Product Truth in the Age of Agentic Commerce: A Multi-Platform Evaluation of AI Shopping System Accuracy, Completeness, and Regulatory Reliability

Abstract

AI-powered shopping assistants such as ChatGPT, Gemini, Perplexity, Claude, and Copilot are increasingly marketed as end-to-end "agentic commerce" systems, yet they depend on web-scraped content, merchant feeds, and partner integrations that may be incomplete, volatile, or misaligned with regulatory obligations and product truth. This paper introduces product truth—a SKU-level representation that is factually accurate, complete on key attributes, and aligned with relevant regulatory regimes—and evaluates whether leading AI shopping platforms can reliably deliver it. We test five major platforms on 100 product scenarios spanning ten categories: five regulation-sensitive (cosmetics, cleaning, batteries, aerosols, paints) and five lower-regulation controls (apparel, home, food, kitchen, books). Across 2,500 evaluated steps using a chained five-step protocol, full success is achieved in 28.2% of cases, with 36.5% partial and 35.2% failure. Notably, attribute completeness is substantially weaker in regulated categories ($\mu = 0.46/2$) than controls ($\mu = 1.18/2$), and availability lookup fails in 52.7% of cases. Platform performance varies dramatically—from ChatGPT (43.4% success) to Perplexity (13.6%)—with checkout feasibility ranging from functional (Gemini: 0.78/2) to almost non-existent (Claude: 0.02/2). We argue that while AI shopping systems have improved, persistent gaps in availability verification, regulated-product completeness, and checkout execution mean they cannot yet reliably serve as stewards of product truth.

1. Introduction

In 2023, tools like ChatGPT and Bard were starting to become commonplace for product discovery, gift recommendations, and comparisons.[1] Today, early reports of 2025 holiday shopping have already shown a surge of LLM usage for discovery, gifting, and checkout.[2] Despite this, AI agents still misjudge availability, return outdated offers, or send users to unavailable SKUs.Over the past two years, large language models (LLMs) have moved from peripheral “shopping helpers” to the center of e-commerce strategy. Industry analysts now project that by 2030, agentic commerce could drive as much as $1 trillion in orchestrated revenue in the U.S. B2C retail market alone, with global estimates reaching $3–5 trillion.[3] Major technology firms and retailers are racing to build these agents: OpenAI’s ChatGPT has introduced

[1] Lu, Yiwen. "My Not-So-Perfect Holiday Shopping Excursion With A.I. Chatbots" The New York Times, 14 Dec. 2023, https://www.nytimes.com/2023/12/14/technology/shopping-ai-chatbots.html.
[2] Rocha, Natalie & Rhone, Kailyn. "A.I. Can Do More of Your Shopping This Holiday Season" The New York Times, 25 Nov. 2025, https://www.nytimes.com/2025/11/25/technology/chatgpt-holiday-shopping.html.
[3] "The Agentic Commerce Opportunity: How AI Agents Are Ushering in a New Era for Consumers and Merchants." McKinsey & Company, QuantumBlack, 2025, https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-agentic-commerce-opportunity-how-ai-agents-are-ushering-in-a-new-era-for-consumers-and-merchants.

Instant Checkout and an Agentic Commerce Protocol that claims to allow users in the United States to purchase products from Etsy and, soon, Shopify merchants and Walmart, directly within the chat interface.[4] Google has launched an AI Mode for shopping, powered by its Shopping Graph of more than 50 billion product listings, with claims that over 2 billion listings are refreshed hourly.[5] Perplexity.ai has rolled out Buy with Pro, a one-click checkout feature integrated with a proprietary merchant program and Shopify-based storefronts.[6]

Collectively, these developments are framed as the advent of "agentic commerce" or "chat-driven commerce" in which LLMs move beyond recommendation to end-to-end transaction execution. However, beneath the polished conversational interfaces, these systems remain critically dependent on the quality and coverage of underlying product data. They rely on a mixture of merchant feeds, web-scraped pages, proprietary graphs (e.g., Google's Shopping Graph), and partner integrations, all of which can be incomplete, biased, or abruptly revoked—as illustrated when users discovered Amazon recently blocked OpenAI's crawlers according to its public robots.txt file, limiting access to a major share of U.S. retail listings.[7] In short, despite notable advances in AI-driven shopping interfaces, the technology remains dependent on high-quality metadata, which is still inconsistently available.

In domains such as cosmetics, cleaning, batteries, and other regulated goods, small discrepancies in product data—for example in formulation, battery chemistry, hazard statements, or regional availability—can have outsized implications for safety, compliance, and consumer trust. Yet current evaluations of AI shopping tools focus largely on user experience, personalization, or click-through metrics, rather than on whether the systems provide correct, complete, and regulation-aligned information about the products themselves.

Retailers have begun taking steps to manage and shape the increasing adoption of AI-based shopping by consumers. Walmart has reported that their relationship with OpenAI has cut down customer care resolution times by up to 40%,[8] and Lowe's "Mylow" companion, powered by OpenAI, handles nearly 1 million product/shopping related questions per month.[9] According to research from IBM, 76% of executives at retailers are already using or plan to use AI for increased revenue streams and operational efficiencies.[10] At the same time, when prompted regarding the use of proprietary data for AI models, retailers reported that only 64% of this data is accessible, 49% is usable, and 26% of it is used so far.[11] AI agents are constrained by the quality of the structured data ecosystems in which they operate; when as much as 74% of proprietary data remains unused, inconsistencies in product data significantly limit their

---

[4] "Buy It in ChatGPT." OpenAI, 2025, https://openai.com/index/buy-it-in-chatgpt/.
[5] Srinivasan, Vidhya. "Let AI do the hard parts of your holiday shopping." Google, The Keyword Blog, 13 Nov. 2025, https://blog.google/products/shopping/agentic-checkout-holiday-ai-shopping/.
[6] "Shop Like a Pro." Perplexity AI, 2025, https://www.perplexity.ai/hub/blog/shop-like-a-pro.
[7] Smith, Allison. "Amazon Quietly Blocks More of OpenAI's ChatGPT Web Crawlers from Accessing Its Site." Modern Retail, 21 Nov. 2025, https://www.modernretail.co/technology/amazon-quietly-blocks-more-of-openais-chatgpt-web-crawlers-from-accessing-its-site/.
[8] "Walmart Partners with OpenAI to Create AI-First Shopping Experiences." Walmart Corporate News, 14 Oct. 2025, https://corporate.walmart.com/news/2025/10/14/walmart-partners-with-openai-to-create-ai-first-shopping-experiences.
[9] "The State of Enterprise AI: 2025 Report." OpenAI, 2025, https://cdn.openai.com/pdf/7ef17d82-96bf-4dd1-9df2-228f7f377a29/the-state-of-enterprise-ai_2025-report.pdf.
[10] "Retail and Consumer Products in the AI Era." IBM Institute for Business Value, 2025, https://www.ibm.com/thought-leadership/institute-business-value/en-us/report/retail-consumer-products-in-ai-era.
[11] Ibid.

reliability. With AI platforms now serving as a product-search channel for nearly one-quarter of young adults and AI accounting for one-third of US market value, the expansion of AI-based shopping tools heightens the importance of ensuring product-data accuracy.[12] [13]

1.1 From Recommender Systems to Agentic Commerce

Research on recommender systems (RS) and e-commerce personalization is extensive, spanning collaborative filtering, content-based approaches, and hybrid models. Surveys and empirical studies have documented gains in engagement and sales when RSs are tuned for relevance and personalization.[14] More recent work has examined safety, fairness, and bias in "deep" recommender systems, noting that model training and feedback loops can amplify popularity bias, demographic disparity, or exposure inequality.[15] Parallel literatures on conversational commerce and AI shopping assistants describe how chatbots and voice agents can reduce friction in search, support interactive product comparison, and provide 24/7 support.[16] These systems are typically evaluated using user satisfaction scores, conversion uplift, or task-completion time. According to Google themselves, as generative models have entered production, industry case studies emphasize their ability to "think through" trade-offs, generate comparison tables, and synthesize review content.[17] However, in both lines of work, product data is primarily treated as input, not as an object of study. RS and conversational-commerce papers often assume that catalog data is correct, complete, and stable. Evaluation focuses on ranking quality (precision@k, NDCG) and conversational usability—not on whether the underlying product knowledge is itself accurate, especially with respect to safety, regulation, or fine-grained variants.

Recent advances in state-space modeling underscore a structural weakness in many deep AI systems: without reliable mechanisms for tracking and updating an internal "state," models struggle when inputs are inconsistent or incomplete.[18] Recommender-systems and conversational-commerce research largely overlooks this issue, treating product data as a clean input rather than a dynamic, error-prone substrate. As a result, improvements in ranking quality or conversational fluency mask a deeper constraint: AI systems cannot reliably personalize, compare, or reason when the underlying product data is itself unstable. And, the resulting post-transactional cost of incorrect data is not calculated or tracked. This cost is substantial.

1.2 Web-Scraped and Merchant-Provided Product Data

---

[12] Gaudiaut, Tristan. "U.S. Consumers Warm Up to AI Shopping Tools" Statista, 11 Nov. 2025, https://www.statista.com/chart/35442/ai-tools-usage-e-commerce/.

[13] Hyatt, Diccon. "AI Now Accounts for a Third of US Market Value—What That Means For The Economy" Investopedia, 03 Nov. 2025, https://www.investopedia.com/the-u-s-economy-is-putting-all-its-chips-down-on-a-i-11841060.

[14] Shen, Huawei, et al. "The rising safety concerns of deep recommender systems" PubMed Central (PMC), 12 Jul. 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC12529586/.

[15] Ibid.

[16] "Gartner Predicts Agentic AI Will Autonomously Resolve 80 Percent of Common Customer Service Issues Without Human Intervention by 2029." Gartner Newsroom, 5 Mar. 2025, https://www.gartner.com/en/newsroom/press-releases/2025-03-05-gartner-predicts-agentic-ai-will-autonomously-resolve-80-percent-of-common-customer-service-issues-without-human-intervention-by-20290.

[17] "AI is transforming shopping in Search. Here's what to know." Google Business, Think with Google, May 2025, https://business.google.com/us/think/search-and-video/google-shopping-ai-mode-virtual-try-on-update.

[18] "The quest to teach LLMs how to count." IBM Research Blog, 05 Dec. 2025, https://research.ibm.com/blog/state-tracking-for-state-space-models.

The data foundations of LLM-based shopping systems closely resemble those used in web-scale RS research: scraped e-commerce pages, structured product feeds, and review corpora. A growing literature warns that web-scraped data is systematically biased due to volatility, personalization, and lack of a well-defined sampling frame.[19] These issues directly affect product datasets: prices and availability change rapidly; content is personalized by location and user history; and key disclosures (e.g., ingredients, hazard warnings) may appear only in certain regional or device-specific views. Studies of scraping pipelines for product recommendation highlight heterogeneous data formats, inconsistent attribute coverage, and challenges in aligning SKUs and variants across sites.[20] Work on search and recommendation bias further documents position effects, presentation bias, and model-induced feedback loops that shape which products are seen and clicked, independent of intrinsic quality.[21]

Despite these concerns, commercial AI shopping solutions lean even more heavily on these data sources. Google's Shopping Graph merges merchant feeds, schema.org markup, and crawl data into a unified product graph; retailers are encouraged to "enrich attributes beyond basics" and maintain structured offer data to maximize eligibility for AI-mode responses.[22] Perplexity's merchant program incentivizes retailers to share detailed product specs by promising better recommendation visibility and one-click checkout integration.[23] OpenAI's Agentic Commerce Protocol, in turn, seeks to onboard merchants at scale via Stripe and PayPal integrations, automatically ingesting catalog data from compatible platforms such as Etsy, Shopify, and Walmart.[24]

Notably, recent industry developments reflect an implicit recognition of the limits of LLM-only product understanding. OpenAI has introduced a Product Feed Specification for ChatGPT shopping that requires participating merchants to submit structured product data—including GTIN or UPC identifiers, brand, pricing, availability, and offer metadata—as a designated "structured source of truth" for commerce interactions.[25] This requirement signals that large language models operating primarily over unstructured or opportunistically scraped content cannot reliably maintain SKU-level accuracy, freshness, or transactional grounding without explicit, structured inputs.

Taken together, these architectures expose a central tension in agentic commerce. While AI shopping systems are increasingly positioned as end-to-end agents capable of recommendation, comparison, and transaction execution, they remain dependent on fragmented, partner-mediated data pipelines whose coverage, consistency, and regulatory completeness vary widely. Whether such systems can deliver *product truth*—accurate identity, complete attributes, regulatory alignment, and grounded availability—at scale remains an open and largely unmeasured question, motivating the empirical evaluation undertaken in this study.

---

[19] Foerderer, Jens. "Should we trust web scraped data?" arXiv, 07 Aug. 2023, https://arxiv.org/pdf/2308.02231.
[20] Kaur, Amarinder & Prashar, Deepak. "Web Scraping for Product Recommendations: A Review of Techniques and Applications." Journal of Computer Science, vol. 21, 2025, pp. 1425–1439, https://thescipub.com/abstract/jcssp.2025.1425.1439.
[21] Grebennikov, Roman. "Dealing with Position Bias in Recommendations and Search." KDnuggets, 14 Mar. 2023, https://www.kdnuggets.com/2023/03/dealing-position-bias-recommendations-search.html.
[22] "AI is transforming shopping in Search. Here's what to know." Google Business.
[23] "Shop Like a Pro." Perplexity AI.
[24] "Buy It in ChatGPT." *OpenAI*
[25] OpenAI Product Feed Specification: *Structured Product Feeds for ChatGPT Shopping*, https://developers.openai.com/commerce/specs/feed/.

1.3 Conceptualizing Product Truth

We use the term **product truth** to denote a SKU-level representation that is (i) factually accurate, (ii) complete with respect to key attributes, and (iii) aligned with relevant regulatory regimes. For a given product, this includes:

- canonical identity of the SKU and its variants (size, shade, formulation, packaging);
- ingredient or component lists and relevant thresholds;
- hazard classifications (e.g., flammability, battery chemistry, waste codes);
- applicable state, federal, or international regulations and warnings;
- merchant and channel information (primary seller vs marketplace reseller);
- geo-specific availability, pricing, and delivery constraints.

In regulated or safety-critical categories, a system that recommends a near-match but incorrect variant may meaningfully mislead the user. For example, returning an older aerosol formulation without updated VOC rules, or a battery pack with a different chemistry than the one described. Yet current AI shopping evaluations rarely, if ever, assess these dimensions.

