Findability Test Standard

The method every Evaluate follows, published so results can be checked by buyers, partners and the vendors whose engines we test.

Version 1.0 Published November 5, 2026 Change log

Both pillars use one intent set and one catalog truth set, so their results can be compared and a re-test can repeat the original exactly. Plain-language summary; the engagement letter refers to this version by number.

Intents come from real shoppers

Sources, in order of preference: the store's own search logs (top queries, zero-result queries, queries followed by exits), its customer-service questions, category questions on Reddit and review sites, and Google's "People also ask" for the category. We keep the shopper's wording, including typos and loose units ("75 inch tv", "seats 6").

Each intent is tagged by the constraint it carries, so results can be broken down by what failed:

Constraint	Example	A correct answer must respect
Size or fit	"tv stand for 75 inch tv"	Stand width at or above the TV's width
Price	"oak desk under $500"	Price at or below $500 on the test date
Material	"solid wood dining table"	Material attribute or stated construction

Constraint	Example	A correct answer must respect
Use or audience	“couch good for kids and dogs”	Performance or pet-friendly fabric
Capacity	“dining table that seats 8”	Stated seating, or a length that fits 8
Style or room	“small apartment sectional”	Dimensions under a stated threshold
Combination	“grey sofa under 80 inches for a dog”	All constraints at once

A standard Evaluate uses 30 intents with at least 4 constraint types; an extended Evaluate uses 60–80 intents covering all seven. The intent set is frozen before testing and kept with the report.

Catalog truth is built from the store’s own data

For each intent we list the products that should match, from the store’s product feed or a crawl of its product pages, checked by a person against the page (dimensions, material, price, availability). A product counts as a match only if the store’s own page supports it. If the store stocks nothing that fits, the intent is marked “no match in catalog” and scored separately, because the right answer is to say so and suggest the nearest product.

Method A: on-site natural-language search

1. **Run each intent** in the store’s search box on desktop and mobile, logged out, in a clean browser session, from a US IP address. We record the full results page, the autocomplete suggestions and any “did you mean” rewrite.
2. **Capture evidence:** a timestamped full-page screenshot, the result list as text (product, price, position) and the URL, named by intent ID and run time.

3. **Adjudicate by hand.** A reviewer marks each of the top 10 results as relevant, partly relevant (fits some constraints) or not relevant, against the catalog truth. A second reviewer checks every failure and a 20% sample of passes. Disagreements are settled against the product page and noted.
4. **Classify each intent** as Good (a fitting product in the top 3), Partly right (fitting products only below position 3, or results that ignore one constraint) or Failed (zero results, or no fitting product in the top 10 while the catalog has one).
5. **Find the root cause** for every failure: missing attribute, attribute present but not indexed, synonym or unit gap, constraint ignored by the engine, merchandising rule overriding relevance, or no matching product.

In an extended Evaluate we run the same intents on two competitors chosen with the buyer. Competitor results are shown only to the buyer and never published by name.

Method B: AI findability

Prompts. Each intent becomes one or two prompts in the shopper's words: an open request ("best sofa for a small apartment with a dog, under $1,500") and, for a subset, a store-aware request ("does [store] sell a sofa that works for a dog?"). Open requests measure whether the assistant finds the store at all; store-aware requests measure whether what it says about the store is true. Prompts do not contain the store's name unless the form calls for it.

Surfaces. Standard: ChatGPT (search and shopping results), Perplexity, Google AI Mode, Google AI Overviews and Microsoft Copilot. Extended adds Gemini. We record the model or mode where the product shows it.

Runs. Each prompt runs 3 times per surface, on different days within one week, in a fresh logged-out or new-account session with memory and personalization off where the product allows, from a US IP address in one fixed state. A standard Evaluate is about 450 answers.

Evidence. Every answer is saved as a timestamped full screenshot plus the answer text, the cited links and the product cards (name, price, merchant).

Adjudication. A reviewer checks every answer that mentions the store, and every answer to a prompt where the catalog has a fitting product but the store does not appear. Facts are checked against the store's product page on the same day. A product counts as present only if it is the store's own listing or links to the store; a brand the store resells, sold elsewhere, does not count.

Technical checks behind the AI results

1. **Crawler access, tested by actual response.** We read robots.txt for OAI-SearchBot, GPTBot, PerplexityBot, Googlebot, Google-Extended and Bingbot, then request a sample of product, category and policy pages with each bot's user agent and record the status, any challenge page and the page size. User-agent tests cannot reproduce checks that verify a bot's IP range, so a 200 response is reported as "not blocked by user agent", and a block is reported only when we see it.
2. **Feed and catalog eligibility.** For Shopify stores: Shopify Catalog requirements, whether attributes kept in metafields are mapped on the Channel mapping page, and the ChatGPT channel's policy requirements. For all stores: Google and Microsoft Merchant Center feed presence and disapprovals (with access granted by the store), and Perplexity Merchant Program enrollment.
3. **Structured data.** Product, Offer and AggregateRating markup present, valid, and matching the visible price and availability.
4. **Fact-stating content.** Whether size, material, use and care facts appear as attributes or plain text on the product page, not only in images or PDFs.

Each failed intent gets a root cause from this list, or "unexplained" when none of the checks accounts for it.

Metrics and how rates are printed

Failure rate

Failed intents divided by intents with a match in the catalog

Zero-result rate

Intents returning nothing, with how many had a match in the catalog

Precision at 3 and at 10

Share of the top 3 and top 10 results that are relevant

Constraint handling

Pass rate by constraint type: size, price, material, use, capacity

First fitting position

Position of the first relevant product, median across intents

Missed-sale rate and revenue-at-risk range

On-site only, stated as a range with its assumptions

Presence rate

Share of runs in which the store or one of its products appears, per surface and overall

Recommendation share

The store’s share of all merchants recommended across runs for the intent set

Rank

Position of the store’s first product among recommended products, median across runs where it appears

Fact accuracy

Share of stated facts (price, availability, dimensions, material, policies) that match the product page

Citation rate

Share of runs where the answer links to a page on the store's domain

Competitor displacement

Intents where the catalog has a fitting product but assistants recommend named competitors instead

Every rate is printed with its sample size ("10 of 22 intents", "present in 4 of 15 runs"). With 30 intents a single failure moves a rate by more than 3 points, so we do not report decimals or compare stores on differences under 10 points. Single-run assistant results are never reported as findings. The AI sections do not estimate revenue.

Evidence retention

Screenshots, result lists and answer text are kept for 24 months so a re-test can be compared with the original, and are available to the tested store on request.

Disputes and corrections

The store can dispute any finding for 30 days after delivery, through the link beside each finding or the dispute form. We answer within 5 business days. Any error we confirm is corrected in the report within 5 business days and listed in the public correction log (date, what changed, no store name). Our liability is capped at the fee paid (terms).

Limits stated on every report

- Assistants are not deterministic. Three runs show a range, not a stable rank. We never promise a position.
- Personalization and location change answers. Our sessions are clean and US-based.
- Platforms change without notice. Results are true on the dates shown.
- Influence is indirect. The store controls catalog data, feeds, structured data, crawler access, policies and content, not model behavior, ranking, reviews or press.
- Causation is inferred. The re-test after a fix is the evidence; until we have several, the link is presented as likely, not proven.
- On-site tests do not cover logged-in personalization, vendor A/B tests, in-app search, or measured conversion impact.

Change log

Version	Date	What changed
1.0	November 5, 2026	First published version: shared intents and catalog truth, Method A (on-site search), Method B (AI findability), technical checks, metrics, evidence, disputes and limits.

We review the standard quarterly against platform changes. Each report names the version it followed.