DEEPLUMEN WHITE PAPER

AI Shelf Benchmark: Measuring Recommendation Readiness in AI Shopping

The new ecommerce shelf is not a search results page, a marketplace category, or a social feed. It is the generated answer inside AI shopping assistants. This white paper defines a benchmark framework for measuring which brands are visible, understandable, comparable, and recommendable when shoppers ask AI what to buy.

DeepLumen | AI-readable commerce infrastructure for agentic discovery

Executive Summary

AI shopping assistants are changing where product selection happens. A shopper can now describe a goal, constraint, budget, skin type, room size, technical requirement, or gift context inside an AI assistant and receive a short list of recommended products before visiting a store. The brand may win or lose the sale upstream, inside the answer.

This shift creates a measurement gap. Ecommerce teams are used to tracking search rankings, ad impressions, marketplace placement, social reach, and conversion rate. Those signals still matter. But they do not fully answer the new question: when an AI assistant evaluates products for a real shopper intent, does the brand appear on the AI shelf?

The AI Shelf Benchmark is designed to measure that upstream selection layer. It evaluates whether AI shopping systems can discover a brand, understand its products accurately, match products to natural-language buying needs, compare them against alternatives, surface a trustworthy purchase path, and avoid unsafe or exaggerated recommendations.

This document is a benchmark framework, not a fabricated leaderboard. The first useful step is to define a reproducible method. A public leaderboard should come only after repeated testing, stable prompts, clear scoring rules, and human review of model outputs.

The Thesis

The central thesis is simple: ecommerce is moving from search presence to agentic selection. In search presence, the user sees many links and performs the comparison manually. In agentic selection, the AI assistant compresses the comparison and presents a smaller set of options.

This changes the commercial unit of competition. The old unit was a ranked result, an ad slot, a product grid position, or a marketplace listing. The new unit is answer inclusion. A brand needs to be not only available to the AI system, but also useful enough for the AI system to mention, explain, compare, and recommend.

OpenAI's shopping help documentation describes product results that can appear when a question suggests shopping intent. It says product options may include imagery, product details, and links to sites where users can learn more or purchase. It also states that product results are selected independently and may consider intent, context, structured metadata, price, reviews, and safety standards. This is a strong signal that AI shopping surfaces are recommendation systems, not simple link directories.

For Shopify merchants, Shopify Catalog is an important distribution foundation. Shopify documentation explains that eligible products can be available to agentic storefronts and that agent discovery files such as /agents.md, /llms.txt, and /llms-full.txt can help AI agents understand store context. But Shopify also distinguishes Catalog from agent discovery files and notes that open-web discoverability remains a separate path. Catalog inclusion is therefore a starting point, not the whole competitive layer.

AI shelf presence is the state in which a product or brand appears inside an AI-generated shopping answer with enough accuracy, context, and actionability to influence a buyer's decision.

Why This Benchmark Now

The first wave of AI search tools measured whether brands appeared in generated answers. That was useful, but too broad for ecommerce. A brand mention does not equal a product recommendation. A citation does not equal a purchase path. A crawler visit does not equal revenue.

Ecommerce teams need a more commercial benchmark because AI shopping decisions happen at product level. A shopper is not only asking "What brands exist?" The shopper is asking "Which one should I buy for my situation?" That requires product facts, category logic, constraints, reviews, policy context, availability, and trust boundaries.

Existing SEO tools are not built for this. They measure rankings, backlinks, technical crawlability, page speed, and keyword demand. Existing AI visibility tools can measure share of voice across prompts. But the emerging business question is sharper: can the AI assistant recommend the product for a specific buying job?

The AI Shelf Benchmark exists to close that gap. It turns the vague phrase "AI visibility" into a measurable commerce question. It also gives brands a way to compare their AI readiness by category, not only by domain authority or generic brand awareness.

What Is the AI Shelf?

The AI shelf is the set of brands, products, merchant options, and purchase paths surfaced by an AI assistant when a user expresses shopping intent. It is not a physical shelf and it is not a static web page. It is dynamically generated from user intent, model knowledge, search retrieval, product metadata, merchant data, reviews, policies, safety rules, and available integrations.

This matters because the AI shelf is compressed. A marketplace page can show dozens of products. A generated answer may show three to six options, or even one recommended choice. That compression creates a new bottleneck. Brands that fail to enter the answer may never receive the visit.

Benchmark Methodology

A credible AI shopping benchmark needs to separate visibility from recommendation quality. The benchmark should not ask only whether a brand appears. It should ask whether the appearance is useful, accurate, justified, and actionable.

The proposed pilot uses five measurement steps.

