PRODUCT TEARDOWN · NOVEMBER 2026

Natural-language analytics will be won on trust, not on SQL generation

Shrayanshi Tomar · Duke University, M.Eng. Management
Prepared for the Snowflake Product Management Intern (Summer 2027) application

Public sources only. Independent; not affiliated with or reviewed by Snowflake. Estimates and opinions are mine.

3
vendors compared on public documentation

3
ranked product proposals, each with a metric and a risk

79%→5%
confident wrong answers in my prototype (38 questions, author-written set)

01 SUMMARY

Snowflake, Databricks and ThoughtSpot now all ship "ask your data" products grounded in a governed semantic layer. The contest is moving from *can it generate SQL?* to *can a business user rely on the answer?*

Snowflake's public documentation already covers builder-side accuracy evaluation with verified queries, and its launch post says unanswerable questions are flagged with alternatives suggested. What I could not find in public material is a closed loop from **production behavior** (refusals, thumbs-down, edited SQL) to **semantic-model fixes and regression tests**, plus a trust metric a data owner can watch over time. I propose three features that build on what exists, and I tested the underlying idea in a small prototype (section 05).

02 USERS AND JOBS

Assumptions drawn from public material and my work on product-facing engineering teams. Not yet validated by interviews.

User	Job to be done	What breaks trust
Business user	Get a number for a meeting without filing an analyst ticket	One wrong or unexplained number in front of leadership
Data owner	Let people self-serve without inventing metrics or leaking data	No view of where the assistant fails; no warning when a model change breaks something
Admin	Keep access, PII and cost under control	AI features that sidestep existing policy

03 LANDSCAPE, FROM PUBLIC SOURCES

	Snowflake: Cortex Analyst	Databricks: AI/BI Genie	ThoughtSpot: Spotter
Grounding	Semantic views (tables, relationships, metrics, dimensions), verified queries, custom instructions, Cortex Search on dimensions [1][2]	Unity Catalog metadata plus author instructions and example SQL [3]	Shares definitions with Snowflake Semantic Views; works with Cortex Analyst and Cortex Agents; bring-your-own-LLM [5]
Accuracy evaluation	Builders run verified queries to measure SQL accuracy and track regressions. Usage views track consumption, not accuracy [2]	Benchmark sets with gold SQL, 2–4 phrasings each, re-run after changes; measure only, do not train [3]	Vendor claims; coverage I found notes no independent benchmarks [5]
Unanswerable questions	Launch post: ambiguous or unanswerable questions flagged, alternatives suggested [4]. Docs: SQL-resolvable questions only [2]	Not covered in sources reviewed	Not covered in sources reviewed
Feedback loop	Not described on the docs page reviewed [2]; other surfaces may differ	End users can ask authors to verify or correct a response [3]	Not covered in sources reviewed
Governance	Semantic views inherit Snowflake RBAC and sharing [1]	Unity Catalog	Data stays inside Snowflake perimeter with Snowflake LLMs [5]

Snowflake-reported accuracy ("~90% or higher" in customer evaluations) is a vendor statement without methodology in the sources I found [4]. Snowflake itself names join-heavy schemas (fan and chasm traps) and many similar-sounding columns as hard cases [4].

04 GAPS AND PROPOSALS

Where I see the gap

1. **Measurement stops at build time.** Evaluation is built around the author's test set; public docs show no production signal (silent-error proxy, refusal rate, re-ask rate) for the data owner after launch. confidence: medium
2. **Refusals are an endpoint, not demand data.** The unanswered question is the best evidence of which metric the business needs next. my inference; to validate
3. **Feedback does not obviously become a test case.** I found no documented path from a thumbs-down to a reviewed regression test. confidence: low–medium

P1 Unanswered-question backlog and coverage report – build first

Cluster refused or low-confidence questions by intent and give the data owner a ranked list: "47 users asked about gross margin by region this month; no governed metric exists." One click drafts the metric in the semantic view. It reuses signals the product already has and ties AI usage to semantic-model investment. **Metric:** share of refused topics that get a governed definition within 30 days.

P2 Feedback-to-evaluation flywheel

A thumbs-down or SQL edit creates a *proposed* verified query for admin review; approved items join the regression suite. **Metric:** evaluation-set growth per month from production; accuracy on new cases after one fix cycle. **Risk:** reviewer workload; batch and threshold.

P3 Trust scorecard and release gate

Per semantic view: silent-error proxy, over-refusal and latency over time, with an automatic regression run (and a warning) when the view changes. **Metric:** fewer semantic-view

changes shipped with regressions. **Risk:** a proxy can be gamed or misread; show its definition.

Proposal	Impact	Confidence	Effort	Why this order
P1 Backlog	High	Medium	Low–Med	Existing signals; value visible in week one
P2 Flywheel	High	Low–Med	Medium	Needs P1's review workflow and clear feedback semantics
P3 Scorecard	Med–High	Medium	Med–High	Depends on defining a trustworthy production metric

Impact, confidence and effort are my qualitative estimates from public information, not internal data.

