seed, action trace, and failure-mode labels.

What this isn't (yet):

A claim of SOTA training or a full-scale simulator.

Soon we will create CommerceBench-Train, a Gymnasium-compatible RL environment used for post-training agents on realistic commerce semantics at scale. This comes after we prove synthetic realism tracks real-store failure signatures.

Following this, we will create CommerceBench-Generator, a storefront and task synthesis system that will generate synthetic catalogs and realistic commerce primitives (variants,

promos, inventory, shipping rules), task variants and difficulty curricula, and controlled randomization across meaningful axes (not scraped UI noise).

Engagement Model

CommerceBench is designed to function as infrastructure and ideally be built and validated through direct co-development with frontier labs and commerce platforms rather than as a closed or speculative research project.

We're looking for investment, access, and resources that will help to accelerate CommerceBench's development. In return we offer partners early access to a custom evaluation environment with direct influence over its design so that the environment is aligned with your research goals.

If interested, please reach out to us at **founders@lazertechnologies.com**.

What labs/partners get:

- Direct influence over environment design
- Alignment with agent architecture
- Early / exclusive access window
- Evaluation environment tailored to research goals
- Speed of execution and depth of expertise in this domain

The Problem

Agents are increasingly used for web tasks. Commerce represents one of the most economically critical applications, with platforms like Shopify processing $300B+ in annual GMV. But we don't have reliable measurements on how agents perform on real storefronts.

As well, autonomous commerce agents will only be as good as the environments they are trained in.

Commerce tasks expose a unique combination of challenges:

- Long-horizon, multi-step execution
- Partial observability (hidden inventory, shipping logic, pricing rules)
- Non-stationary environments (promotions, stock changes)
- High action branching factor (navigation, filtering, configuration)
- Safety and policy surfaces (fraud, restricted items, deceptive UX)

We need measurements that reflect real shopping behavior. Current approaches fall short because the open internet is unusable for controlled evaluation (non-deterministic, slow, unsafe, legally risky), and generic web benchmarks abstract away commerce semantics.

Why The Problem Matters

From a Commerce Platform Perspective

From a commerce platform (starting with Shopify) perspective, agentic commerce is a platform-defining shift. If autonomous agents become a primary interface for shopping, the platforms those agents are trained and evaluated on will define where value accrues.

Platform-compatible agents create a defensive moat against closed assistants by ensuring agents work best on merchant-owned storefronts rather than centralized marketplaces. They unlock a new developer surface area for agent-aware extensions and potentially a new class of "agent apps" (e.g. best discount-finding agent, sustainability-optimized agent, etc), while increasing GMV through improved conversion and reduced friction.

CommerceBench provides the training and evaluation infrastructure required to make this transition safe, reliable, and merchant-aligned—positioning platforms such as Shopify as the default substrate for agent-mediated commerce.

From a Frontier AI Research Lab Perspective

Shopify and other key commerce platforms represent the largest and most strategically important entry points for agent-mediated commerce. Shopify merchants, for example, process roughly $300B+ in annual GMV, spanning millions of stores across consumer, enterprise, subscription, and omnichannel commerce.

If autonomous agents meaningfully mediate even 5–10% of Shopify GMV—through discovery, comparison, replenishment, or end-to-end purchase execution—that corresponds to $15–30B in agent-mediated commerce on Shopify alone.

At a conservative 0.5–1.0% take rate, this implies $75–300M in annual revenue potential within the Shopify ecosystem, before expanding to other platforms. This makes reliable, aligned commerce execution a first-order strategic capability for frontier AI labs.

Overall

If autonomous agents become a primary interface for commerce:

- The agents that win will be trained correctly
- The labs that win will own execution benchmarks
- The commerce platforms that are agent-friendly will succeed
- The environments that matter will be semantic, not scraped

Why Shopify First?

Shopify represents the most structurally sound and economically meaningful substrate for benchmarking autonomous commerce agents. Shopify merchants process $300B+ in annual GMV across millions of independently operated stores spanning DTC, enterprise, subscription, and omnichannel commerce.

By anchoring CommerceBench in Shopify semantics, we get real-world relevance, variation, and economic impact.

Following Shopify, we plan on expanding coverage to other commerce platforms such as Salesforce Commerce Cloud, Adobe Commerce (Magento), Commercetools, WooCommerce, and more. We will then be able to understand which platforms are agent-friendly and help agents perform better across all key platforms that power commerce around the world.

Why Synthetic Environments Matter

Real sites are slow, fragile, and legally tricky. Synthetic sites allow scale, ablations, and controlled difficulty.

More important: synthetic storefronts can be generated offline using real Shopify semantics, not scraped HTML. This means we can randomize across multiple axes while preserving commerce logic:

- Industry vertical
- Layout and navigation patterns
- Inventory volatility
- Variant complexity
- Pricing and promotion logic
- Subscription mechanics
- Checkout architecture

If these evaluations become trusted, a conversion from pure Eval to a Training Environment could follow naturally.

Evidence Plan: Real vs Synthetic Storefront Correlation

CommerceBench-Eval can test whether synthetic storefronts reproduce the main failure modes and approximate difficulty observed on real storefronts.