- 1. Select a narrow category. Begin with one vertical where purchase decisions require meaningful context. The recommended first category is clean beauty because it exposes ingredient, skin type, claim, routine, and safety issues.
- 2. Build a brand set. Include a balanced sample of large incumbents, high-growth DTC brands, and smaller Shopify challenger brands. The goal is not to punish small brands, but to measure whether AI systems can understand them.
- 3. Run real shopper prompts. Use non-branded prompts that reflect how consumers ask AI assistants for help: budget constraints, skin concerns, gifting, ingredient exclusions, comparison requests, and purchase-path questions.
- 4. Score output quality. Evaluate whether the assistant included the brand, described the product accurately, matched the user need, compared alternatives fairly, provided a usable purchase path, and avoided risky claims.
- 5. Repeat and review. Run prompts across multiple AI assistants and repeated sessions. Use human review to catch hallucinated facts, stale prices, unsafe recommendations, and false confidence.

The benchmark should be transparent about uncertainty. AI shopping surfaces change frequently, and generated answers can vary by user context, region, model, retrieval availability, personalization, and product data freshness. For that reason, the first benchmark should report methodology, task design, scoring definitions, and observed patterns before over-indexing on a single ranking.

The AI Shelf Score

The proposed AI Shelf Score is a 100-point framework. It measures downstream commercial readiness rather than raw technical accessibility. A brand with good SEO may still score poorly if the AI system cannot connect products

to shopper intent. A brand with strong product pages may still score poorly if the assistant invents claims or cannot locate the correct purchase path.

This score is intentionally product-centered. It is designed for categories where a single brand can have strong and weak SKUs. A brand-level score is useful for the leaderboard. A product-level score is more useful for optimization.

Dimension	Weight
AI discoverability	20%
Product understanding	20%
Intent fit	20%
Comparison readiness	15%
Actionability	15%
Trust and risk control	10%

Pilot Category: Clean Beauty

The first pilot should focus on health and beauty, specifically clean beauty and skincare. This category is suitable because AI assistants must reason across several dimensions that are easy to get wrong: skin type, ingredients, claims, price, routine order, contraindications, texture, fragrance, reviews, and user risk tolerance.

Clean beauty also maps well to DeepLumen's existing industry coverage. The category has many DTC and Shopify brands, rich product detail pages, heavy use of claims, and strong reliance on trust. It is exactly the kind of market where "being readable by AI" can affect whether a brand enters a recommendation shortlist.

The pilot should not claim that the benchmark covers all commerce. It should explicitly say that category-specific measurement is required. A skincare benchmark should not use the same rubric details as a USB-C hub benchmark, bedding benchmark, or fashion benchmark. The 100-point structure can stay consistent, but the task set and attribute requirements must change by category.

Pilot Task Set

The first version should use 25 tasks. These tasks are not keyword queries. They are buying jobs. Each one tests whether an AI assistant can move from natural-language intent to a commercially useful recommendation.

- Recommend gentle hydrating serums for sensitive skin.
- Find fragrance-free clean beauty moisturizers under $40.
- Suggest beginner-friendly skincare gifts under $50.
- Recommend travel-friendly skincare sets for a weekend trip.
- Name lesser-known clean skincare brands with strong customer trust signals.

- Extract ingredients, size, price, skin type, and availability from a product detail page.
- Determine whether a product is more suitable for oily, dry, or sensitive skin.
- Check whether a product contains fragrance, alcohol, retinol, or common irritants.
- Explain the difference between a serum, moisturizer, toner, and treatment in a brand's routine.
- Create a morning and evening routine using only products from the selected brand.
- Select one product for a sensitive-skin shopper who wants hydration and avoids strong actives.
- Choose the safest first product for someone new to skincare.
- Identify the single best starter SKU and justify the choice.
- Recommend a lightweight texture option for users who dislike heavy creams.
- Find vegan, cruelty-free, or refillable options and verify the evidence.
- Compare the challenger brand against three mainstream competitors.
- Rank options for sensitive skin using ingredients and stated product use cases.
- Create a table comparing budget, ingredients, texture, skin type, and routine role.
- Explain why a smaller brand deserves consideration instead of a legacy brand.
- Identify who should not use a specific product and why.
- Find the correct place to buy the selected product.
- Determine whether inventory or availability is stated clearly.
- Check current price, bundle options, and whether the answer avoids stale information.
- Extract return policy, shipping context, and purchase risk information.
- Make one final recommendation and explain the decision in plain language.

Each task should be scored with evidence. The reviewer should record the prompt, assistant output, brand inclusion status, product facts used, incorrect claims, missing context, cited or linked sources, and the final score. The strongest benchmark asset will not be the score alone. It will be the explanation of why brands won or lost the AI shelf.

Score Interpretation

A benchmark should help teams prioritize action, not only create a ranking. The score bands below are designed to map directly to operational work.

How It Connects to ACCC

DeepLumen's ACCC scoring system measures whether a site is accessible, crawlable, structured, and high-quality enough for AI search visibility. The AI Shelf Benchmark should sit downstream of ACCC.