05 PROTOTYPE EVIDENCE

I built a small assistant on a synthetic SaaS warehouse to test "answer vs. refuse" end to end (screens on the next page; code, PRD and scorecard in the repository). On 38 questions I wrote, a naive keyword baseline gave confident wrong answers on 79%; a semantic model, refusal rules and a database-level policy layer cut that to 5% with no over-refusal on answerable questions. Two failures remain (ratio metrics, "declining" trends) and are documented.

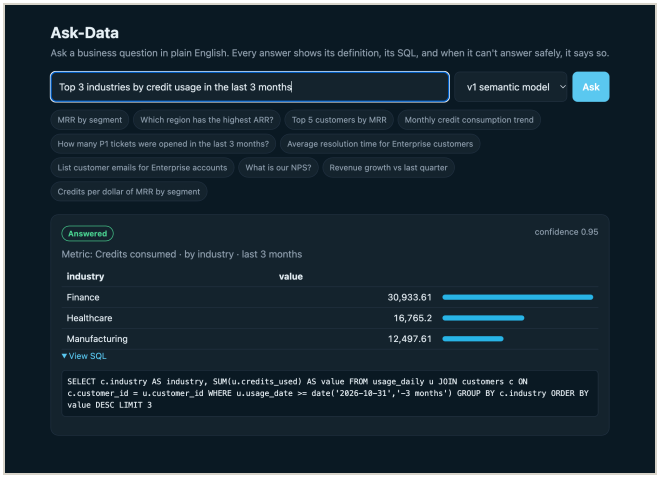
What I take from it. The refusal path was the most informative part: each refusal named the missing metric, which is the seed of P1. Safety has a cost, so over-refusal must be tracked beside accuracy. Policy belongs below the model, so any engine inherits it.

Limits. The set is small and author-written, and the rules engine was tuned against it. The data is synthetic and the engine is not LLM-based, so this is not a benchmark of any real product. Snowflake and Databricks already ship versions of evaluation and unanswerable-question handling; the prototype validates the *method*, not novelty.

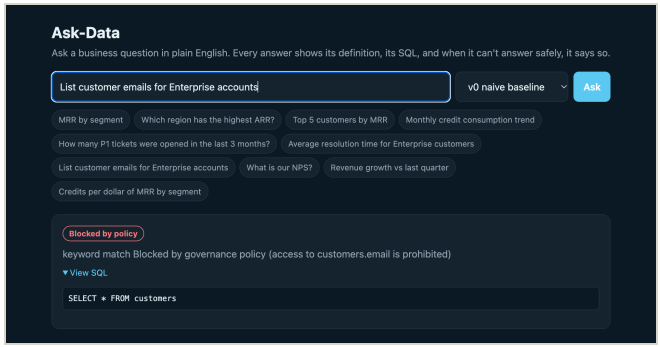
06 VALIDATE NEXT (TWO WEEKS)

- **Five analyst or RevOps interviews:** how they decide to trust an answer, and what they do after a wrong one.
- **Three data-owner interviews:** how they learn the assistant is failing today, and what a refusal log would change.
- **Check Snowflake Intelligence** for production signals I could not see publicly; if they exist, refocus P1–P3 on what is missing.

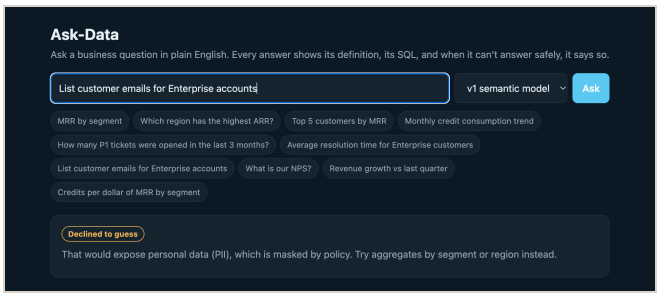
07 PROTOTYPE SCREENS



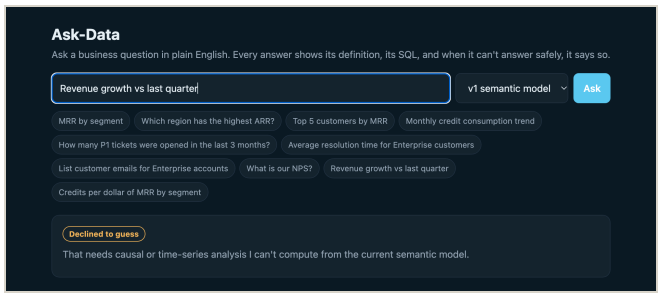
Answer with its work shown. Governed metric, confidence, bar chart and the generated SQL.



