

AI VISIBILITY AUDIT

How AI assistants describe, recommend and cite Northwind Evals

Prepared for Northwind Evals, a fictional LLM evaluation platform used here to show the format of an AEOLedger audit. Every company name, number and finding in this document is illustrative.

CLIENT	Northwind Evals (fictional)
PROMPT SET	24 buyer prompts in 4 clusters (Appendix)
ENGINES	ChatGPT, Claude, Perplexity, Gemini · 3 runs each · 288 answers
GOOGLE	AI Overviews and AI Mode, checked by hand, 24 queries each
FIELDWORK	06–10 October 2026
DELIVERED	12 October 2026
PREPARED BY	AEOLedger

SECTION 01

Executive summary

NAMED IN	SHARE OF VOICE	BEST ENGINE	WEAKEST CLUSTER
38 / 288	9.1%	Perplexity	Compliance
answers named Northwind (13%)	5th of 5 tracked brands	named in 17 of 72 answers	named in 1 of 72 answers

Assistants know Northwind as a tool for a narrow task (regression tests) but not as a platform in its category, and almost never when a buyer asks about security or self-hosting.

What we found

- 01

Visibility depends on live web retrieval.
Perplexity named Northwind in 17 of 72 answers; ChatGPT in 5 of 72. Engines that answered mostly from their training data rarely knew the product.
- 02

How-to content is the only cluster where Northwind competes.
Cluster B accounts for 20 of the 38 mentions. Two docs pages on regression testing are cited repeatedly.
- 03

Competitors own the comparison layer.
41% of all citations in category and comparison answers point to competitor-owned pages or third-party “best tools” lists. Northwind has no comparison or alternatives pages.
- 04

Key facts are invisible to crawlers.
SOC 2 status, self-hosting and EU data residency exist only in a gated PDF and in client-side-rendered docs that do not appear in the initial HTML.

What to do first (detail in Section 07)

#	ACTION	WINDOW
1	Publish a crawlable security and deployment page (SOC 2, self-hosting, EU region)	Days 1–7
2	Server-render the docs and add them to the sitemap; verify in Bing Webmaster Tools	Days 1–14
3	Publish three honest comparison pages against the most-named alternatives	Days 8–21

SECTION 02

Method and limits

This audit measures one thing: when a buyer asks an AI assistant a question that Northwind should be relevant to, does the answer name Northwind, which competitors does it name instead, and which sources does it rely on. It does not measure traffic, rankings or revenue.

STEP	WHAT WE DID
Prompt set	24 prompts written from Northwind's ICP, sales call notes and competitor positioning, grouped into 4 clusters of 6. Approved by Northwind before fieldwork. Full list in the Appendix.
Engines	ChatGPT, Claude, Perplexity and Gemini, consumer web apps, default model on each, web search left at its default setting. Each prompt run 3 times in a new session, logged out where possible.
Google	AI Overviews and AI Mode checked by hand for all 24 prompts as search queries, once each, US English, signed-out browser. Recorded separately; not included in the 288 count.
Scoring	An answer counts as "named" if Northwind Evals appears by name anywhere in it. We also logged position, whether the mention is a recommendation, and every cited URL.
Review	Every answer was read by a person. Automated extraction was used only to collect citations, then checked by hand.
Share of voice	Mentions of a brand divided by mentions of all five tracked brands across the 288 answers. Brands were agreed with Northwind before fieldwork.