ACCC answers: can AI systems reach and read the site? AI Shelf answers: can AI systems use that information to recommend a product? Both are necessary. A store may fail AI Shelf because it fails ACCC first: blocked crawlers, JavaScript-heavy content, missing structured data, buried differentiators, or poor citation readiness. But a store can also pass many ACCC checks and still lose the AI shelf because the product context is not specific enough for real buying tasks.

ACCC

Measures the upstream AI readability layer: accessibility, crawlability, content structure, and content quality.

AI Shelf Score

Measures the downstream commercial layer: answer inclusion, product understanding, intent fit, comparison, actionability, and risk control.

The practical workflow is: diagnose with ACCC, optimize the AI-readable layer, verify with AI Shelf tasks, then monitor AI traffic and answer inclusion over time.

Merchant Playbook

Brands do not need to wait for a public leaderboard to start improving. The benchmark framework already points to the work that matters.

- Map priority prompts. Identify the non-branded buying questions your products should win. These are not keywords; they are shopper jobs.
- Make product facts explicit. Put ingredients, materials, variants, use cases, exclusions, policies, reviews, and purchase constraints in forms AI systems can extract.
- Reduce noisy corpus units. Keep the human storefront polished, but give AI systems a compact route to commercial truth.
- Build comparison context. AI assistants often need to explain why one product is better for a specific buyer. Give them structured trade-offs, not only brand adjectives.
- Separate claims from evidence. Especially in health, beauty, wellness, and supplements, unsupported claims can reduce trust or create unsafe recommendations.
- Connect availability to action. Price, stock, variant, shipping, return policy, and correct product URL matter because AI shopping is closer to purchase than ordinary search.
- Measure repeatedly. AI answers change. Treat benchmark testing as a loop, not a one-off audit.

Where DeepLumen Fits

DeepLumen's position is that agentic commerce requires an AI-readable commerce layer. A human storefront is designed to persuade people after they arrive. An AI-readable layer is designed to help AI systems understand,

compare, and recommend products before the visit.

That is why the AI Shelf Benchmark is strategically useful for DeepLumen. It converts a broad market narrative into a measurable operating question. The question is not only "Can AI crawl the site?" It is "Can AI use the site as a reliable source when deciding what to recommend?"

Agentic Page

Creates a machine-readable layer that exposes product facts, use cases, evidence, and purchase context without replacing the human storefront.

Corpus unit reduction

Reduces the amount of noisy or ambiguous context AI systems must process before reaching recommendation-relevant facts.

Automatic structured markup

Organizes product and merchant information so AI systems can parse identity, offers, attributes, policies, and trust signals.

Traffic monitoring

Tracks AI crawler visits, user-triggered retrieval, and downstream answer inclusion signals where available.

The benchmark also gives DeepLumen a distribution asset. A standard measurement framework can become a reason for brands, agencies, investors, and AI commerce teams to discuss the category using DeepLumen's language: AI shelf, recommendation readiness, AI-readable ecommerce, and corpus unit efficiency.

FAQ

What is the AI Shelf Benchmark?

The AI Shelf Benchmark is a measurement framework for evaluating whether ecommerce brands appear in AI shopping assistants, whether AI systems understand their products accurately, and whether those products are recommended for real shopper intents.

Is AI shelf presence the same as SEO ranking?

No. SEO ranking measures visibility in search results. AI shelf presence measures whether an AI assistant includes, explains, compares, and recommends a product inside a generated shopping answer.

Does Shopify Catalog inclusion guarantee AI shelf presence?

No. Shopify Catalog can help eligible products become available to agentic channels, but recommendation readiness depends on product context, intent fit, comparison evidence, trust signals, and AI readability.

How is AI Shelf Score different from ACCC?

ACCC measures whether a site is accessible, crawlable, structured, and citation-ready for AI systems. AI Shelf Score measures the downstream commercial outcome: whether AI shopping assistants can use that context to recommend products.

Why start with clean beauty?

Clean beauty is a strong pilot category because purchase decisions depend on ingredients, skin type, claims, exclusions, reviews, safety boundaries, and routine fit. Those factors reveal whether AI systems can reason beyond basic product availability.

Should this become a public leaderboard?

Yes, but only after the methodology is stable. A leaderboard should be based on repeated tests, consistent prompts, clear scoring rules, and human review. Otherwise it risks becoming a marketing artifact rather than an industry benchmark.

Sources and Source Notes

OpenAI Help Center, Shopping with ChatGPT Search: https://help.openai.com/en/articles/11128490-shopping-with-chatgpt-search

OpenAI Developers, Overview of OpenAI Crawlers: https://developers.openai.com/api/docs/bots

Shopify Help Center, Shopify Catalog and product discovery for agentic storefronts: https://help.shopify.com/en/manual/online-sales-channels/agentic-storefronts/products

Shopify Help Center, Requirements for being included in Shopify Catalog: https://help.shopify.com/en/manual/promoting-marketing/seo/shopify-catalog/requirements

DeepLumen, ACCC Scoring System: https://www.deeplumen.com/blog/accc-scoring-system/