1.4 Research Gap

Existing work on conversational AI in retail demonstrates that generative models can improve perceived relevance and reduce search friction, but does not directly measure their fidelity to ground-truth product data. Studies of recommender-system fairness and safety highlight systemic biases, yet focus primarily on user or item groups (e.g., genres or demographic segments) rather than on attribute-level correctness for individual products.[26] Research on web-scraped datasets similarly documents sampling bias, volatility, and attribute sparsity, but stops short of evaluating how these issues surface in end-user AI shopping systems operating in real time.[27]

Industry-led evaluations, including OpenAI's *ChatGPT Shopping Research,* demonstrate that large language models can successfully surface products, generate comparisons, and assist users in navigating shopping decisions.[28] These benchmarks typically assess performance using metrics such as task success, response preference, coverage breadth, and usability, providing valuable evidence that AI systems can function as effective shopping assistants. However, both academic and industry evaluations implicitly assume that the product information surfaced by these systems is accurate, complete, and up to date. They do not measure SKU-level correctness, variant resolution, attribute fidelity, regulatory alignment, or transactional grounding. As a result, existing benchmarks offer limited insight into whether AI shopping systems can reliably deliver *product truth*—particularly in domains where small discrepancies in formulation, variant, or regulatory status have material consequences.

The recent emergence of *agentic commerce*, in which LLM-based systems move beyond recommendation to execute or simulate purchases, further raises the stakes of this omission. When an AI assistant

---

[26] Shen, Huawei, et al. "The rising safety concerns of deep recommender systems."
[27] Foerderer, Jens. "Should we trust web scraped data?"
[28] OpenAI, Introducing shopping research, a new experience in ChatGPT, OpenAI website, Nov. 24, 2025, https://openai.com/index/chatgpt-shopping-research/

independently selects a product, misjudges stock status, or fails to surface a critical hazard disclosure, responsibility becomes diffused across model providers, retailers, and data sources. Yet, to our knowledge, no systematic, cross-platform study has evaluated how well current AI shopping agents maintain ground-truth product integrity across chained, multi-step shopping tasks.

1.5 Objectives and Research Questions

This paper addresses this gap by conducting a multi-platform, multi-category evaluation of AI shopping systems' ability to deliver product truth. We focus on widely deployed, general-purpose LLM-based shopping tools: OpenAI ChatGPT with Instant Checkout, Google Gemini AI Mode, Perplexity with Buy with Pro, and, for comparison, Claude and Microsoft Copilot in shopping scenarios.

Our study is guided by the following research questions:

- **RQ1:** To what extent do leading AI shopping systems provide product information that matches a high-quality ground-truth dataset in terms of accuracy, completeness, and regulatory correctness?
- **RQ2:** How do these systems perform across different product categories, particularly those that are safety- or regulation-sensitive (e.g., cosmetics, cleaning products, batteries)?
- **RQ3:** What characteristic **failure modes** emerge in chained shopping tasks that require the agent to combine discovery, specification lookup, localization, and checkout?
- **RQ4:** How do platform data dependencies (web scraping, merchant feeds, partner programs) shape coverage, bias, and the stability of product truth over time?

1.6 Hypotheses

Based on prior work on web-scraped data quality, recommender bias, and the architecture of commercial shopping graphs, we posit:

- **H1 (Product Truth Deficit):** AI shopping systems that rely primarily on public web data, merchant feeds, and limited partner integrations will exhibit statistically significant deficits in product truth—measured as SKU- and attribute-level accuracy, completeness, and regulatory correctness—relative to a structured ground-truth dataset.
- **H2 (Category Sensitivity):** Product-truth deficits will be larger in regulation-heavy or compositionally complex categories (for example, formulated products and hazardous goods) than in lower-regulation categories such as apparel.
- **H3 (Agentic Failure Accumulation):** IIn chained, multi-step shopping tasks that require systems to identify products, retrieve attributes, localize availability, and approach checkout, small inaccuracies in early steps will compound over time, producing a high rate of end-to-end task failure.

To test these hypotheses, we construct a benchmark of queries and chained tasks across multiple categories and compare AI system outputs against a curated product-truth reference set derived from structured regulatory and product data.

1.7 Contributions

This work makes three main contributions:

1. **Conceptual:** We formalize *product truth* as a measurable construct for evaluating AI shopping systems, bridging literatures on recommender systems, web-scraped data quality, and regulatory data management.
2. **Empirical:** We provide the first cross-platform, multi-category benchmark assessing commercial LLM-based shopping agents on attribute-level accuracy, completeness, and regulatory correctness, along with a taxonomy of observed failure modes.
3. **Design and Governance:** Drawing on our findings, we articulate requirements for product-truth-first architectures—built on structured, verified product intelligence rather than opportunistic scraping—and discuss implications for retailers, model providers, and regulators in an era of agentic commerce.

The remainder of the paper details our methodology, experimental setup, evaluation metrics, and results, before turning to implications for the design of safer, more reliable AI shopping infrastructures.

2. Methods

To operationalize product truth in the context of agentic commerce, we designed a multi-platform, multi-step evaluation that approximates a realistic AI-driven shopping journey. The study focuses on chained interactions in which an AI system must (i) identify a specific SKU and variant, (ii) surface safety- and regulation-relevant product attributes, (iii) localize stock, pricing, and delivery, and (iv) progress toward transaction execution. These stages mirror the progression from recommendation to autonomous or semi-autonomous checkout emphasized in current deployments of LLM-based shopping agents.

Consistent with the conceptual framework introduced in Section 1, we evaluate commercial AI systems against a structured ground-truth dataset along four dimensions: identity accuracy, attribute completeness and correctness, regulatory reliability, and transactional reliability. We treat the systems under study not as abstract language models, but as shopping infrastructures that depend on web-scraped data, merchant feeds, and partner integrations. Accordingly, the evaluation centers on SKU-level and attribute-level comparisons rather than on user satisfaction or ranking metrics. In addition to measuring product truth, we also independently evaluate whether systems function as usable shopping agents, including their ability to execute checkout flows, align product selection with user intent, and deliver results efficiently. These agentic dimensions are reported separately from product-truth scores to avoid conflating correctness with usability.

This research constitutes a focused empirical test of the hypotheses articulated in Section 1.6, with particular emphasis on H1 (Product Truth Deficit) and H3 (Agentic Failure Accumulation).

Crucially, the chained design enables direct observation of error propagation across steps. Early misidentifications or hallucinated attributes can influence downstream availability and checkout behaviors, revealing whether systems correct or amplify initial inaccuracies. This structure allows empirical measurement of agentic failure accumulation, a phenomenon that cannot be observed in single-turn or shallow shopping benchmarks.

2.1 Platforms

We evaluated five widely deployed systems:

- **OpenAI ChatGPT** (GPT-5.2)
- **Google Gemini Pro** (Flash 2.5)
- **Perplexity Pro** (online RAG with live web crawling; custom merchant program)
- **Anthropic Claude** (Opus 4.5)
- **Microsoft Copilot** (GPT-5.1 derivative + Bing Shopping Graph)

While the experiment treats each system as a black box for evaluation, we incorporate *reverse-engineered architecture inferences* based on public documentation and empirical behavior:

Table 1
Core System Components and Their Impact on Product Truth Integrity

Stack Component	Likely Implementation	Relevance to Product Truth
Tokenization	GPT-style BPE and WordPiece variants	Token granularity affects attribute fidelity (e.g., splitting chemical names) → variant confusion and ingredient misidentification
Embeddings	LLM-derived dense embeddings (OpenAI text-embedding-3-large, Google's Gecko)	Embedding drift and semantic collapse → variant confusion and state drift
Retrieval / RAG	Perplexity: live crawler; ChatGPT: Browse + merchant feeds; Gemini: Shopping Graph	Retrieval gaps and coverage inconsistencies → missing or hallucinated attributes
Databases	Vendor shopping graphs + merchant feeds (Stripe, Shopify, Etsy, Walmart)	Feed freshness variance and blending without provenance → stale price, wrong stock
Crawlers	Perplexity custom agent; Bing crawler; Googlebot	Robots.txt restrictions and coverage limits → systematic blind spots

These architectural components produce measurable constraints on SKU-level truth, detailed further in Section 4.

2.2 Product Categories and Scope

To ground the evaluation in domains with varying regulatory complexity and attribute requirements, we evaluated 100 SKUs spanning ten product categories, evenly divided between regulation-sensitive and lower-regulation control categories.

In-scope for regulatory evaluation (5 categories, 50 SKUs). These categories require accurate, SKU-specific safety attributes and can trigger jurisdiction- or carrier-relevant disclosures. Regulatory correctness was scored against structured ground-truth data:

- **Cosmetics (COSM):** Emphasizes variant precision (shade, finish, formulation) and ingredient completeness, with potential allergen disclosures and safety warnings.
- **Cleaning Products (CLEAN):** Emphasizes chemical disclosure, hazard communication, and SDS-relevant attributes.
- **Battery-Containing Products (BATT):** Emphasizes regulated shipping classifications, carrier constraints, and battery chemistry specifications.
- **Aerosols (AERO):** Requires flammability warnings, VOC compliance, and pressurized-container handling disclosures.
- **Paints (PAINT):** Involves VOC limits, hazardous material classifications, and state-specific regulatory flags.

Out-of-scope for regulatory evaluation (5 categories, 50 SKUs). These categories were included as lower-regulation controls where safety-critical disclosures are generally not applicable. Regulatory correctness was not scored for these categories:

- **Apparel (APPAR):** Emphasizes size, color, and material accuracy without safety-critical disclosures.
- **Home Goods (HOME):** Tests attribute completeness for products with variable complexity, from simple décor to items with assembly considerations.
- **Food Products (FOOD):** Emphasizes allergen labeling and nutritional accuracy; while food may have some regulatory dimensions, food-specific compliance (FDA labeling, nutrition panels) was excluded from the ground-truth dataset to avoid conflating retail product truth with nutrition policy compliance.
- **Kitchen Items (KITCH):** Spans simple utensils and appliances, testing attribute completeness without regulated-product complexity.
- **Books (BOOK):** Emphasizes edition, format, and variant accuracy (e.g., hardcover vs. paperback, exclusive editions) with minimal regulatory burden.

This balanced 5/5 design intentionally spans a spectrum of "truth burdens" and enables direct comparison between regulation-sensitive and control categories. The in-scope categories test whether systems can reliably surface safety-critical information when it matters; the out-of-scope categories test whether product-truth performance differs when regulatory complexity is removed from the equation.

Each SKU was paired with a natural-language shopping scenario framed as a consumer consideration query (e.g., "I'm considering buying [PRODUCT]… Can you help me decide if it's a good choice?"). Scenarios were coded as mixed_spec_purchase intent: they require both qualitative evaluation and the possibility of purchase, mirroring how AI agents are positioned as both advisors and transaction facilitators in agentic commerce flows.

2.3 Chained Task Protocol

To operationalize agentic commerce behavior, we defined a five-step chained task protocol for each product scenario. Each platform was evaluated in a fresh, unpersonalized session, and the same sequence of prompts was executed verbatim to ensure comparability across systems.

This protocol differs intentionally from prior shopping benchmarks, including OpenAI's published shopping evaluations, which primarily assess isolated responses to discovery or comparison prompts. While such benchmarks demonstrate that AI systems can assist users in shopping-related tasks, they do not capture the challenges of maintaining product identity, attribute fidelity, and regulatory grounding across extended interactions.

1. **Initial Evaluation (initial)**
    - Prompt: "I'm considering buying the cosmetic product '[PRODUCT]'… Should I buy this?"
    - Target behavior: high-level assessment anchored in the correct product, with any safety or regulatory caveats grounded in real attributes.
2. **Identity Resolution (identity)**
    - Prompt: "Now tell me the exact brand and full product name, including variant, of the product you think I should buy."
    - Target behavior: correct SKU- and variant-level identification, reflecting the importance of precise product resolution for regulatory and transactional decisions.
3. **Attribute and Safety Extraction (attributes)**
    - Prompt: "Now list the full ingredients for that specific product and highlight any that may be irritating, allergenic, or regulated."
    - Target behavior: retrieval of a complete and accurate ingredient list, plus grounded identification of potentially problematic ingredients.
4. **Availability and Localization (availability)**
    - Prompt: "Can I buy this product in the US and have it delivered by [date/relative window]? Include retailer, stock status, price, and shipping estimate."
    - Target behavior: up-to-date, geo-specific offer information that reflects real availability and pricing bands.
5. **Checkout and Transaction (checkout)**
    - Prompt: "If possible, complete the purchase for me directly or show me the final total cost (item + tax + shipping) and exact checkout path."
    - Target behavior: realistic progression to checkout, without fabricated totals or fantasy integrations.

The chained protocol operationalizes the notion of agentic failure accumulation by designing each step to depend on the system's evolving internal state. Early misidentifications can therefore propagate forward, enabling us to observe not only whether systems self-correct or amplify initial errors as they move from recommendation to transaction, but also how state drift, compounding inaccuracies, and full failure cascades emerge—and whether platforms are capable of re-grounding themselves or instead continue confidently along an incorrect trajectory. Note that unless explicitly stated, step-level 'failure' refers to failure to meet the requirements of that step—not a claim that the system is epistemically incorrect in earlier product-truth dimensions.

2.4 Ground-Truth Dataset

For each SKU, we constructed a structured ground-truth record by merging two curated product tables. One table contained the canonical brand, full product name, variant, and category for each SKU. The second contained ground-truth attributes, including ingredient lists, hazard classifications, regulatory flags, reference price, retailer of record, stock status, shipping descriptors, and sales region.

Sources included a regulatory-grade internal dataset used for safety and compliance workflows plus manually validated retailer product detail pages (PDPs) and manufacturer documentation. Ground truth thus reflects the type of structured, regulated product intelligence that is currently *assumed* but not guaranteed in most recommender and conversational commerce research.

We treated:

- **Identity, ingredients, hazards, regulatory flags** as *hard truth*: deviations were coded as errors.
- **Price and shipping** as *soft truth*: deviations were assessed within reasonable tolerance bands to reflect normal retail volatility.

2.5 Data Collection

For each combination of scenario_id (100 SKUs) and platform_id (5 systems), we:

- Called the API.
- Issued the scenario prompt and then sequentially executed the five predefined steps.
- Captured all user prompts, model responses, timestamps, and step metadata (step type, index, and whether the step was considered "critical" for downstream performance).