Limits you should know

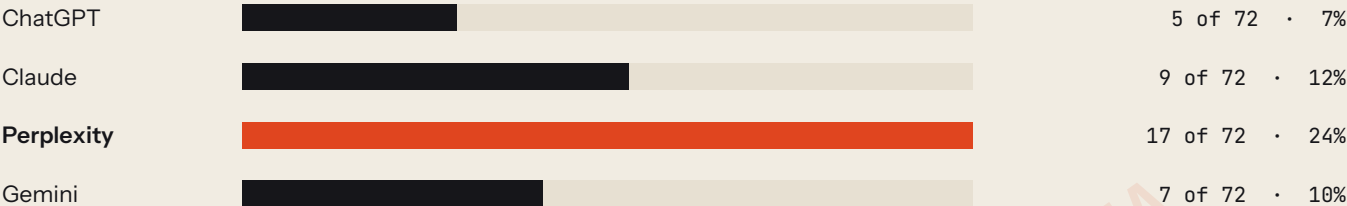
- AI answers are not deterministic. The same prompt can return a different answer a minute later. Three runs reduce, but do not remove, that noise. Treat differences of one or two mentions as noise.
- Results reflect the models and settings available during fieldwork. Engines update without notice.
- Personalisation, location, account history and paid tiers change answers. We controlled what we could and recorded the rest.
- 24 prompts describe a slice of buyer demand, not all of it. We chose them with you; we did not measure how often real buyers ask them.
- Month-over-month comparisons are valid only on this exact prompt set, run the same way. The free Snapshot used different prompts and is not comparable.
- We report what we observed and recommend changes. No one can guarantee how an AI assistant will answer, and we do not.

SECTION 03

Visibility by engine and prompt cluster

Northwind was named in 38 of 288 answers (13%). The spread across engines is wide, and it follows how much each engine relied on live web search for these prompts.

ANSWERS NAMING NORTHWIND, BY ENGINE (OF 72 EACH)



ANSWERS NAMING NORTHWIND, BY CLUSTER AND ENGINE (OF 18 PER CELL)

CLUSTER	CHATGPT	CLAUDE	PERPLEXITY	GEMINI	ALL ENGINES
A Category discovery	1 / 18	2 / 18	4 / 18	1 / 18	8 / 72 11%
B Problem / how-to	3 / 18	5 / 18	8 / 18	4 / 18	20 / 72 28%
C Comparison / alternatives	1 / 18	2 / 18	4 / 18	2 / 18	9 / 72 12%
D Requirements / compliance	0 / 18	0 / 18	1 / 18	0 / 18	1 / 72 1%
All clusters	5 / 72	9 / 72	17 / 72	7 / 72	38 / 288 13%

Darker cells = named more often. Each cell is 6 prompts × 3 runs.

CLUSTER	EXAMPLE PROMPTS	READING
A Category discovery	“best LLM evaluation tools”, “LLM observability platforms for startups”	Named mostly as a footnote in long lists. Never in a top-three recommendation.
B Problem / how-to	“how to catch hallucination regressions before release”	The only cluster where Northwind is recommended, driven by two docs pages.
C Comparison / alternatives	“Lanternbench alternatives”, “Tallyforge vs Orrery AI”	Rarely named even when the prompt names a direct competitor.
D Requirements / compliance	“self-hosted LLM eval tool with SOC 2”, “EU data residency”	Effectively absent. Assistants say they “could not find” Northwind’s compliance status.

Google AI Overviews and AI Mode (checked by hand)

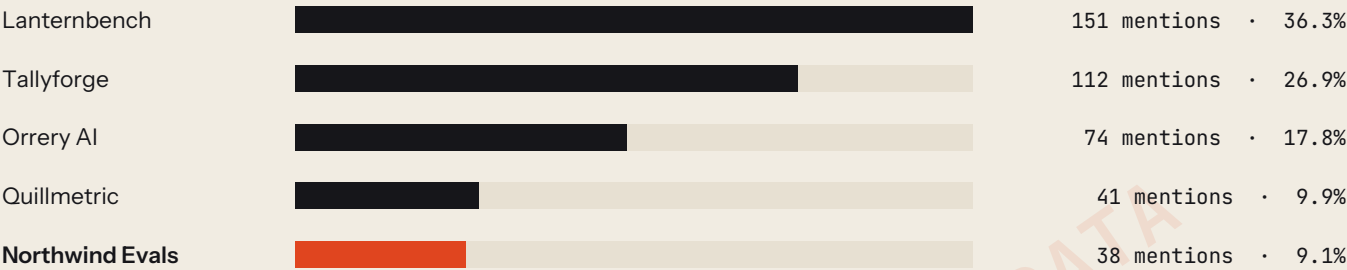
SURFACE	SHOWN FOR	NORTHWIND NAMED OR CITED	NOTE
AI Overviews	15 of 24 queries	2 of 15	Both in cluster B, citing the regression-testing guide.
AI Mode	24 of 24 queries	3 of 24	Competitor comparison pages cited in 11 of 24.

