The Golden-Query Eval Template

Score any AI shopping agent on your own catalog. One afternoon. One hundred queries. One honest answer: is it actually better than what you have?

A demo runs the queries an agent already handles. Your shoppers run the ones it doesn't. This template is the fixed query set, the scoring rubric, and the regression tracker that turn "it looked great in the demo" into a real, repeatable evaluation. Build it once, and it becomes your standing regression suite forever.

Published benchmarks like ShoppingBench and ShoppingComp show even frontier models fail most real shopping tasks. The only way to know whether an agent works on *your* products is to score it on them.

How to run it

1. **Build your golden set.** Fill the grid on page 2 from your own search logs and catalog. Aim for 100 queries in two bands: about 85 in-catalog queries spread across volume tiers (roughly 22 head, 30 torso, 33 tail) and the eight intent types, plus 10 to 15 out-of-catalog controls in their own band.
2. **Write the expected outcome** for each query. Not the exact product, the acceptance criteria.
3. **Run the set through each system** you are comparing: your current or native search, and the candidate agent. Same set, same catalog snapshot, same conditions.
4. **Score blind** on the five-axis rubric (page 3). Whoever scores should not know which system produced a result set.
5. **Compare axis by axis**, then re-run the whole thing after any change and log it in the regression tracker (page 6).

Score it with one merchandiser and one technical lead in half a day. The merchandiser owns the queries and the conversion-potential calls. The technical lead owns the harness and consistency.

Page 2: The golden-set grid

Fill each cell with that many queries from your own logs. The counts below are a starting allocation for a Plus catalog. Adjust to your traffic, but keep the tail over-weighted, because that is where agents quietly fail.

Two bands make up the 100. The grid below holds the ~85 in-catalog queries, split by volume tier and intent type. The out-of-catalog control band underneath holds the other ~15. Do not split the control band by volume tier; a query for something you do not carry has no head, torso, or tail. The two bands sum to 100.

Band 1: in-catalog queries (~85), by volume tier and intent type

Intent type	Head	Torso	Tail	Row total
Exact-product / SKU	___	___	___	___
Attribute-constrained	___	___	___	___
Natural-language intent	___	___	___	___
Comparison	___	___	___	___
Synonym / regional	___	___	___	___
Negation	___	___	___	___
Branded	___	___	___	___
Ambiguous	___	___	___	___
Column total	**~22**	**~30**	**~33**	**~85**

Band 2: out-of-catalog control queries (~15), not tiered by volume

Control query	Why your catalog can't satisfy it	Expected behavior
_____	_____	`no-match / labeled alternatives`
_____	_____	`no-match / labeled alternatives`
_____	_____	`no-match / labeled alternatives`

Repeat to 10 to 15 control rows. Band 1 (~85) plus Band 2 (~15) sums to 100.

> **Out-of-catalog controls are the band most evals skip.** These are queries your catalog cannot satisfy, like "navy midi dress" when you carry none. The correct behavior is a graceful "no match" or honestly-labeled alternatives, never a confident wrong product. Reserve 10 to 15 of your 100 queries here. An agent that hallucinates confidently fails the safety axis even if everything else looks fine.

Intent type, what it tests, and an example to crib from:

- **Exact-product / SKU:** lookup and typo correction. "Arizona sandal," "SKU 4471-BLK"
- **Attribute-constrained:** hard-constraint handling. "waterproof hiking boots size 9"
- **Natural-language intent:** does the engine read meaning. "something for a beach wedding"
- **Comparison:** superlative and intent ranking. "warmest winter jacket"
- **Synonym / regional:** query expansion across vocabularies. "trainers" vs "sneakers" vs "runners"
- **Negation:** constraint exclusion. "jacket, no logo"
- **Branded:** correct brand surfaced. "[your brand] crew tee"
- **Ambiguous:** disambiguation for your catalog. "gold," "shell," "tank"
- **Out-of-catalog control:** honest no-match, no hallucination. "[a thing you don't carry]"

Page 3: The five-axis rubric

Score the result *set* (the top 10 to 20), not a single product. Each axis is 0 to 3. The anchors are what make two people score the same result the same way, so read them before every scoring session.

Axis 1: Relevance (0 to 3)

Do the results match the query topic?

- **0:** Off-topic. The set is mostly or entirely wrong.
- **1:** A few on-topic results buried among wrong ones.
- **2:** Mostly on-topic, ranking is imperfect.
- **3:** All on-topic and ranked sensibly.

Axis 2: Intent match (0 to 3)

Do the results match what the shopper was trying to do, not just the words they typed?