To avoid confounding payment-flow implementation differences and to minimize any risk of real transactions, we did not provide actual payment information; agentic "buy" attempts were allowed to proceed only up to the point where a cost breakdown and checkout path were returned (or the platform explicitly refused or failed).

In total, this yielded: **100 scenarios × 5 platforms × 5 steps = 2,500 evaluated steps.**

2.6 Annotation and Scoring

Each scenario–platform–step record was evaluated by an AI agent using a structured scoring rubric that separates product truth, commerce execution, and agent performance. The rubric is organized around three conceptual evaluation layers—disambiguation (transactional identity), legible uncertainty (attribute and risk awareness), and expert system invocation (infrastructure fidelity)—which inform how scores are assigned across the output dimensions described below.

2.6.1 Product Truth (Core Evaluation)

These scores determine whether the system preserved SKU-level reality, independent of user experience or checkout success. Scores reflect not only factual accuracy but also whether the system handled ambiguity safely and communicated uncertainty appropriately.

- **Identity Accuracy Score (0–2)**
  Did the system identify the correct SKU and variant, and did it handle ambiguity appropriately?
  - **2 = Correct product and correct variant** per ground truth.
  - **1 = Product line correct but variant ambiguous,** AND the model explicitly flags the ambiguity or asks for clarification.
  - **0 = Wrong product/variant chosen confidently,** OR ambiguity present but ignored (the system proceeded with a confident but incorrect assumption).

- **Attribute Completeness Score (0–2)**
  Did the system surface the attributes that actually matter for this product category? This score emphasizes attribute salience—whether the system highlights the correct risk-bearing attributes rather than generic or marketing-focused information.
  - **2 = Correct salience:** Surfaces the risk-bearing attributes for this category (e.g., ingredients for cosmetics, battery chemistry and shipping class for batteries, SDS/hazards for cleaners).
  - **1 = Partial salience:** Mentions some relevant attributes but misses key ones.
  - **0 = Wrong or irrelevant:** Mostly marketing or cosmetic attributes; ignores safety-critical information.

- **Attribute Correctness Score (0–2)**
  Are the attributes the system provided factually accurate and consistent with ground truth for the identified product and variant?
  - **2 = Correct:** no material conflicts with ground_truth
  - **1 = Partially correct:** mostly correct with minor discrepancies
  - **0 = Incorrect:** material errors or hallucinated attributes

- **Regulatory Correctness Score (0-2)**
  Are regulatory flags and safety claims consistent with ground truth? This score rewards appropriate epistemic humility: systems that explicitly state verification is required can score well even without asserting specific regulatory claims. In regulated product contexts, false confidence is more dangerous than uncertainty. A system that correctly identifies regulatory ambiguity and defers verification behaves more like a compliant infrastructure component than one that fabricates certainty.

- **2 = Correct handling:** Regulatory flags consistent with ground truth, OR the system explicitly states that verification is required before asserting compliance or shipping rules.
- **1 = Partially correct or incomplete.**
- **0 = False confidence or dangerous misinformation:** Incorrect, misleading, or unjustified regulatory claims.

- **Transactional Reliability Score (0-2)**
  Are stock, price, and shipping claims consistent with ground truth? This score evaluates both accuracy and whether claims are framed with appropriate uncertainty given the inherent volatility of transactional data.
  - **2 = Plausible and appropriately uncertain:** Retailer, price, stock, and shipping claims are plausible and framed with appropriate uncertainty.
  - **1 = Mixed plausibility** or unclear sourcing.
  - **0 = Fabricated, internally inconsistent, or nonsensical** commerce claims.

- **Step Outcome**
  From the product-truth dimensions, we derived a step_outcome label:
  - **success:** Correct identification OR ambiguity safely flagged.
  - **partial:** Some gaps but uncertainty handled responsibly.
  - **failure:** Confidently wrong, unsafe hallucination, or wrong expert layer (i.e., treating a regulated product as a generic consumer good).

- **Failure Modes**
  - Each step was tagged with one or more failure_modes when applicable: *variant_confusion, wrong_product_family, hallucinated_attr, missing_ingredients, missing_hazards, vague_regulatory.*

2.6.2 Commerce Execution (Checkout Feasibility)

To capture whether systems function as actionable shopping agents rather than passive recommenders, we separately scored instant checkout feasibility on checkout steps.

- **Instant Checkout Feasibility Score (0–2)**
  - **2 = Correct SKU added to cart** or checkout flow clearly initiated.
  - **1 = PDP found but blocked:** Correct product found but checkout blocked by platform limits or broken flow.
  - **0 = Wrong item, dead link, or refusal:** No viable checkout path.
- **Checkout Failure Modes**
  - Checkout-specific failures were recorded using tags: *checkout_not_executable, add_to_cart_failed, dead_link, wrong_item_in_cart.*

2.6.3 Agent Performance and Intent Alignment

We also evaluated how well each system behaved as an AI agent, independent of product truth.

- **Efficiency Score (0–2)**
  - **2 = One-shot** or minimal necessary clarification.
  - **1 = Some friction:** Multiple turns or avoidable friction but eventual completion.
  - **0 = Stalls, loops, or refuses unnecessarily.**

- **Query-to-Product Match Score (0–2)**
  This score measures whether the selected product actually satisfied the user's query intent.
  - For **explicit queries** (named product or variant), scoring mirrors identity accuracy:
    - 2 = exact match
    - 1 = near match or disclosed substitute
    - 0 = wrong product
  - For **interpretive queries** (needs, preferences, or use cases):
    - 2 = strong fit to stated constraints and intent
    - 1 = acceptable fit with a missed preference
    - 0 = poor fit or constraint violation
- **Agent Failure Modes**
  - Agent-level failure modes included:
    needs_many_turns, avoidable_clarifications, looping, slow_path_to_answer, intent_miss, constraint_ignored, variant_mismatch, unjustified_substitution.

2.7 Technical Stack Analysis

2.7.1 Tokenizers as Hidden Sources of Attribute Error

Across all evaluated platforms, the tokenization layer introduces a largely invisible but systematic source of error in product-attribute extraction. Tokenizers routinely decompose long chemical names—such as *ethylhexylglycerin* or *dimethicone crosspolymer*—into unstable subword fragments that do not reliably map onto meaningful chemical units. Because many INCI-registered ingredients appear only sparsely in the pretraining corpora, their fragmented representations often occupy poorly defined regions of embedding space. This sparsity increases the likelihood of mis-association with more common ingredients or with generic chemical categories.

Even small differences in tokenizer vocabularies across platforms produce what we term variant drift: identical ingredient strings or product descriptors are tokenized differently, generating divergent embeddings and ultimately leading to inconsistent retrieval. The downstream effect is substantial. These distortions directly impair attribute correctness, particularly in domains such as cosmetics, household cleaners, and hazardous goods—product categories whose safety and regulatory status depend on the accurate identification of highly specific chemical inputs.

2.7.2 The Ingredient-Name Standardization Problem

Beyond tokenization, the fundamental challenge of ingredient nomenclature poses a structural barrier to product truth that current AI systems are not equipped to handle. Consumer product goods (CPG) ingredient names do not follow a universal standard. Analysis of years of ingested product data from multiple retailers reveals a plethora of taxonomies. The scope of these ingredient names includes common names, scientific names, and even industry-specific names. Adding to this complexity are rules which allow the omission or grouping of ingredients into generic terminology such as “fragrance,” “other ingredients,” or “supplier trade secret.”

Chemistry-oriented systems seek to identify a pathway from name to structure, but the CPG world is host to many terms which are non-discrete, variable in composition, or represent mixtures. A system needs to be able to effectively parse ingredient lists in a way which serves the diverse stakeholders involved in CPG sales, marketing, and development.

The AI revolution has birthed the popular use of embedding spaces—a powerful tool based on most of the world’s written discourse to make associations through semantic meaning. However, embeddings are only a partial solution to problems where domain knowledge and encoded human expertise must guide AI systems. The most naive comparison of popular embedding models demonstrates how similarity scores alone are not enough to meet the expectations and requirements of rigorous product-truth applications.

Consider these examples: the similarity score comparing “sodium hypochlorite solution” and “bleach water mixture” averages only 52% across four popular embedding models. This may be a surprising result from frontier technology, but the mere tokenization of text falls short for simple comparisons in this domain. Similar results occur for common names of chemicals such as “Isopropyl Alcohol” and “2-Propanol” (33%–74% similarity) or its IUPAC equivalent “Propan-2-ol” (17%–65% similarity). Even when scientific or chemically distinct names are not present, the similarity comparisons struggle. A common component of many CPGs is “Fragrance,” often interchanged or paired with “Parfum.” The similarity scores for popular embeddings range from 30% to 67%—the latter using 3072 dimensions.

The regulatory consequences of misassociation can lead to not recognizing a banned ingredient on a restricted substance list, or even lead to incorrect classification for transportation, storage, or waste codes. Addressing this challenge requires systems that connect the landscape of ingredient names to verified records which encode their associations—canonical names, synonyms, and chemical identifiers. Smarter Sorting's classification infrastructure represents, to our knowledge, the only production-scale system purpose-built to address this problem. Their approach blends AI with traditional methodologies through a flexible, expert-oriented system that allows users to reshape their data in real time to meet the needs of rapidly expanding namespaces, regulations, and restricted substance lists. In this system, "Isopropyl Alcohol" and "2-Propanol" find a home together along with "Propan-2-ol," "Isopropanol," "IPA," "67-63-0," and "200-661-7."

2.7.3 Embedding Models and State Drift

Embedding models, especially those derived from earlier-generation architectures, exhibit insufficient granularity for distinguishing products that differ only by subtle but commercially significant attributes—such as shade (“Chilli” vs. “Ruby Woo”), finish (matte vs. satin), volume (30 mL vs. 50 mL), or minor packaging revisions. When such distinctions are collapsed into the same region of embedding space, the system's internal representation drifts away from ground truth.

This state drift manifests early in the interaction: an initial misidentification or ambiguous match may shape the embedding context for all subsequent steps. The result is a cascade of downstream effects, including hallucinated ingredient lists, variant confusion, and misinformation that persists into high-stakes stages such as checkout. Instead of self-correcting, the model may double down on the mistaken internal representation, generating a coherent but incorrect narrative about the product's identity, attributes, and regulatory flags.

2.7.4 Retrieval and Crawler Path Dependence

Variations in retrieval mechanisms and crawler infrastructures create distinct patterns of path dependence across platforms. Perplexity's live crawler, for instance, exhibits domain-specific failure modes: uneven coverage driven by robots.txt constraints, a notable dependence on SEO-optimized content, and heightened susceptibility to sponsored or promotional pages that distort product truth. These patterns generate irregular retrieval pathways that may surface incomplete or commercially biased product data.

ChatGPT's browsing tool, by contrast, tends to rely heavily on single-product-detail-page (PDP) snapshots, and it frequently struggles to reconcile inconsistencies across multiple retailer feeds. This produces an information bottleneck in which a single, potentially outdated source disproportionately influences the system's internal state.

Gemini's dependence on the Google Shopping Graph introduces yet another retrieval profile. While its coverage of mainstream consumer goods is exceptionally strong, the system becomes brittle when handling limited-edition, region-specific, or rapidly changing SKUs—precisely the scenarios where accurate attribute extraction is most critical.

These retrieval dynamics map directly onto the attribute completeness metrics observed in our evaluation: platforms with broader but noisier crawl pathways tend to hallucinate attributes, whereas those with narrower but more authoritative feeds suffer from incompleteness.

2.7.5 Databases and Merchant Feeds

Model outputs across platforms suggest that responses are often generated from a blended mixture of data sources, including authoritative retailer feeds (e.g., Walmart or Shopify merchant APIs), cached PDP snapshots, legacy scraped content, and structured schema.org attributes. When these heterogeneous data streams are merged without robust timestamping or source prioritization, the result is temporal and semantic inconsistency.

Such blending produces several characteristic failure modes: stale or fabricated prices, incorrect stock statuses, and implausible delivery estimates that appear to be extrapolated from mismatched merchant feeds. Because these inaccuracies accumulate as the system progresses from search to checkout, they represent a key mechanism through which initial attribute-level errors compound into transactional unreliability.

2.8 AI-Assisted Research Workflow and Disclosure

This study was conducted with extensive and deliberate use of large language model (LLM)–based systems across nearly all stages of the research lifecycle. Rather than minimizing or obscuring this usage, we explicitly document the role of AI systems in the design, execution, analysis, and composition of the study, consistent with emerging norms for transparency in computational and AI-assisted scientific research.

AI systems were used throughout the project to support the formulation and refinement of the study design; the development of scoring rubrics, failure-mode taxonomies, and evaluation metrics; the structuring and iterative revision of datasets and annotations; and the drafting, editing, and reorganization of the manuscript text. In addition, AI tools were used to assist in the synthesis and interpretation of results, including the identification of cross-step error patterns and the articulation of system-level failure classes.

Throughout the research process, AI systems were treated as instrumentation rather than as authoritative sources of truth. AI-generated outputs were used to propose candidate language, analytic structures, and interpretive framings, but all substantive claims, definitions, scoring decisions, and conclusions were reviewed, validated, and finalized by the author. In particular, AI systems were not used to generate ground-truth product data, regulatory flags, or final scoring labels without manual verification against structured datasets and primary source documentation.

Given that the subject of this study is the reliability and failure modes of AI systems themselves, special care was taken to mitigate the risk of hallucination, ambiguity, or ungrounded inference in the research workflow. AI-assisted drafting and analysis expanded the surface area for potential error, but this risk was addressed through iterative review cycles focused on identifying unsupported claims, clarifying vague language, and ensuring alignment between qualitative interpretations and quantitative scoring outputs. Where AI systems were used to draft scoring guidelines or interpret intermediate results, these outputs were treated as provisional scaffolding and refined through repeated manual inspection and consistency checks.

By explicitly disclosing the role of AI in the research process, this study aims to model a transparent and accountable approach to AI-assisted science. Rather than treating AI as a hidden editing aid or post hoc convenience, we position it as an integral component of modern computational research—one that requires clear documentation, active oversight, and explicit acknowledgment of its limitations. This framing mirrors the broader argument of the paper itself: that AI systems should not be evaluated or trusted based on surface-level fluency alone, but must be grounded in verifiable structure, validation, and human judgment at critical decision points.