SECTION 04

Competitor share of voice

Across 288 answers we counted 416 mentions of the five tracked brands. Northwind holds 9.1% of them. Lanternbench is named roughly four times as often.

MENTIONS ACROSS 288 ANSWERS AND SHARE OF VOICE



Competitor names in this sample are fictional.

MENTIONS BY ENGINE, AND HOW OFTEN THE BRAND WAS IN THE TOP THREE WHEN NAMED

BRAND	CHATGPT	CLAUDE	PERPLEXITY	GEMINI	TOP-3 POSITION
Lanternbench	41	38	37	35	61%
Tallyforge	30	27	29	26	44%
Orrery AI	19	18	20	17	28%
Quillmetric	9	11	12	9	17%
Northwind Evals	5	9	17	7	16%

What explains the gap

- 01

Lanternbench is cited from its own site.

Its comparison pages (“Lanternbench vs X”) are cited in 27 answers. Assistants reuse the framing of those pages, including the shortlist.
- 02

Third-party lists set the shortlist.

Four “best LLM evaluation tools” articles are cited in 38 answers. Northwind appears in one of the four.
- 03

Positioning is inconsistent.

Northwind describes itself as an “eval harness”, “AI QA platform” and “LLM testing suite” on different pages. Assistants that did name it used three different category labels.

SECTION 05

Cited-source map

Engines that search the web show their sources. We logged 612 citations across the 288 answers and grouped them by who owns the page. This is the map of where assistants get their facts about this category.

SHARE OF 612 CITATIONS BY SOURCE TYPE



TOP CITED DOMAINS

DOMAIN	TYPE	CITES	ENGINES	NORTHWIND PRESENT?	OPPORTUNITY
lanternbench.com	Competitor	88	ALL four	—	Answer with your own comparison pages
g2.com	Review site	51	P, G, C	Listed, 3 reviews	Grow reviews; complete category fields
reddit.com	Community	47	P, C, G	2 threads, both 2024	Answer active threads honestly, as staff
tallyforge.ai	Competitor	44	ALL four	—	Comparison page
mlops-weekly.dev (example)	Editorial	29	P, CG	Not listed	Pitch inclusion with a factual brief
github.com	Docs / code	26	P, C	SDK repo, no README detail	Expand README; add examples
news.ycombinator.com	Community	21	P	1 launch thread	No action; earned only
capterra.com	Review site	17	P, G	Not listed	Claim and complete listing
northwindevals.com	Own site	18	P, C	—	Fix rendering; add facts pages

Engines: C = ChatGPT, CG = ChatGPT with search, P = Perplexity, G = Gemini. Domain names other than well-known public sites are fictional.

Only 3% of citations point to Northwind’s own site. Assistants are describing Northwind, when they do at all, from other people’s pages.

SECTION 06

Technical readiness

Checks that decide whether AI crawlers and search-backed assistants can read and trust Northwind's own pages. Each finding lists the evidence we saw.