Item	Specification
Task set	10–20 end-to-end tasks spanning search/browse → product detail → variant selection → cart → checkout initiation
Real sites	Using 2–4 production Shopify stores (read-only; stop at checkout)
Runs	3–5 runs per task with fresh sessions
Metrics	Success rate, median steps, time-to-success, failure taxonomy
Artifacts	JSON traces with state transitions, Replay scripts for deterministic reproduction, 3–5 "golden failure" traces with screenshots and state diffs

Claim we're testing: Synthetic storefronts track real-world failures in a way that makes them useful for scaled evaluation and eventual training.

This is the unlock. If we can show that agents fail on synthetic environments the same way they fail on real ones, we've earned the right to talk about training infrastructure.

Real Storefront Safety & Compliance Acknowledgment

Our use of real storefronts for evaluation is limited by safety and compliance protocols. All runs are **read-only**, strictly designed to **stop at checkout (no purchase)**. We ensure **no PII is captured**, and agents are configured to **respect rate limits** and **not bypass bot defenses** to maintain site stability. Furthermore, we explicitly ban any actions that could be construed as **illegal** or as a Denial of Service (**DDoS**).

Our Initial Solution: CommerceBench-Eval

CommerceBench-Eval is a reference evaluation suite for commerce agents, built to measure performance on realistic commerce workflows.

Core Capabilities

Task suite: Structured tasks covering discovery, constraint reasoning, variant selection, cart management, and checkout initiation. Each task is defined as a program with explicit intent, hard constraints, and success criteria.

Scoring and replay: Deterministic environment resets, reproducible trajectories, and sparse terminal rewards. Every run produces machine-readable traces that can be replayed exactly.

Failure taxonomy: Clear categorization of semantic failures—wrong variant selected, constraint violated, checkout dead-end, infinite loop—that explain where agents break down.

Observation Modality

We use browser accessibility trees as observations. This is the standard semantic representation exposed for assistive technologies—roles, labels, hierarchy for buttons/inputs/links. It provides a structured, stable view of actionable UI elements that maps cleanly to agent actions and is less brittle than pixel-only approaches.

Researchers can also opt into screenshot streams for debugging or multimodal experiments. Raw DOM is intentionally excluded due to instability and noise.

Action Space

Agents interact via discrete, human-equivalent actions:

- Click(node_id)
- Type(node_id, text)
- Scroll(direction, amount)
- Navigate(semantic_intent | url)

- Terminate(task_complete)

All actions emit step-level observations, rewards, and termination signals.

Task Programs

Tasks are defined as structured programs, not natural language alone. This enables curriculum learning, constraint reasoning, failure attribution, and controlled generalization tests.

Field	Example
intent	"Purchase running shoes"
constraints	"price_max": 120, "size": 10, "shipping_days": "<=3"
success	"add_to_cart": true, "checkout_started": true

More About Lazer

Deep Expertise

160+ Senior Engineers, Designers, and Product Builders

At Lazer, our team of 160+ consists of only senior, product-minded, full-stack engineers, architects, product managers, and designers. We have a blend of startup and enterprise experience, fully focused on solving the most important problems enterprises are dealing with. Our team members have backgrounds from early-stage startups, large enterprises (e.g. Coinbase, Google, Meta, Shopify, etc), and some of the fastest growing companies in key tech hubs such as Silicon Valley.

Vast Commerce & Shopify Expertise

We are experts in every aspect of Shopify as a platform and are one of Shopify's most preferred agency partners. We are one of the most active developers in the Shopify app ecosystem, build products internally for Shopify around core departments, and help some of the most important and largest enterprises on Shopify build complete experiences and be as successful as possible across online, retail and back-office. Some of these enterprises include Mattel, SKIMS, Gibson, Alo Yoga, October's Very Own, Bombas, Eddie Bauer, Athletic Greens, PayPal, Uniswap, Authentic Brands Group (Champion & Reebok), Kraken, Coinbase, Cox Automotive, Royal Bank of Canada, Shopify, Alo Yoga, VF Corporation, and more.

Over the past 6 years, we also have worked very closely with Shopify executives, VPs of product and engineering, Shopify's professional services team and others on building internal products for Shopify as well as ensuring their largest and most complex enterprises build and launch successfully. We are now one of Shopify's most trusted engineering and design agencies.

Deep AI Capabilities

We have an entire division that helps fast growing startups and major enterprises integrate AI directly into their core value propositions, ranging from mental health and resource management to generative media and next-generation hardware interfaces. We design the pipelines and agentic workflows that represent the scalable infrastructure required to leverage LLMs for business operations.

Global Presence

In terms of global presence, our entire team is based in Canada, USA, and the UK. Our global presence gives us a strong understanding of regional nuances to commerce, e.g. GDPR, internationalization, etc., and allows us to have high communication and high collaboration throughout the engagement regardless of timezones and stakeholder locations.

Why Lazer

Lazer is uniquely positioned to deliver in this context due to a wealth of experience and industry integration across both commerce and AI.

For more information on our commerce and AI work, feel free to explore some of our portfolio here:

- https://www.lazertechnologies.com/industries/artificial-intelligence
- https://www.lazertechnologies.com/industries/commerce

Lazer Clients & Case Studies

Some of our clients include: Hinge, AG1, Reckitt, Retool, Ohalo, Salt, Polymarket, Uniswap, Google, Abridge, Census, Readwise, Ro, Coinbase, Shopify, ClassDojo, Extropic, Motion, GoBolt, Retool, Berachain, XMTP, Alchemy, Stanford, Iconiq, Kraken, General Catalyst, OpenSea, Axio

For more information, please visit our website: https://www.lazertechnologies.com/
Email: founders@lazertechnologies.com