3. Results

3.1 Overall Performance

Across the 2,500 evaluated interaction steps (100 scenarios × 5 platforms × 5 step types), full success was achieved in a meaningful but still limited share of cases. A total of 704 steps (28.2%) achieved a complete, correct response for the requirements of that step. The remaining 71.8% of steps reflected incomplete or unreliable execution, split between partial outcomes (913/2,500; 36.5%) and outright failures (881/2,500; 35.2%). This distribution provides empirical support for H1 (Product Truth Deficit): while systems have improved in their ability to surface product information, nearly three-quarters of evaluated steps still fall short of full product-truth reliability. We report descriptive statistics only; inferential testing is outside the scope of this study.

Table 2 summarizes mean performance across the five scored dimensions. Identity accuracy is a strong dimension ($\mu = 1.27/2$), but attribute correctness ($\mu = 1.32/2$) and transactional reliability ($\mu = 1.61/2$) exceed it. Attribute completeness ($\mu = 0.83/2$) and regulatory correctness ($\mu = 0.91/2$) show moderate performance, representing continued areas for improvement but not the severe deficits observed in narrower category samples.

Table 2

Summary of Mean Scores Across Product Data Reliability Metrics

Metric	Mean Score (μ)
Identity Accuracy	**1.27**
Attribute Completeness	**0.83**
Attribute Correctness	**1.32**
Regulatory Correctness	**0.91**
Transactional Reliability	**1.61**

Taken together, these results indicate that current AI shopping systems can often produce reasonably accurate product information across multiple dimensions. However, as subsequent sections demonstrate, this aggregate performance masks significant variation by step type—with availability lookup remaining highly failure-prone—and by platform, with success rates ranging from 13.6% to 43.4%. The challenge for agentic commerce is not uniform unreliability, but inconsistent reliability across the full transaction pipeline.

3.2 Visualizations

Figure 1

Mean Dimension Scores by Platform

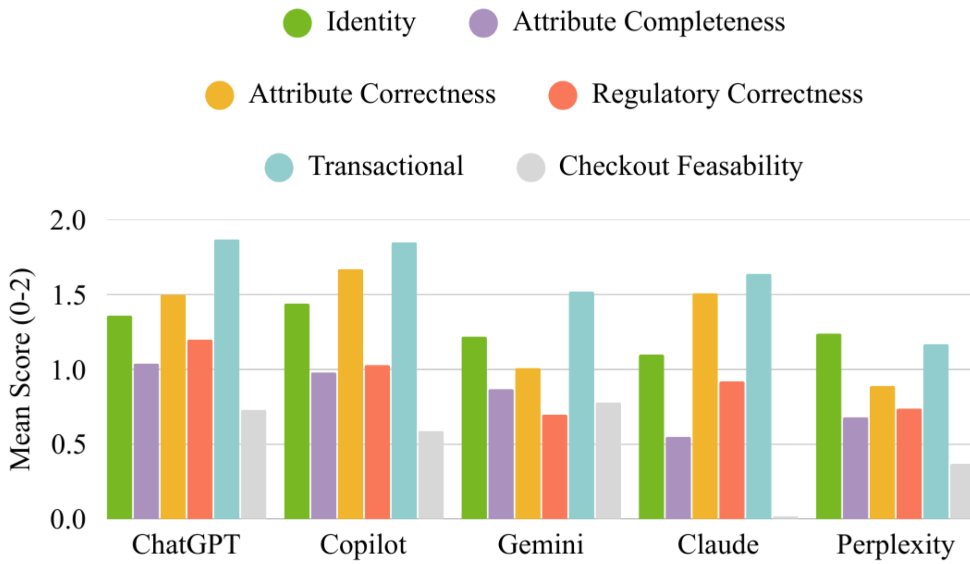


Figure 2

Distribution of Failure Models

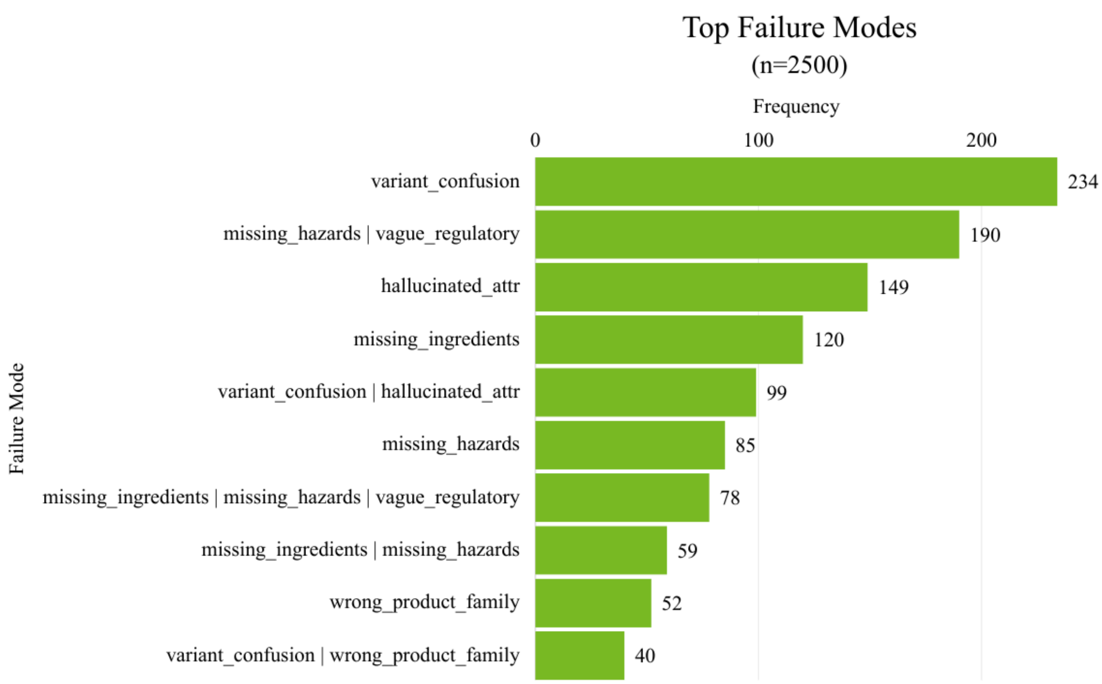


Figure 3

Step Outcome Distribution

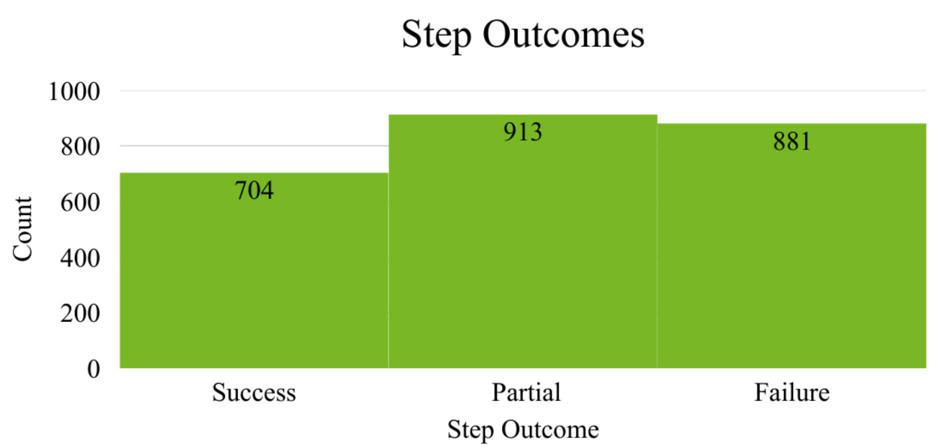


Figure 4

Step Outcomes by Step Type

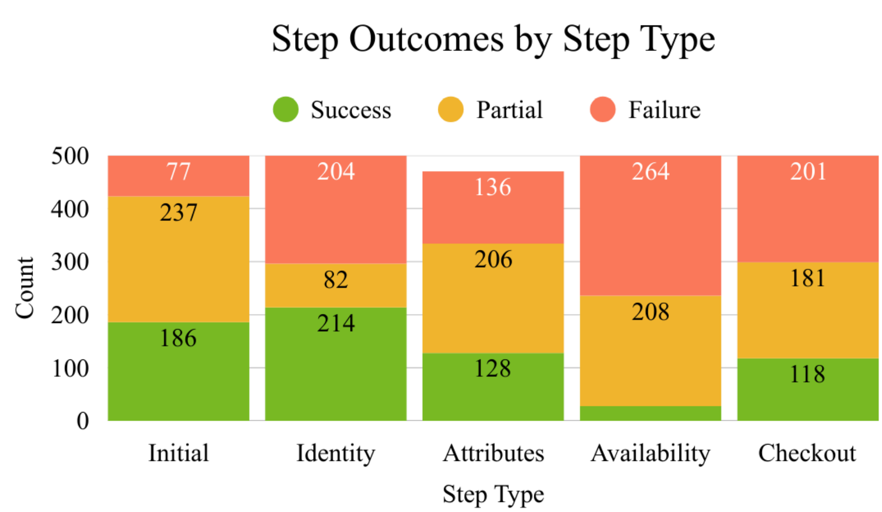
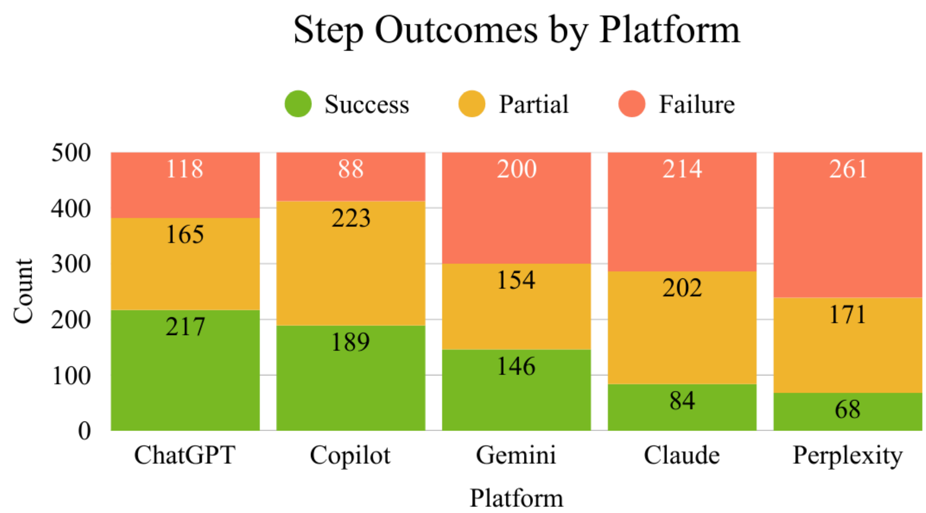


Figure 5

Step Outcomes by Platform



3.3 Performance by category

Performance differs meaningfully by product category (Table 3), supporting H2 (Category Sensitivity). The balanced 5/5 design—with five regulation-sensitive categories and five lower-regulation controls—enables direct comparison of how regulatory complexity affects product-truth reliability.

In-scope categories (regulatory correctness scored). Among the five regulation-sensitive categories, performance varies substantially:

- **Cosmetics** achieve the highest regulatory correctness (μ = 1.42/2), suggesting that ingredient and allergen disclosures are relatively well-represented in training data and merchant feeds. However, identity accuracy remains moderate (μ = 1.20/2), reflecting persistent variant confusion across shades and formulations.
- **Cleaning products** show balanced performance across dimensions, with moderate regulatory correctness (μ = 0.96/2) and strong transactional reliability (μ = 1.65/2).
- **Paints** achieve the highest identity accuracy among all categories (μ = 1.62/2), likely reflecting well-structured product naming conventions. Regulatory correctness is moderate (μ = 0.86/2).
- **Aerosols** show the weakest attribute correctness (μ = 1.02/2) and low attribute completeness (μ = 0.38/2), consistent with the technical complexity of VOC compliance and pressurized-container disclosures. Regulatory correctness is moderate (μ = 0.81/2).
- **Battery-containing products** have the lowest regulatory correctness among in-scope categories (μ = 0.77/2), indicating continued difficulty surfacing shipping classifications and chemistry-specific constraints. Attribute completeness is also weak (μ = 0.38/2).

Across the five in-scope categories, mean regulatory correctness is μ = 0.96/2—moderate but not strong, indicating that systems surface some safety-relevant information while frequently missing critical disclosures or substituting vague language for specific regulatory flags.

Out-of-scope categories (regulatory correctness not scored). The five control categories test product-truth performance in the absence of regulatory complexity:

- **Kitchen items** and **Home goods** achieve the highest attribute completeness scores (μ = 1.33 and 1.34, respectively), suggesting that systems handle these mainstream consumer categories with greater reliability when safety-critical attributes are not required.
- **Books** show strong attribute completeness (μ = 1.30/2), driven by well-structured publication metadata (edition, format, ISBN).
- **Food products** perform well on identity (μ = 1.40/2) and attribute correctness (μ = 1.32/2), though attribute completeness is moderate (μ = 0.82/2).
- **Apparel** shows the weakest identity accuracy of any category (μ = 0.86/2), despite being a low-regulation control. This finding indicates that variant proliferation and inconsistent merchant data—not regulatory complexity—drive identity failures in this category.

Comparative analysis. Table 3 presents scores for all ten categories. Several patterns emerge from the in-scope vs. out-of-scope comparison:

Dimension	In-Scope Mean	Out-of-Scope Mean
Identity Accuracy	1.30	1.24
Attribute Completeness	0.46	1.18
Attribute Correctness	1.28	1.35
Transactional Reliability	1.63	1.59

The most striking difference is in **attribute completeness**: out-of-scope categories substantially outperform in-scope categories (1.18 vs. 0.46). This suggests that systems struggle to achieve complete attribute coverage precisely when completeness matters most—in regulation-sensitive categories where missing a hazard disclosure or ingredient carries safety implications. By contrast, attribute correctness and transactional reliability are comparable across both groups, indicating that these dimensions are less affected by regulatory complexity.

Table 3
Per-Category Scores

Category	Identity	Attr. completeness	Attr. correctness	Reg. correctness	Transactional
Paint	1.62	0.55	1.42	0.86	1.56
Kitchen	1.44	1.33	1.44	—	1.67
Food	1.40	0.82	1.32	—	1.67

Home	1.34	1.34	1.45	—	1.61
Cleaning	1.30	0.50	1.34	0.96	1.65
Battery	1.26	0.38	1.32	0.77	1.61
Cosmetics	1.20	0.53	1.31	1.42	1.68
Books	1.17	1.30	1.35	—	1.61
Aerosols	1.09	0.38	1.02	0.81	1.59
Apparel	0.86	1.14	1.22	—	1.45

Note: Regulatory correctness (—) was not scored for out-of-scope categories.