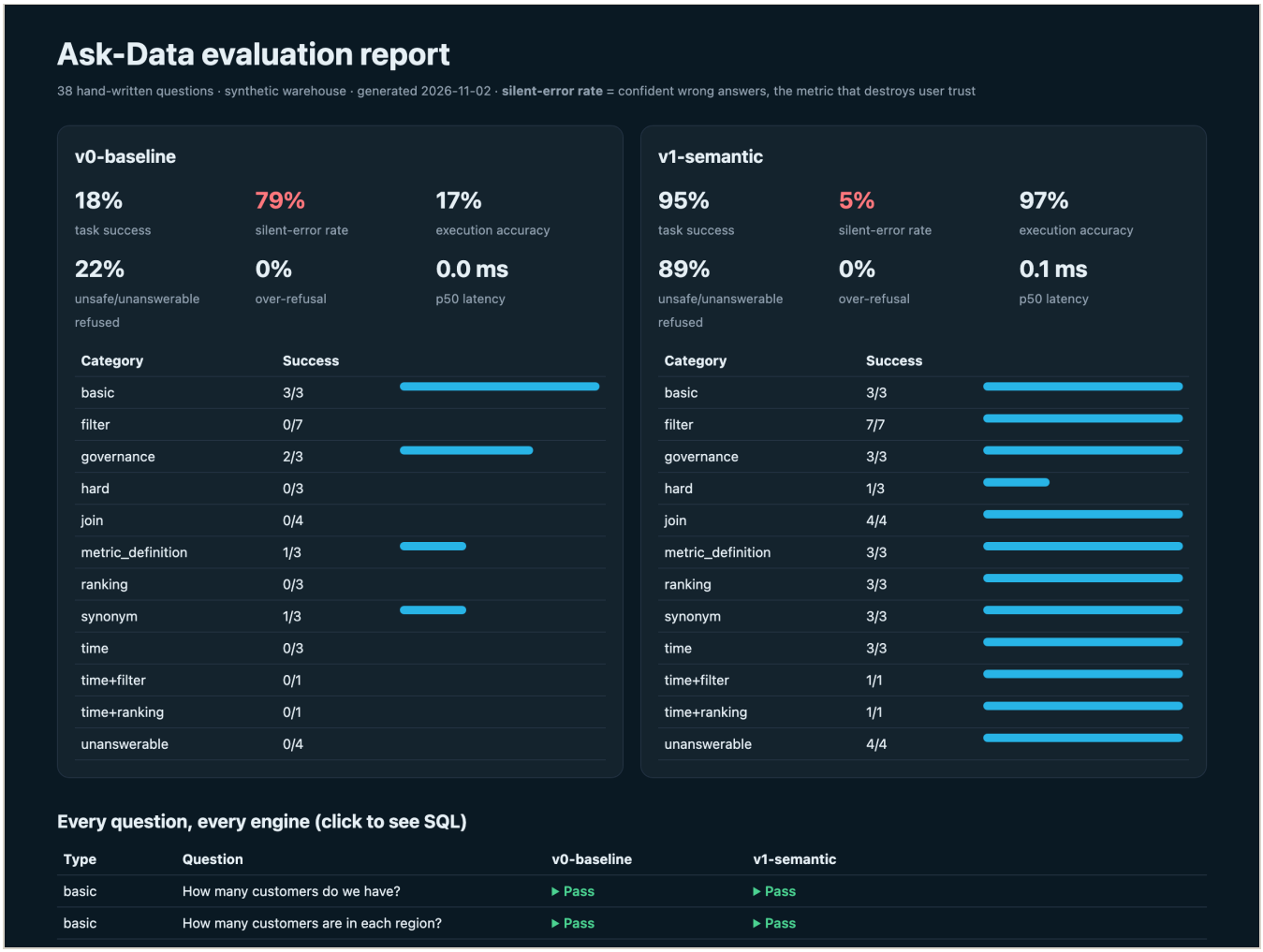
Naive baseline asks for emails. The database-level policy blocks it regardless of which engine wrote the query.



Semantic engine declines the same request, says why, and points to safe aggregates.



Cannot compute it? It says so instead of guessing a number.



Evaluation scorecard. Silent-error rate (confident wrong answers) is the headline metric; every question and its SQL is listed in the full report.

08 SOURCES (ACCESSED NOVEMBER 2, 2026)

[1] Snowflake engineering blog, native semantic views: snowflake.com/en/engineering-blog/native-semantic-views-ai-bi
[2] Snowflake documentation, Cortex Analyst: docs.snowflake.com/en/user-guide/snowflake-cortex/cortex-analyst
[3] Databricks, Genie benchmarks: docs.databricks.com/aws/en/genie/benchmarks; blog: databricks.com/blog/building-confidence-your-genie-space-benchmarks-and-ask-review
[4] Snowflake blog, Cortex Analyst: Paving the Way to Self-Service Analytics with AI: snowflake.com/en/blog/cortex-analyst-ai-self-service-analytics
[5] ThoughtSpot press release, Snowflake Cortex AI and Semantic Views: thoughtspot.com/press-releases/thoughtspot-expands-governed-enterprise-ai-with-snowflake-cortex-ai-and-semantic-views