- **0:** Keyword-matched, intent ignored. "Beach wedding" returns anything tagged "beach."
- **1:** Partial intent. Some results fit the goal, most don't.
- **2:** Intent mostly understood, a few literal misses.
- **3:** The set clearly understands the goal behind the words.

Axis 3: Attribute fidelity (0 to 3)

Are the hard constraints respected: size, color, material, price, negation?

- **0:** Constraint violated. A "no logo" query returns logoed products.
- **1:** One constraint respected, others ignored.
- **2:** Most constraints respected, one slips.
- **3:** Every stated constraint respected across the set.

Axis 4: Conversion potential (0 to 3)

Would a real shopper actually buy from this set? In-stock, on-brand, sensibly priced, decent imagery.

- **0:** Nothing here converts. Out of stock, off-brand, or wrong tier.
- **1:** A buyable result or two, mostly weak.
- **2:** A solid buyable set with a few weak slots.
- **3:** A set a real shopper would buy from immediately.

Axis 5: No-wrong-results (safety) (0 to 3)

Are there clearly wrong results near the top, and on out-of-catalog controls, did the agent avoid a confident hallucinated answer?

- **0:** Confident wrong answer, especially on an out-of-catalog control.
- **1:** Wrong results in the top positions.

- **2:** Clean top positions, one questionable result lower.
- **3:** No wrong results, and out-of-catalog queries get an honest no-match.

> **The quantitative bridge (optional, for the technical reader).** Convert per-product judgments into the standard IR metrics: nDCG@10 (are good results near the top), Recall@10 (did the relevant products make the top 10), MRR (how far down the first good result sits). The rubric is the human-readable layer; these are the comparable layer, one number per metric per system.

Pages 4 to 5: The scoring sheet

One row per query. Copy as many sheets as you need. Score each system on the same query, blind to which system produced the set.

#	Query	Intent type	Volume tier	Expected outcome (acceptance criteria)	Current search: R / I / A / C / S	Candidate agent: R / I / A / C / S	Notes
1	_____	_____	H / T / Tail	_____	_ / _ / _ / _ / _	_ / _ / _ / _ / _	_____
2	_____	_____	H / T / Tail	_____	_ / _ / _ / _ / _	_ / _ / _ / _ / _	_____
3	_____	_____	H / T / Tail	_____	_ / _ / _ / _ / _	_ / _ / _ / _ / _	_____
4	_____	_____	H / T / Tail	_____	_ / _ / _ / _ / _	_ / _ / _ / _ / _	_____
5	_____	_____	H / T / Tail	_____	_ / _ / _ / _ / _	_ / _ / _ / _ / _	_____

R = relevance, I = intent match, A = attribute fidelity, C = conversion potential, S = safety. Each 0 to 3. Max 15 per query per system. Repeat rows to 100.

Roll-up (fill after scoring all 100):

Bucket	Current search avg (of 15)	Candidate agent avg (of 15)	Delta	Winner
Head	___	___	___	___
Torso	___	___	___	___
Tail	___	___	___	___
Branded	___	___	___	___
Out-of-catalog control	___	___	___	___
All queries	___	___	___	___

> **Read the buckets, not just the bottom row.** A candidate can win the "all queries" average while losing the tail or branded bucket. That is the change that ships and quietly costs you conversion. Gate on the buckets.

Page 6: The regression tracker

The golden set's real payoff is the second run. After any change (model update, synonym tweak, re-index, agent version bump), re-score the *same* frozen set and diff against the prior run. Do not pull fresh queries; that breaks comparability.

Run date	What changed	Improved (queries)	Regressed (queries)	Unchanged	Worst-hit bucket	Ship? (Y/N)
_____	_____	___	___	___	_____	___
_____	_____	___	___	___	_____	___
_____	_____	___	___	___	_____	___
_____	_____	___	___	___	_____	___

The rule: count a query as improved or regressed only if its total moves past your score-delta threshold (a 2-point move on the 15-point scale is a sensible starting threshold). Block any change that regresses a tail or branded query past the threshold, even if the aggregate went up.

> **Run it on a schedule.** Before every deployment, and weekly as a watch. The first time the tracker catches a regression you were about to ship, it pays back the afternoon you spent building the set.

Want us to run this live on your catalog?

We will take your golden set, run it through a search engine built for Shopify Plus catalogs alongside your current search, and score every query with you. 30 minutes, your queries, your products.

Book a demo →

Built by the Layers team. Search and merchandising for Shopify Plus. Last updated June 2026.