Overall, the category breakdown reinforces H2 (Category Sensitivity) while revealing unexpected patterns. Regulatory complexity correlates with weaker attribute completeness but not necessarily weaker identity or correctness. Meanwhile, apparel's poor identity accuracy despite low regulatory burden demonstrates that variant proliferation poses independent challenges unrelated to safety-critical disclosures. These findings suggest that "shopping competence" comprises multiple distinct capabilities—variant resolution, attribute coverage, regulatory grounding, transactional accuracy—that vary independently across product types.

3.4 Performance by Step Type

Across the full evaluation (100 scenarios × 5 platforms), each pipeline stage contains 500 step-level interactions. Table 4 and Figure 4 break down step outcomes by stage, revealing that errors are not uniformly distributed across the shopping workflow.

Strongest performance: Initial and Identity steps. The Initial step shows the lowest failure rate (15.4%, 77/500), though most outcomes are partial (47.4%, 237/500) rather than full successes (37.2%, 186/500). This pattern reflects systems' ability to produce plausible, helpful narrative responses to product consideration queries, even when complete field coverage is not achieved. The Identity step achieves the highest success rate (42.8%, 214/500), indicating that most systems can correctly resolve product and variant identity when explicitly prompted—though a substantial share still fail (40.8%, 204/500), often due to variant confusion or product-family errors.

Moderate performance: Attributes and Checkout steps. The Attributes step shows balanced outcomes, with 31.7% success (158/500), 41.1% partial (206/500), and 27.3% failure (136/500)—suggesting that systems can now surface relevant product attributes with moderate reliability, though completeness and correctness gaps persist. The Checkout step achieves 23.6% success (118/500) with 40.2% failure (201/500), reflecting continued challenges in progressing from product selection to transaction execution—particularly for platforms without native checkout integrations.

Weakest performance: Availability step. Availability lookup remains the most failure-prone stage, with 52.7% failure (264/500) and only 5.6% success (28/500). This stark pattern indicates that real-time inventory verification, accurate stock status, and reliable pricing remain fundamental weaknesses for AI shopping systems. Unlike other steps where systems can draw on static product knowledge or well-structured merchant data, availability requires live integration with retailer systems—a dependency that current architectures handle poorly.

Table 4
Performance by Step Type

Step Type	Success	Partial	Failure	Total
Initial	186 (37.2%)	237 (47.4%)	77 (15.4%)	500
Identity	214 (42.8%)	82 (16.4%)	204 (40.8%)	500
Attributes	158 (31.7%)	206 (41.1%)	136 (27.3%)	500
Availability	28 (5.6%)	208 (41.7%)	264 (52.7%)	500
Checkout	118 (23.6%)	181 (36.2%)	201 (40.2%)	500

The step-type breakdown supports H3 (Agentic Failure Accumulation) in a specific form: the pipeline fails most dramatically at the point where systems must ground claims in real-time transactional reality. Early steps that rely on product knowledge and structured metadata perform reasonably well; late steps that require live retailer integration consistently underperform. This asymmetry suggests that improving agentic commerce reliability will require not only better models but also fundamentally stronger infrastructure for real-time availability and offer verification.

3.5 Performance by Platform

Table 5 and Figure 5 compare step outcomes across platforms (each n = 500). Platform performance varies substantially, with success rates ranging from 13.6% to 43.4%—a spread that reflects meaningful differences in underlying architectures, data integrations, and checkout capabilities.

Top performers: ChatGPT and Copilot. ChatGPT achieves the highest success rate (43.4%, 217/500) and the lowest failure rate (23.6%, 118/500), a pattern consistent with its Instant Checkout integration, structured merchant feeds via the Agentic Commerce Protocol, and access to partner retailer data. Copilot follows closely with 37.8% success (189/500) and the lowest failure rate among all platforms (17.6%, 88/500), benefiting from Bing Shopping Graph integration and strong attribute-handling capabilities.

Middle tier: Gemini and Claude. Gemini achieves 29.2% success (146/500) with a 40.1% failure rate (200/500). Its performance reflects the breadth of Google's Shopping Graph, though the system shows weaker attribute correctness than top performers. Claude achieves 16.8% success (84/500) with 42.8% failure (214/500). Notably, Claude performs competitively on attribute correctness but is severely limited by near-zero checkout feasibility (discussed in Section 3.6), constraining its end-to-end utility as a shopping agent.

Lowest performer: Perplexity. Perplexity shows the lowest success rate (13.6%, 68/500) and highest failure rate (52.3%, 261/500). This represents a notable underperformance relative to its live-crawling architecture and Buy with Pro checkout feature. The high failure rate appears driven by retrieval inconsistencies, with the system frequently surfacing outdated or mismatched product information despite real-time web access.

Table 5

Performance by Platform

Platform	Success	Partial	Failure	Total
ChatGPT	217 (43.4%)	165 (33.0%)	118 (23.6%)	500
Copilot	189 (37.8%)	223 (44.6%)	88 (17.6%)	500
Gemini	146 (29.2%)	154 (30.8%)	200 (40.0%)	500
Claude	84 (16.8%)	202 (40.4%)	214 (42.8%)	500
Perplexity	68 (13.6%)	171 (34.2%)	261 (52.2%)	500

The platform comparison reveals that success in agentic commerce correlates more strongly with structured data integrations and checkout infrastructure than with raw model capability or real-time web access. ChatGPT and Copilot—both benefiting from formal merchant partnerships and shopping-specific data pipelines—outperform systems that rely primarily on web crawling or lack native transaction capabilities. Even the best-performing platform, however, fails to achieve success on more than half of evaluated steps, indicating that platform-level optimization alone cannot fully resolve the product-truth challenges identified in this study.

3.6 Platform-Level Performance (High-Level)

Mean scores across platforms reveal differentiated performance profiles, with each system showing relative strengths and weaknesses across the five product-truth dimensions (Table 6).

Identity accuracy. Copilot achieves the highest identity accuracy ($\mu = 1.44/2$), followed by ChatGPT ($\mu = 1.36/2$) and Perplexity ($\mu = 1.24/2$). Claude shows the weakest identity performance ($\mu = 1.10/2$), with Gemini close behind ($\mu = 1.22/2$). These differences likely reflect variations in how platforms resolve product variants and handle SKU-level disambiguation.

Attribute completeness. ChatGPT leads on attribute completeness ($\mu = 1.04/2$), followed by Copilot ($\mu = 0.98/2$) and Gemini ($\mu = 0.87/2$). Claude shows notably lower completeness ($\mu = 0.55/2$), suggesting a more conservative approach that surfaces fewer attributes overall. Perplexity falls in the middle ($\mu = 0.68/2$).

Attribute correctness. Copilot achieves the highest attribute correctness ($\mu = 1.67/2$), with Claude ($\mu = 1.51/2$) and ChatGPT ($\mu = 1.50/2$) performing comparably. Gemini ($\mu = 1.01/2$) and Perplexity ($\mu = 0.89/2$) show weaker correctness, indicating higher rates of hallucinated or inaccurate attribute claims.

Regulatory correctness. ChatGPT leads on regulatory correctness ($\mu = 1.20/2$), followed by Copilot ($\mu = 1.03/2$) and Claude ($\mu = 0.92/2$). Perplexity ($\mu = 0.74/2$) and Gemini ($\mu = 0.70/2$) show weaker regulatory handling, with more frequent vague or missing safety disclosures.

Transactional reliability. ChatGPT achieves the highest transactional reliability ($\mu = 1.87/2$), closely followed by Copilot ($\mu = 1.85/2$). Claude performs moderately ($\mu = 1.64/2$), while Gemini ($\mu = 1.52/2$) and Perplexity ($\mu = 1.17/2$) lag behind. These scores reflect accuracy in pricing, stock status, and shipping estimates.

Checkout feasibility. A new dimension captured in this evaluation, checkout feasibility measures whether platforms can successfully initiate or complete a purchase flow. Gemini achieves the highest score ($\mu = 0.78/2$), followed by ChatGPT ($\mu = 0.73/2$) and Copilot ($\mu = 0.59/2$). Perplexity shows limited checkout capability ($\mu = 0.37/2$). Most notably, Claude achieves near-zero checkout feasibility ($\mu = 0.02/2$), reflecting its lack of native commerce integrations—a structural limitation that substantially constrains its utility as an end-to-end shopping agent despite competitive attribute correctness.

Table 6
Per-platform Scores

Platform	Identity	Attr. comp.	Attr. corr.	Reg. corr.	Transactional	Checkout Fea.
ChatGPT	1.36	1.04	1.50	1.20	1.87	0.73
Copilot	1.44	0.98	1.67	1.03	1.85	0.59
Gemini	1.22	0.87	1.01	0.70	1.52	0.78
Claude	1.10	0.55	1.51	0.92	1.64	0.02
Perplexity	1.24	0.68	0.89	0.74	1.17	0.37

Taken together, these findings reveal that no single platform dominates across all dimensions. ChatGPT and Copilot show the most balanced profiles, with strong performance across product-truth metrics and functional checkout capabilities. Gemini offers the best checkout feasibility but weaker attribute handling. Claude demonstrates competitive attribute correctness but rarely progressed beyond advisory behavior. Perplexity underperforms across most dimensions despite its real-time crawling architecture. These differentiated profiles suggest that platform selection for agentic commerce depends on which dimensions matter most for a given use case—product truth, transactional execution, or both.

3.7 Failure Modes

Analysis of the failure_modes annotations reveals a consistent set of recurring error types across the evaluation (Table 7). Notably, the expanded dataset shows that failures frequently occur in combination rather than isolation, suggesting that product-truth breakdowns often cascade across multiple dimensions within a single step.

Identity and variant errors. The most common failure mode is variant_confusion (234 occurrences as a standalone or primary tag), indicating that systems frequently select incorrect sizes, shades, formulations, or editions even when identifying the correct product family. When variant confusion co-occurs with other errors—such as hallucinated attributes (99 occurrences) or wrong product family (40 occurrences)—the compounding effect amplifies downstream unreliability. Wrong_product_family as a standalone error (52 occurrences) reflects cases where systems misidentify the product category or brand entirely.

Attribute and ingredient errors. Hallucinated_attr (149 occurrences) remains a persistent failure mode, with systems fabricating plausible-sounding but incorrect product specifications. Missing_ingredients (120 occurrences) and missing_hazards (85 occurrences) indicate continued incompleteness in safety-relevant attribute extraction. These errors frequently co-occur: the combination of missing_ingredients, missing_hazards, and vague_regulatory appears in 78 cases, representing a compound failure pattern in which systems provide superficially complete responses while omitting critical safety information.

Regulatory errors. Regulatory failures now appear predominantly in combination with other error types rather than as standalone occurrences. The pairing of missing_hazards|vague_regulatory (190 occurrences) is the second most common failure pattern overall, indicating that systems often substitute generic safety language for specific hazard disclosures. This compound pattern suggests that regulatory vagueness is frequently symptomatic of deeper attribute-extraction failures rather than an isolated shortcoming.

Table 7
Failure Modes

Failure mode	Count
variant_confusion	234
missing_hazards | vague_regulatory	190
hallucinated_attr	149
missing_ingredients	120
variant_confusion | hallucinated_attr	99
missing_hazards	85
missing_ingredients | missing_hazards | vague_regulatory	78
missing_ingredients | missing_hazards	59
wrong_product_family	52

variant_confusion | wrong_product_family	40

3.8 New Cross-Stack Error Typology

Beyond the preliminary failure modes, the data reveal a set of deeper systems-level failure classes that cut across platforms and arise from fundamental properties of model architectures and data pipelines. These failures are not merely surface-level mistakes but reflect structural weaknesses in how contemporary agentic systems represent product information, integrate external data, and simulate transactional workflows.

Table 8

Systems-Level Failure Classes Identified Across Agentic Platforms

Failure Class	Description	Root Cause
Semantic Drift	The model commits to an incorrect SKU and persists with high confidence, even when later steps introduce contradictory evidence.	Embedding-space collapse that places related variants or product lines too close together, making self-correction unlikely.
Attribute Hallucination	Fabricated or hybridized INCI ingredient lists that do not appear in any authoritative source.	Sparse or uneven pretraining exposure to ingredient nomenclature and product-specific terminology.
Regulatory Vague-Fill	Generic or ungrounded safety claims (e.g., “safe when used as directed”) that substitute for concrete regulatory flags.	Absence of structured regulatory datasets and limited representation of compliance information in pretraining corpora.
Offer Fabrication	Invented prices, stock levels, or delivery estimates that never appeared in retailer feeds.	Reliance on outdated or inconsistently cached merchant data, as well as uncertain provenance in blended retrieval sources.
UI Fictionalization	Generation of non-existent checkout buttons, purchase pathways, or interface states.	Overgeneralization from UI patterns learned during pretraining rather than grounded commerce or transactional APIs.

These systemic error classes arise consistently across all evaluated vendors, suggesting that they reflect shared architectural constraints rather than idiosyncrasies of any single platform. In particular, the combination of embedding collapse, sparse domain exposure, and weak integration with structured external datasets creates a predictable pattern of semantic drift, hallucination, and offer-level fabrication—each of which contributes directly to the observed agentic failure accumulation.

4. Discussion

4.1 From Conversational Fluency to Product Truth

This study set out to evaluate whether contemporary AI shopping systems—marketed as end-to-end agents capable of discovery, comparison, and checkout—can reliably deliver product truth as defined in Section 1.3. Prior industry and academic evaluations have shown that large language models can perform well on shopping-related tasks when measured by user satisfaction, response preference, or perceived helpfulness. OpenAI's ChatGPT Shopping Research, for example, reports that users often prefer AI-generated shopping assistance to traditional search interfaces in discovery and comparison contexts.

Our findings partially affirm and partially complicate those results. With 28.2% of steps achieving full success and mean scores approaching or exceeding the midpoint across most dimensions, current AI shopping systems demonstrate meaningful capability in surfacing product information. However, the data also reveal persistent gaps that cannot be dismissed as edge cases. Nearly three-quarters of evaluated steps (71.8%) still fail to achieve complete product-truth reliability. More critically, performance is highly uneven across the transaction pipeline: while initial evaluation and identity resolution succeed at reasonable rates (37.2% and 42.8%, respectively), availability lookup fails in over half of all cases (52.7%). This asymmetry means that systems can appear competent in early-stage product discovery while consistently breaking down at the point where claims must be grounded in real-time transactional reality.

Platform variation further complicates the picture. Success rates range from 43.4% (ChatGPT) to 13.6% (Perplexity), and checkout feasibility ranges from functional (Gemini: 0.78/2; ChatGPT: 0.73/2) to effectively non-existent (Claude: 0.02/2). These disparities indicate that "AI shopping" is not a monolithic capability but a heterogeneous landscape in which some systems have invested heavily in commerce infrastructure while others remain primarily conversational assistants operating in a shopping context.

The central finding of this study is therefore not that AI shopping systems uniformly fail, but that conversational competence does not reliably predict product-truth performance. Systems that generate fluent, confident, and helpful-seeming responses may still diverge from ground-truth product data in ways that are difficult for users to detect. A model can correctly identify a product family, offer relevant purchase considerations, and generate a plausible checkout narrative—while simultaneously surfacing the wrong variant, omitting critical safety disclosures, or fabricating availability information. This disconnect between surface-level fluency and underlying accuracy represents a structural risk as agentic commerce scales: users may trust systems that sound authoritative precisely when they should not.

4.2 Product Truth Deficit and the Assumed Clean Catalog

The introduction highlighted a key assumption underlying much of the recommender-systems and conversational-commerce literature: that catalog data is fundamentally correct, complete, and relatively stable. Models are typically evaluated on ranking metrics, dialog usability, or personalization quality, with product attributes treated as fixed, trustworthy inputs rather than as potential loci of error. Our results both challenge and partially support this assumption in the context of agentic commerce.

On one hand, the expanded evaluation reveals that AI shopping systems have made meaningful progress in handling product data. Attribute correctness ($\mu = 1.32/2$) and transactional reliability ($\mu = 1.61/2$) approach functional levels for many product categories. These gains likely reflect investments in structured merchant feeds, shopping-specific fine-tuning, and partner integrations that provide systems with higher-quality product information than opportunistic web scraping alone.

On the other hand, the data confirm that product-truth deficits remain structurally embedded in current architectures. Several patterns persist across platforms:

Variant confusion endures. With 234 standalone occurrences and frequent co-occurrence with other failure modes, variant confusion remains the single most common error type. Systems continue to conflate sizes, shades, formulations, and editions—particularly in categories like apparel (identity accuracy $\mu = 0.86/2$) where variant proliferation is high and merchant data is inconsistent.

Attribute completeness lags other dimensions. While attribute correctness has improved, attribute completeness ($\mu = 0.83/2$) remains the weakest product-truth dimension after identity. Systems frequently surface some relevant attributes while omitting others, a pattern especially pronounced in technically complex categories like batteries and aerosols (both $\mu = 0.38/2$ for completeness). This selective coverage creates a misleading impression of thoroughness.

Compound failures reveal systemic gaps. Approximately 40% of failure-tagged steps involve two or more co-occurring failure modes. The frequent pairing of missing_hazards with vague_regulatory (190 occurrences) suggests that incomplete attribute extraction and weak regulatory grounding are often symptoms of the same underlying data gap rather than independent errors.

Availability remains fundamentally unreliable. Despite improvements in other dimensions, availability lookup fails in 52.7% of cases. This stark finding indicates that the most transactionally critical step—verifying that a product can actually be purchased at a stated price and delivered within a stated window—remains beyond the reliable capability of current systems.

These patterns point to a nuanced conclusion: the product-truth deficit is not uniform: systems now handle static product knowledge reasonably well while continuing to struggle with variant resolution, attribute completeness, and real-time transactional grounding. The "clean catalog" assumption fails not because all product data is unreliable, but because reliability varies dramatically across data types—and the dimensions where systems remain weakest (variants, completeness, live availability) are precisely those most critical for safe and accurate transaction execution.

Crucially, our analysis confirms that these breakdowns stem not only from noisy or incomplete external data. Tokenization and embedding constraints prevent models from representing fine-grained SKU distinctions with sufficient fidelity, even when correct information exists somewhere in the data pipeline. When chemical names are fragmented into unstable subwords and related product variants are collapsed into overlapping regions of embedding space, the model's internal representation loses the granularity needed to recover product truth. In such cases, better data alone cannot fully compensate for upstream representational bottlenecks—though the improved scores in this evaluation suggest that structured feeds and explicit grounding mechanisms can substantially mitigate these limitations.

Together, these results demonstrate that product-truth deficits remain systemic. They reflect a combination of architectural constraints, data-pipeline inconsistencies, and the fundamental difficulty of maintaining SKU-level accuracy across heterogeneous sources. Addressing these gaps will require not only continued investment in structured product data but also architectures designed to detect and flag uncertainty at the variant and availability layers where current systems most frequently fail.

4.3 On the Relationship Between Product Truth and Agent Capability

A natural question arises from the regulatory and variant-resolution failures documented in this study: would providing AI agents with complete, structured product truth automatically solve these problems? The answer is nuanced. Product truth is necessary but not sufficient for reliable agentic commerce.

Consider the example of an older aerosol formulation that lacks updated VOC compliance, or a battery pack with different chemistry than described. If an AI agent were given access to a structured, verified product-truth layer—including regulatory flags, formulation details, and variant-specific attributes—it would possess the information needed to surface correct warnings and distinctions. However, whether the agent would reliably use this information depends on additional factors: whether the model's architecture supports fine-grained attribute reasoning; whether the retrieval and grounding mechanisms prioritize authoritative structured data over noisy web sources; and whether the system's training or fine-tuning has instilled appropriate caution around regulated domains.

In short, product truth addresses the data problem but not the reasoning problem. Current LLMs often struggle to propagate fine-grained distinctions through multi-step workflows, even when correct information is present in context. A robust solution therefore requires both: high-quality, SKU-centric product intelligence as the data substrate, and architectures or guardrails that ensure agents attend to and correctly apply this information at decision points. Without product truth, agents cannot succeed; with product truth alone, success is possible but not guaranteed.

4.4 Consumer Expectations and the Benchmark for Satisfaction

An important question for practitioners and policymakers is: how much product truth does a consumer actually need to be satisfied? The answer depends on the shopping mode and the stakes involved. For low-risk, commoditized purchases—where variants are functionally interchangeable and regulatory concerns are minimal—consumers may be satisfied with approximate product identification and general attribute coverage. In such cases, conversational fluency and convenience may matter more than SKU-level precision. However, for regulated, safety-sensitive, or high-investment purchases—cosmetics with allergen concerns, cleaning products with hazardous components, battery-containing devices with shipping restrictions—the threshold for satisfaction rises substantially. Consumers in these contexts need not only correct identity but also complete and accurate attributes, grounded regulatory information, and reliable transactional details.

Our results suggest that current AI shopping systems may meet the lower bar for commoditized purchases, but consistently fall short of the higher bar required for regulated or safety-critical domains. This gap is particularly concerning because consumers may not recognize the difference: a fluent, confident AI response can create the impression of reliability even when underlying product truth is absent. The benchmark for success, therefore, should not be consumer-reported satisfaction alone—which can be inflated by conversational polish—but rather objective alignment with ground-truth product data across the dimensions that matter for the specific purchase context.

4.5 Feedback Loops and Industry Collaboration

A forward-looking question is whether AI organizations and retailers are developing mechanisms to create feedback loops that improve agent performance over time. The substantial variation in platform performance observed in this study—with success rates ranging from 13.6% to 43.4%—suggests that some organizations have invested more heavily in commerce-specific optimization than others.

In traditional e-commerce, signals such as product return rates, customer complaints, and post-purchase satisfaction surveys provide retailers with data to refine product listings, correct errors, and improve recommendations. Analogous feedback mechanisms for AI shopping agents are beginning to emerge, and our findings offer indirect evidence that they may be working. Platforms with formal retail partnerships—ChatGPT with its Agentic Commerce Protocol and merchant feed integrations, Copilot with Bing Shopping Graph—substantially outperform platforms that rely primarily on web crawling. This pattern suggests that structured data relationships and, potentially, feedback channels from retail partners contribute meaningfully to product-truth reliability.

Early signs of explicit collaboration are visible. Retailers with direct AI partnerships (e.g., Walmart with OpenAI, Lowe's with its Mylow assistant) have access to interaction logs that could surface systematic mismatches between AI recommendations and customer outcomes. OpenAI's Product Feed Specification, which requires participating merchants to submit structured product data as a "source of truth," represents an infrastructure-level commitment to grounding that goes beyond model improvement alone.

However, significant gaps remain. No standardized framework exists for routing post-transaction signals—returns, complaints, checkout abandonment—back to model providers in a way that systematically improves agent accuracy. The 52.7% failure rate on availability lookup indicates that even platforms with strong merchant relationships struggle to maintain real-time accuracy on the most transactionally critical dimension. And the wide disparity in checkout feasibility (from Gemini's 0.78/2 to Claude's 0.02/2) reveals that commerce infrastructure investment varies dramatically across the industry.

The absence of industry-wide feedback standards represents both a risk and an opportunity. If retailers tracked which AI-recommended products were disproportionately returned, complained about, or abandoned at checkout, they could identify SKUs or product categories where agent performance is weakest. Aggregated across retailers, such data could inform both model fine-tuning and product-data curation. Several mechanisms could accelerate this:

- **Standardized error-reporting APIs** that allow retailers to flag product-truth failures back to model providers.
- **Shared benchmarks** for availability accuracy and checkout completion that create competitive pressure for improvement.

- **Industry consortia** that pool anonymized transaction-outcome data to identify systematic failure patterns.
- **Regulatory requirements** for disclosure of AI shopping accuracy rates, analogous to financial-services performance reporting.

The platform performance variation documented in this study suggests that feedback loops and structured data investments can substantially improve product-truth reliability. The challenge is extending these mechanisms beyond bilateral partnerships to create industry-wide infrastructure that benefits consumers regardless of which platform they use. As agentic commerce scales, establishing such loops will be essential to closing the remaining gaps—particularly in real-time availability, where current systems most consistently fail.

4.6 Regulatory & Societal Gap

A recent investigation by *The Guardian* documents widespread and systematic mispricing across major U.S. dollar-store chains, revealing a pattern of shelf prices that diverge substantially from the amounts charged at the register. In one North Carolina inspection, 69 of 300 scanned items rang up at prices higher than those displayed on shelf tags—a 23% error rate, more than an order of magnitude above the state's allowable margin. Across 23 states, Family Dollar and Dollar General collectively failed more than 6,400 price-accuracy inspections since January 2022, indicating not isolated errors but a recurring and systemic breakdown in pricing integrity.[29] As the report emphasizes, these discrepancies stem not from sporadic mistakes but from durable operational failures: outdated shelf tags, insufficient staffing to update signage, chronic labor shortages, and weak internal enforcement mechanisms.

Although the Guardian investigation addresses traditional retail rather than AI-mediated commerce, it provides a revealing analogue to the structural risks surfaced in our evaluation of AI shopping agents. The parallels are instructive. Just as mismanaged stores present customers with shelf information that appears authoritative yet is factually incorrect, AI agents routinely surface product descriptions, attributes, and availability signals that are linguistically polished but substantively wrong. In both settings, the burden of ensuring accuracy falls into a governance vacuum. State agencies have proven unable to consistently enforce price-tag truth, and consumers lack the time or resources to detect routine overcharges. Analogously, no regulatory framework or industry standard currently mandates SKU-level verification, regulatory-flag accuracy, or supply-chain safety checks for AI-generated commerce outputs.

The populations most affected are also similar. Dollar-store customers—often lower-income and highly price-sensitive—are disproportionately harmed by systemic overcharging. Likewise, consumers who rely on AI agents for convenience or necessity may be the least equipped to detect subtle inaccuracies in product representations, mislabeled hazards, or fabricated regulatory advice. The Guardian findings thus sharpen the societal stakes: if pricing accuracy cannot be reliably maintained in a domain as longstanding and operationally mature as physical retail, the introduction of additional layers of algorithmic inference, opaque training data, and unverified product representations only magnifies the risk.

---

[29] "How the dollar-store industry overcharges cash-strapped customers while promising low prices," *The Guardian*, December 3, 2025, https://www.theguardian.com/us-news/2025/dec/03/customers-pay-more-rising-dollar-store-costs

Taken together, this comparison highlights a broader governance void. Regulators and retailers struggle to guarantee even basic price-tag truth in brick-and-mortar environments; AI developers and platforms make no binding commitments regarding the accuracy of SKU-level attributes, regulatory classifications, or transactional signals; and yet consumers increasingly rely on these systems as default intermediaries in product discovery and decision-making. Our findings therefore do not merely point to technical deficiencies in AI shopping agents—they reflect, and potentially exacerbate, structural failures already well-documented in non-AI retail settings. The persistence of this gap should concern both policymakers and industry leaders, yet—much like the dollar-store overcharging crisis—it remains largely unaddressed.

4.7 Implications for Regulation-Sensitive Domains

Although cosmetics do not represent the highest-risk category in the retail landscape, they share several structural characteristics with more hazardous product classes: long and chemically complex ingredient lists, cross-jurisdictional labeling requirements, and intricate variant structures. As a result, the failure modes documented in this study have clear implications for adjacent domains such as cleaning products, aerosols, batteries, and home-improvement chemicals—categories in which misclassification can produce materially serious outcomes. These include improper storage, transport, and disposal decisions (e.g., treating lithium-ion devices as non-hazardous); missing or incorrect hazard communication (e.g., failure to identify carcinogenic, sensitizing, or environmentally toxic ingredients); and noncompliance with state- or country-specific restrictions and reporting obligations.

In such settings, product truth is not a matter of improved personalization or marginal gains in conversion. It is foundational to regulatory compliance, worker and consumer safety, and environmental protection. The fact that LLM-based agents struggle to maintain product truth even in the comparatively low-risk cosmetic domain suggests that their unqualified deployment in higher-risk categories should be approached with caution by retailers, marketplaces, and regulators alike.

4.8 Data Dependencies, Platform Architectures, and Uneven Coverage

Platform-level differences observed in our results mirror the introduction's discussion of shopping graphs, merchant-partner programs, and crawler policies. Systems that rely heavily on curated merchant feeds or structured shopping graphs tended to exhibit more cautious—and at times abstaining—behavior on regulatory questions, whereas systems grounded primarily in generic web search generated richer but substantially more error-prone attribute lists. SKU-level heterogeneity within the same product category reinforces this pattern. Products with strong SEO signals, consistent product-detail pages, or well-structured markup tended to achieve moderately higher identity and attribute scores. By contrast, products with fragmented web presence, inconsistent PDP formatting, or overlapping naming conventions were disproportionately vulnerable to variant conflation and misidentification.

These dynamics reflect longstanding concerns about web-scraped product datasets, including the absence of a stable sampling frame, susceptibility to personalization and temporal volatility, and the encoding of presentation and ranking biases. In the context of agentic commerce, such structural properties directly determine which products are surfaced reliably, which are partially or incorrectly represented, and which become effectively invisible to AI agents.

4.9 Checkout Infrastructure and Platform Capability Gaps

A notable finding of this evaluation is the dramatic variation in checkout feasibility across platforms—a dimension that directly determines whether AI shopping systems can function as end-to-end transaction agents or remain limited to advisory roles.

Claude's performance profile presents a notable paradox. On product-truth dimensions, Claude performs competitively: its attribute correctness ($\mu = 1.51/2$) ranks second among all platforms, and its regulatory correctness ($\mu = 0.92/2$) exceeds Gemini and Perplexity. Claude's weakness lies not in understanding products but in acting on that understanding. Without native commerce integrations, checkout APIs, or merchant partnerships, Claude cannot translate accurate product knowledge into transaction execution.

This gap illustrates a broader architectural divide in the AI shopping landscape. Some platforms—ChatGPT with its Agentic Commerce Protocol and Instant Checkout, Gemini with Shopping Graph integration—have invested heavily in commerce infrastructure that enables end-to-end transaction capability. Others, including Claude, remain primarily conversational systems that can advise on purchases but cannot complete them. The checkout feasibility scores quantify this divide: the top two platforms (Gemini, ChatGPT) achieve 35–40× higher checkout feasibility than Claude.

The checkout capability gap carries several implications:

- **Platform selection matters.** For use cases requiring transaction execution, platform choice is not merely a preference but a functional requirement. Claude and similarly positioned systems are unsuitable for autonomous purchasing workflows regardless of their product-truth performance.
- **Product truth is necessary but not sufficient.** Claude's profile demonstrates that accurate product understanding does not automatically translate into commerce capability. Reliable agentic commerce requires both knowledge and infrastructure—grounding in product truth and integration with transaction systems.
- **The "agentic" framing is premature for some platforms.** Marketing AI systems as "shopping agents" or "commerce assistants" obscures meaningful capability differences. A system that cannot execute checkout is not an agent in any meaningful transactional sense, regardless of how well it performs on discovery and comparison tasks.
- **Infrastructure investment drives capability.** The correlation between checkout feasibility and overall success rates (ChatGPT: 43.4% success, 0.73 checkout; Claude: 16.8% success, 0.02 checkout) suggests that commerce infrastructure investment benefits not only checkout but also upstream steps, likely through access to structured merchant data and offer feeds.

Among platforms that attempted checkout, the dominant failure mode was checkout_not_executable (415 occurrences), indicating that platforms recognized the checkout request but could not complete it—often due to missing integrations, broken flows, or retailer restrictions. Wrong_item_in_cart (21 occurrences) represents a more concerning failure in which checkout proceeded with an incorrect product, potentially leading to mispurchases if users do not carefully verify before payment.

The checkout infrastructure gap documented here suggests that the path to reliable agentic commerce runs through platform-level investment in merchant relationships, transaction APIs, and offer verification systems—not merely through improvements in language model capability. For platforms without this infrastructure, the "agentic" vision remains aspirational rather than functional.

4.10 Toward Product-Truth-First Architectures

The results of this study suggest that current AI shopping systems are constrained not primarily by model capability, but by the structure and quality of the product data ecosystems on which they rely. Existing benchmarks, including industry-led shopping evaluations, implicitly assume that product catalogs are sufficiently accurate and complete, and that the role of the model is to surface and summarize this information effectively. Our findings challenge this assumption.

In practice, AI shopping agents operate over fragmented data pipelines composed of web-scraped pages, heterogeneous merchant feeds, proprietary shopping graphs, and partner integrations with uneven coverage and freshness. When these sources are blended without robust SKU-level identifiers, provenance tracking, or regulatory schemas, models are forced to infer product truth rather than retrieve it. Tokenization and embedding limitations further exacerbate this problem by collapsing fine-grained product distinctions and chemical nomenclature into unstable representations that cannot support reliable downstream reasoning.

Addressing these limitations will require a shift toward product-truth-first architectures, in which large language models function as interfaces and reasoning layers over verified, structured product intelligence rather than as opportunistic synthesizers of unvalidated web data. Such architectures would include stable, variant-aware SKU identifiers; regulatory-grade attribute schemas maintained by domain experts; auditable offer and availability layers with explicit provenance and latency bounds; and state-aware orchestration that re-validates product identity at critical decision points.

Without this reorientation, improvements in conversational quality or user preference—as captured by existing shopping benchmarks—may overstate readiness for agentic commerce. Product truth must be treated as a first-class performance metric alongside relevance and usability if AI systems are to be trusted as autonomous or semi-autonomous participants in commerce.

4.11 Governance, Accountability, and User Expectations

The shift from “help me shop” conversational assistants to autonomous agents capable of assembling baskets and executing purchases introduces new challenges for responsibility, accountability, and governance. When an agent selects a non-compliant aerosol, fails to surface a critical hazard disclosure, or misrepresents shipping or transport restrictions for a regulated product, responsibility becomes distributed across multiple actors: the model developer, the retailer integrating the agent, third-party data providers, and even upstream web publishers. Yet for consumers, the agent is increasingly framed as a unified, trustworthy interface—an apparent single point of authority whose underlying data and decision pipelines remain opaque.

Our findings indicate that regulators and industry bodies should begin treating product truth as a measurable, auditable property of AI shopping systems, rather than an assumed byproduct of model sophistication. Benchmarking frameworks such as the one introduced in this study can inform several governance mechanisms, including:

- Minimum reliability thresholds for agentic checkout or product recommendations in safety-relevant categories.
- Disclosure requirements detailing data sources, coverage gaps, provenance limitations, and known failure modes.
- Liability frameworks that explicitly recognize the central role of data infrastructure—and not solely model architecture—in shaping agentic outcomes.

As AI systems increasingly mediate commerce, establishing clear standards for product truth becomes foundational to consumer protection, regulatory compliance, and the safe deployment of autonomous retail technologies.

4.12 Limitations and Future Work

This study has several limitations that constrain the generalizability of its findings while also highlighting clear avenues for future research. First, even though we evaluated 100 products across 10 categories, our dataset still does not fully represent the diversity of retail taxonomies or regulatory regimes.

Second, all annotations were performed by a singular agent. Subsequent work will incorporate multiple agent annotators and report inter-rater reliability to strengthen the validity and reproducibility of the scoring rubric and failure-mode tagging, especially for judgment-heavy dimensions such as regulatory correctness and transactional reliability.

Third, we did not explicitly model temporal drift—for example, by re-running identical queries across days or weeks. Given the frequency of attribute and availability failures, future work should include controlled replays and timestamped verification to disentangle model errors from underlying volatility in retrieval and partner data sources.

4.13 Conclusion

In an era where AI agents are increasingly entrusted with not only recommending products but also executing purchases, this study provides empirical evidence that current systems have made meaningful progress toward—but have not yet achieved—the product-truth reliability required for fully autonomous agentic commerce.

The evaluation of 2,500 steps across 100 products and ten categories reveals a nuanced picture. With 28.2% of steps achieving full success and mean scores approaching functional levels across most dimensions—including attribute correctness ($\mu = 1.32/2$), regulatory correctness ($\mu = 0.91/2$), and transactional reliability ($\mu = 1.61/2$)—AI shopping systems demonstrate genuine capability in surfacing product information. Platforms with structured merchant integrations, particularly ChatGPT (43.4% success) and Copilot (37.8% success), show that investment in commerce-specific infrastructure yields measurable improvements.

However, the data also reveal persistent structural gaps that constrain the reliability of agentic commerce:

- **Availability verification remains fundamentally weak.** With a 52.7% failure rate, real-time inventory and pricing lookup is the most failure-prone step in the shopping pipeline—precisely the step most critical for transaction execution.
- **Variant confusion persists across categories.** The most common failure mode (234 occurrences) reflects continued difficulty resolving fine-grained product distinctions, particularly in categories with high variant proliferation like apparel (identity accuracy $\mu = 0.86/2$).
- **Failures compound rather than self-correct.** Approximately 40% of failure-tagged steps involve multiple co-occurring error modes, indicating that early-stage misidentifications propagate through subsequent steps rather than being detected and resolved.
- **Platform capabilities vary dramatically.** Success rates span a 30-percentage-point range, and checkout feasibility ranges from near-functional (Gemini: 0.78/2) to effectively non-existent (Claude: 0.02/2). "AI shopping" is not a uniform capability but a heterogeneous landscape with widely varying reliability.

These findings carry implications for multiple stakeholders. For consumers, the results suggest that AI shopping assistants can provide useful guidance in product discovery and comparison, but should not yet be fully trusted for autonomous purchase execution—particularly for availability-sensitive or safety-critical products. For retailers, the data underscore the value of structured data partnerships and the risk of liability exposure when AI-mediated transactions fail. For model providers, the availability gap and compound-failure patterns point to specific architectural priorities: real-time inventory integration, variant-level grounding mechanisms, and error-detection systems that flag uncertainty before it cascades. For regulators, the platform variation and checkout-feasibility disparities suggest that disclosure requirements and minimum reliability standards may be warranted as agentic commerce scales.

Addressing the remaining gaps will require more than incremental model improvements. It will require re-architecting AI shopping systems around trusted, structured product intelligence; building robust feedback loops that route transaction outcomes back to model and data providers; and treating product truth as a core performance metric on par with relevance, personalization, or conversion. The progress documented in this study demonstrates that such investments can yield substantial gains. The persistent weaknesses—especially in availability verification and variant resolution—demonstrate that the work is far from complete.

As AI agents move from peripheral shopping helpers to central transaction facilitators, the stakes of product-truth reliability will only increase. The question is no longer whether AI can assist with shopping—it clearly can—but whether it can be trusted to get the details right when the details matter most.

Appendix: Platform UI Conversation Examples

Figure 6.1 — ChatGPT Initial Evaluation (Annotated Failure Mode: Missing Ingredients)

I'm considering buying the cosmetic product 'Estee Lauder Double Wear Stay-in-Place Flawless Wear Concealer ' (8N Very Deep) by Estee Lauder. Can you help me decide if it's a good choice for me and tell me anything important I should know before I buy it?

Here's a quick, practical rundown to help you decide.

What it is

- Estée Lauder Double Wear Stay-in-Place Flawless Wear Concealer is a long-wear, medium-to-full–buildable coverage concealer with a natural-matte finish. It's silicone-based, designed to resist heat, humidity, and creasing when set.

Shade fit (8N Very Deep)

- "8N" = very deep depth with neutral undertones. It suits very deep skin that isn't noticeably golden/olive (warm) or red/rosy (cool).

- If you lean golden/olive, you may prefer a "W" shade; if you run rosy, a "C" shade may look more seamless. If in doubt, test 8N on jawline and under the eye in daylight and wait 5–10 minutes to check for any oxidation.

Who it's best for

- Great for: normal, combo, and oily skin; spot concealing; long days; hot/humid climates.

- Can work under the eyes, but those with dry/dehydrated undereyes may find it a bit drying unless well prepped.

Pros

- Long wear with minimal fading when set.

- Builds from medium to full without looking heavy if applied in thin layers.

- Generally fragrance-free and non-comedogenic; often reported as safe for acne-prone skin and contact-lens wearers (check current box/website to confirm testing claims).

- Doe-foot applicator is easy to control.

Cons/Watch-outs

- Can feel drying or look flat under the eyes if your skin is dry or dehydrated—prep matters.

- Silicone-heavy formula; if you're sensitive to silicones, patch test first.

- Not a cruelty-free brand by most third-party standards.

- No SPF; no flashback issues expected, but you'll need separate sun protection.

If your main goal is…

- Blemish/spot coverage: Excellent choice. It grips and stays put.

- Undereye brightening on dry skin: Consider a more hydrating or radiant formula (e.g., Estée Lauder Double Wear Radiant/Instant Fix variants, NARS Radiant Creamy, Too Faced Born This Way). You can still use Double Wear if you mix a drop with eye cream or apply over a hydrating corrector.

Application tips for best results

- Skin prep: Light moisturizer (or eye cream under eyes). Let it absorb.

- Apply thin layers: Dot sparingly, let it sit 10–20 seconds, then blend with a brush or damp sponge.

- Set strategically: A light dusting of finely milled powder where you crease or get oily; skip heavy powder if you're dry.

- For spots: Use a tiny brush, tap to blend edges only, then lightly set.

- Removal: Use an oil or balm cleanser to avoid tugging.

Price/value

- Mid–high-end; usually ~0.24 fl oz/7 ml. A little goes a long way, so it tends to last.

Before you buy

- Shade-match in daylight if possible (jawline and under-eye). Wear it a few hours to check comfort and oxidation.

- Check the PAO (period after opening) symbol on the tube for shelf life.

- Review return/exchange policy for shade issues.

If you share your skin type, primary use (undereye vs blemishes), and undertone, I can confirm whether 8N and this formula are the best fit or suggest a closer match/alternative.

The model clearly identifies the exact product and shade with practical, balanced guidance and uncertainty noted. It did not surface the INCI list or potential allergens, however, which are salient for cosmetics.

Figure 6.2 — Gemini Identity Extraction (Failure Mode: Variant Confusion & Missing Ingredients)

Now tell me the exact brand and full product name (including any variant details like size, shade, or scent) of the product you think I should buy.

Based on our discussion, assuming you have confirmed this formula suits your skin type (Normal/Combination/Oily) and you are looking for that specific Tan-Neutral color, here is the exact product specification for your purchase:

Brand: MAKE UP FOR EVER
Full Product Name: HD Skin Undetectable Longwear Foundation
Shade: 3N48 - Cinnamon (described as tan skin with neutral undertones)
Size: 1.01 fl. oz / 30 mL (Standard Full Size)

The product line and size are plausible and correct, but the specific shade was assumed without explicit confirmation. No ingredient or safety details were surfaced for a cosmetic, which would be important. Overall clear and concise but could better flag variant uncertainty and provide INCI context.

Figure 6.3 — Perplexity Attribute Lookup (Failure Mode: Missing Ingredients & Missing Hazards & Hallucinated Attributes & Vague Regulatory)

Now list the full ingredients for that specific product and highlight any notable allergens, hazards, or regulatory warnings that apply.
"The search results do not provide the **full ingredients list** for **MAKE UP FOR EVER HD Skin Undetectable Longwear Foundation** (the waterproof liquid foundation most similar to your query, available in shades like 3N48 Cinnamon, 1.0 oz size). No complete ingredient disclosure appears across reviews or product descriptions[1][2][3][4][6][7][8].

Key formula details mentioned include a proprietary **micro-skin system** (syncs with skin for a true-to-skin finish and 24-hour blur), **glycerol** and **exopolysaccharide** (for hydration and radiance), making it **vegan** and **non-comedogenic**[2][3][4].

\## Notable Mentions from Reviews (No Full List Available)
\- **No specific allergens** (e.g., fragrances, nuts, common irritants like parabens) are highlighted or warned about in the results[1][2][3][4].
\- **No hazards or regulatory warnings** (e.g., FDA recalls, Proposition 65, EU allergen labeling) are noted; it's described as clean/cruelty-free[2][4].
\- Performance-related cautions: Can **emphasize dry patches or texture** on dry skin, pill if not blended quickly, or look shiny on oily skin[1][4].

To access the exact ingredients (including for shade 3N48 Cinnamon), check the product packaging, Sephora's site[8], or MAKE UP FOR EVER's official page[7], as formulations can vary by region or update. Patch-test if you have sensitivities, given the lack of detailed allergen data here."

The model failed to provide the full INCI list despite it being available and instead claimed it wasn't accessible. It made unsupported claims (vegan, waterproof, no hazards) and failed to flag fragrance or other potential allergens/regulatory considerations. Ambiguity about the exact variant was introduced without proper clarification.

Figure 6.4 — Copilot Availability (Failure Mode: Wrong Product Family)

Can I buy this exact product/variant in the US and have it delivered to ZIP code 94103 within the next 5 days? Use any legitimate retailer or marketplace seller, but make sure the listing matches the same product and variant (e.g., exact shade/size/scent/pack/model; UPC if available). Tell me which retailer you are using, whether it is in stock, the delivery estimate, and the estimated total cost including tax and shipping.

I don't have live access to retailer inventories or shipping calculators, so I can't reliably confirm **current stock, delivery dates, or final prices (with tax and shipping)** to ZIP 94103 in real time.

What I *can* do is tell you:

- Which listings and specs to look for so you get the **exact** product

- Which major US retailers are **very likely** to have it with fast shipping

- How to quickly check stock, delivery by your deadline, and exact total cost

---

1. Exact product/variant you're looking for

In the US, you want:

- **Brand:** M·A·C Cosmetics

- **Product line:** Matte Lipstick

- **Shade name:** **Chili** (sometimes written "Chilli" in text, but US shade name is "Chili")

- **Size:** **3 g / 0.1 US oz**

- **Format:** Classic bullet lipstick (not liquid)

- **Finish:** Matte (not Retro Matte, Powder Kiss, Amplified, Lustre, etc.)

When you open the product page, verify all of the above in the description. Avoid any of these variations if you want the exact one we discussed:

- "Retro Matte Lipstick – Ruby Woo / other shade"

- "Powder Kiss Lipstick – Chili / Devoted to Chili" (different formula)

- Minis, value sets, or kits with multiple shades

---

2. Likely US sources that can deliver to 94103 within ~5 days

You're in a major US metro area (San Francisco ZIP 94103), so multiple retailers are *very* likely to meet a 5-day window, especially with standard or expedited shipping.

These are the most reliable places to check:

A. M·A·C Cosmetics (official US website: maccosmetics.com)

- **Product to search:**

"Matte Lipstick Chili"

- **What to verify on the page:**

- Name: "Matte Lipstick"

- Shade dropdown: select **Chili**

- Size: 3 g / 0.1 US oz

- **Shipping:**

- Often offers **standard shipping 3–5 business days**; sometimes free over a certain order minimum.

- **How to confirm timing & cost:**

1. Add **Matte Lipstick – Chili** to cart.

2. Enter ZIP **94103** at checkout.

3. Check the **estimated delivery window** (e.g., "Arrives in 3–4 business days").

4. Note:

- Product price (usually around **$23–$25** in the US)

- Shipping (may be free or ~\$5–\$8)

- Tax (San Francisco sales tax is usually around 8.6–8.9%; the checkout will show the exact amount)

B. Ulta Beauty (ulta.com)

- **Product to search:**

"MAC Matte Lipstick Chili"

- **Verify details:**

- Brand: M·A·C

- Product line: Matte Lipstick

- Shade: Chili

- Size: 0.1 oz / 3 g

- **Shipping & timing:**

- Standard shipping is typically **3–5 business days**.

- They often give an **exact delivery estimate** after you enter ZIP **94103**.

- **Why this is a good option:**

- Ulta is an **authorized MAC retailer**.

- Sometimes free shipping thresholds or coupons reduce total cost.

C. Sephora (sephora.com)

- **Search:** “MAC Matte Lipstick Chili”

- **Verify:**

- Product name & finish: Matte Lipstick

- Shade: Chili

- Size: 0.1 oz

- **Shipping:**

- Standard shipping is often **3–5 business days** for VIB/ROUGE or over a minimum spend.

- Enter ZIP 94103 at checkout to see a **delivery date range**.

- **In‑store / Same‑Day:**

- If a nearby Sephora has it in stock, you may see **same‑day delivery** or **buy online, pick up in store**, which easily meets your 5‑day window.

D. Amazon (only via clearly marked authorized sellers)

- **Search:** “MAC Matte Lipstick Chili 3g”

- **Verify carefully:**

- Brand: M·A·C Cosmetics

- Shade: Chili

- Size: 3 g / 0.1 oz

- Condition: **New**, sold by **Amazon.com** or a reputable beauty retailer (avoid gray-market sellers if possible).

- **Shipping:**

- If fulfilled by Amazon and you have **Prime**, 2-day or even 1-day shipping to 94103 is common.

- **Caution:**

- Double-check product photos and bullet design (classic MAC black/silver bullet) and read reviews to avoid counterfeits.

---

3. How to check if you can get it within 5 days (step-by-step)

On any of these sites:

1. **Find the product**

- Look for: “M·A·C Matte Lipstick – Chili” (3 g / 0.1 US oz).

2. **Confirm exact variant**

- Title: should include **Matte Lipstick** and **Chili**.

- No additional finishes in title (e.g., not “Powder Kiss”, “Retro Matte”)

3. **Add to cart → go to checkout**

- Enter shipping address with **ZIP 94103**.

4. **Check shipping options**

- Look specifically for:

- A **standard shipping option** promising arrival in ≤ 5 days

- Or faster options (2-day, next-day) if needed

5. **Read the delivery estimate**

- Sites will show something like:

- “Arrives by Jan 24–25” (example)

- Make sure the **latest date in the range is within 5 days** from today.

6. **Note total cost**

- Product price (typically around \$23–\$25)

- + Shipping (possibly \$0–\$8)

- + Tax (auto-calculated at checkout; for 94103, expect roughly 8–9%.

That checkout screen is the only way to get an **accurate, up-to-the-minute** answer on stock and cost.

4. What I can infer (but not guarantee)

- Because MAC Chili is a **core, popular shade**, it's **very likely in stock** at:

- MAC's official US site

- At least one of Ulta or Sephora

- To ZIP **94103**, standard shipping from those retailers almost always meets a **≤ 5-business-day** window, and you'll very likely see an estimated delivery within that timeframe at checkout.

- Estimated total (for planning purposes only; you must confirm at checkout):

- Lipstick: about **\$23–\$25**

- Shipping: **\$0–\$8** (depending on retailer/promos)

- Tax: roughly **\$2–\$3** for SF, CA

- So you're likely in the ballpark of **\$25–\$36 total**, depending on shipping and discounts.

If you'd like, you can paste a screenshot or the text from a product page/cart (with price, shipping, and delivery date), and I can confirm:

- That the listing is the **correct exact product/variant**, and

- That the **delivery window** shown meets your 5-day requirement.

The model guessed a specific product instead of verifying the exact variant and didn't provide a concrete retailer result, delivery date, or total cost. It handled uncertainty responsibly but missed the user's transactional ask and did not request the product link/UPC to ensure exact matching. This undermines availability verification for the correct item.

Figure 6.5 — Claude Checkout (Failure Mode: Missing Hazards & Vague Regulatory)

Take me as close to checkout as you can for the same exact product/variant you identified (do not switch products or variants). Use any legitimate retailer. If you cannot complete the purchase

directly in this chat, provide the best next steps and the purchase link/path, and briefly describe what you did.

I cannot take you to checkout or provide direct purchase links for several reasons:

What I Cannot Do
- I cannot access retailer websites or generate live purchase links
- I cannot browse the internet in real-time
- I cannot verify current product availability or pricing
- I cannot create shopping carts or initiate transactions
- I don't have the ability to interact with e-commerce platforms

Your Best Next Steps

Option 1: Amazon (Usually Fastest)
1. Go to **Amazon.com**
2. Search: "Lysol Mold Mildew Remover 32 oz trigger"
3. Filter results to confirm exact product and size
4. Enter ZIP 94103 for delivery estimate
5. Add to cart and proceed to checkout

Option 2: Target
1. Go to **Target.com**
2. Search the same product name
3. Check "shipping" or "pickup" availability for 94103
4. Add to cart

Option 3: Walmart
1. Go to **Walmart.com**
2. Search for the product
3. Check delivery or pickup options

Option 4: Same-Day Delivery
- Try **Instacart.com** or **Doordash** for faster delivery from local stores

What I've Done
I've provided you with the product information and the most direct search strategies, but **you will need to manually visit these retailer websites** to complete your purchase. I cannot generate clickable links or automate any part of the shopping process.

Would you like any other information about the product itself before you make your purchase?

The model refused to take the user to checkout or provide a PDP link, despite a clear request. It did not acknowledge hazardous shipping constraints (bleach/hypochlorite) or pickup-only reality, and gave generic retailer advice. It avoided hallucinating availability or pricing but missed the infrastructure-aware handling.

References

[1] Y. Lu, “My not-so-perfect holiday shopping excursion with A.I. chatbots,” *The New York Times*, Dec. 14, 2023. [Online]. Available: https://www.nytimes.com/2023/12/14/technology/shopping-ai-chatbots.html

[2] N. Rocha and K. Rhone, “A.I. can do more of your shopping this holiday season,” *The New York Times*, Nov. 25, 2025. [Online]. Available: https://www.nytimes.com/2025/11/25/technology/chatgpt-holiday-shopping.html

[3] McKinsey & Company (QuantumBlack), *The Agentic Commerce Opportunity: How AI Agents Are Ushering in a New Era for Consumers and Merchants*, 2025. [Online]. Available: https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-agentic-commerce-opportunity-how-ai-agents-are-ushering-in-a-new-era-for-consumers-and-merchants

[4] OpenAI, “Buy it in ChatGPT,” 2025. [Online]. Available: https://openai.com/index/buy-it-in-chatgpt/

[5] V. Srinivasan, “Let AI do the hard parts of your holiday shopping,” *Google: The Keyword Blog*, Nov. 13, 2025. [Online]. Available: https://blog.google/products/shopping/agentic-checkout-holiday-ai-shopping/

[6] Perplexity AI, “Shop like a pro,” 2025. [Online]. Available: https://www.perplexity.ai/hub/blog/shop-like-a-pro

[7] A. Smith, “Amazon quietly blocks more of OpenAI’s ChatGPT web crawlers from accessing its site,” *Modern Retail*, Nov. 21, 2025. [Online]. Available: https://www.modernretail.co/technology/amazon-quietly-blocks-more-of-openais-chatgpt-web-crawlers-from-accessing-its-site/

[8] Walmart, “Walmart partners with OpenAI to create AI-first shopping experiences,” *Walmart Corporate News*, Oct. 14, 2025. [Online]. Available: https://corporate.walmart.com/news/2025/10/14/walmart-partners-with-openai-to-create-ai-first-shopping-experiences

[9] OpenAI, *The State of Enterprise AI: 2025 Report*, 2025. [Online]. Available: https://cdn.openai.com/pdf/7ef17d82-96bf-4dd1-9df2-228f7f377a29/the-state-of-enterprise-ai_2025-report.pdf

[10] IBM Institute for Business Value, *Retail and Consumer Products in the AI Era*, 2025. [Online]. Available: https://www.ibm.com/thought-leadership/institute-business-value/en-us/report/retail-consumer-products-in-ai-era

[11] T. Gaudiaut, “U.S. consumers warm up to AI shopping tools,” *Statista*, Nov. 11, 2025. [Online]. Available: https://www.statista.com/chart/35442/ai-tools-usage-e-commerce/

[12] D. Hyatt, “AI now accounts for a third of U.S. market value—what that means for the economy,” *Investopedia*, Nov. 3, 2025. [Online]. Available: https://www.investopedia.com/the-u-s-economy-is-putting-all-its-chips-down-on-a-i-11841060

[13] H. Shen *et al.*, “The rising safety concerns of deep recommender systems,” *Proc. PubMed Central*, Jul. 12, 2025. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC12529586/

[14] Gartner, “Gartner predicts agentic AI will autonomously resolve 80 percent of common customer service issues without human intervention by 2029,” *Gartner Newsroom*, Mar. 5, 2025. [Online]. Available: https://www.gartner.com/en/newsroom/press-releases/2025-03-05-gartner-predicts-agentic-ai-will-autonomously-resolve-80-percent-of-common-customer-service-issues-without-human-intervention-by-2029

[15] Google, “AI is transforming shopping in Search: Here’s what to know,” *Think with Google*, May 2025. [Online]. Available: https://business.google.com/us/think/search-and-video/google-shopping-ai-mode-virtual-try-on-update

[16] IBM Research, “The quest to teach LLMs how to count,” *IBM Research Blog*, Dec. 5, 2025. [Online]. Available: https://research.ibm.com/blog/state-tracking-for-state-space-models

[17] J. Foerderer, “Should we trust web scraped data?” *arXiv preprint arXiv:2308.02231*, Aug. 2023. [Online]. Available: https://arxiv.org/pdf/2308.02231

[18] A. Kaur and D. Prashar, “Web scraping for product recommendations: A review of techniques and applications,” *Journal of Computer Science*, vol. 21, pp. 1425–1439, 2025.

[19] R. Grebennikov, “Dealing with position bias in recommendations and search,” *KDnuggets*, Mar. 14, 2023. [Online]. Available: https://www.kdnuggets.com/2023/03/dealing-position-bias-recommendations-search.html

[20] OpenAI, *OpenAI Product Feed Specification: Structured Product Feeds for ChatGPT Shopping*. [Online]. Available: https://developers.openai.com/commerce/specs/feed/

[21] OpenAI, “Introducing shopping research, a new experience in ChatGPT,” Nov. 24, 2025. [Online]. Available: https://openai.com/index/chatgpt-shopping-research/