CHECK	STATUS	FINDING	FIX
AI crawler access	PARTIAL	robots.txt disallows /docs/ for all user agents, which blocks GPTBot, ClaudeBot and PerplexityBot from the docs.	Allow /docs/; keep /app/ blocked.
Rendering	FAIL	Docs are client-side rendered. The initial HTML of 41 sampled pages contains navigation only.	Server-render or pre-render docs.
Sitemap	FAIL	sitemap.xml lists 23 marketing URLs. None of the 180 docs URLs are included.	Generate sitemap from the docs build.
Bing index	PARTIAL	Site not verified in Bing Webmaster Tools. 9 of 20 sampled docs URLs not indexed.	Verify; submit sitemap; use IndexNow.
Structured data	PARTIAL	Organization schema present, no sameAs links. No SoftwareApplication markup.	Add sameAs and SoftwareApplication.
Facts pages	FAIL	SOC 2, self-hosting and EU region are only in a gated PDF.	Publish an ungated security page.
Comparison pages	FAIL	None. Competitors publish 14 between them.	Publish 3 honest comparisons.
Entity consistency	PARTIAL	Three different category labels across home, docs and G2.	Pick one; update all profiles.
Pricing page	PASS	Server-rendered, crawlable, plans and limits in plain text.	—
llms.txt	NOTE	Absent. There is no clear evidence yet that major assistants use it.	Optional, low priority.

SECTION 07

Prioritised 30-day fix list

Ordered by expected effect on the findings above, then by effort. “Owner” is a suggestion; AEOLedger can do the items marked AL as part of a Foundation Sprint.

#	FIX	EVIDENCE	EFFORT	OWNER	DUE
1	Publish an ungated Security & Deployment page: SOC 2 Type II status, self-hosting, EU region, data retention.	Cluster D: 1 / 72	S	Client + AL	Day 7
2	Allow /docs/ in robots.txt for AI crawlers.	Tech: crawler access	XS	Client eng.	Day 3
3	Server-render docs; add all docs URLs to sitemap.	Tech: rendering, sitemap	M	Client eng.	Day 14
4	Verify Bing Webmaster Tools; submit sitemap; enable IndexNow.	Tech: Bing index	XS	AL	Day 7
5	Choose one category label and apply it to home, docs, G2, LinkedIn and GitHub.	SoV: inconsistent labels	S	Client mkt + AL	Day 10
6	Publish “Northwind vs Lanternbench” and “vs Tallyforge” with fair, sourced tables.	Clusters A and C; 31% competitor citations	M	AL draft	Day 21
7	Add Organization sameAs and SoftwareApplication schema.	Tech: structured data	XS	AL	Day 14
8	Expand the SDK README with use cases and a quickstart.	github.com: 26 cites	S	Client eng.	Day 21
9	Claim and complete Capterra; ask 10 active customers for G2 reviews.	Review sites: 14% of cites	S	Client mkt	Day 30
10	Send a factual brief to the four “best tools” list authors.	Editorial: 38 answers	S	AL draft	Day 30

Effort: XS under 2 hours · S under 1 day · M 2–5 days. AL = AEOLedger.

These changes make Northwind easier to find, read and cite. They do not guarantee a change in any assistant’s answers; Section 08 sets out how we will check.

SECTION 08

Re-measure plan

We will re-run the same 24 prompts, on the same engines, with the same settings, and report the change against this baseline. Nothing else is comparable.

WHEN	WHAT	REPORTED AS
Day 30	Full re-run: 24 prompts × 4 engines × 3 runs, plus AI Overviews and AI Mode.	Named in X of 288 vs 38 baseline; share of voice; citation map diff.
Day 60	Second re-run, same protocol.	Trend across three points; which fixes coincided with which changes.
Monthly (Retainer)	Same prompt set; new prompts are added as a separate, labelled set.	Month-over-month on the original set only.

How to read the change

- A change of one or two mentions on a single engine is within run-to-run noise.
- A consistent change across engines and both re-runs is a signal worth acting on.
- Engines may update their models between runs. We will note any visible model or interface change in the report.
- If an engine changes so much that the protocol no longer works, we will say so and propose a revised baseline.

Baseline to beat

METRIC	BASELINE (06-10 OCT 2026)
Named in (4 engines, 288 answers)	38 of 288 (13%)
Share of voice (5 brands)	9.1%
Cluster D, compliance	1 of 72
Northwind-owned citations	18 of 612 (3%)
AI Overviews citing Northwind	2 of 15 shown

SECTION A

Appendix: prompt set

ID	PROMPT	CLUSTER
A1	What are the best tools for evaluating LLM applications?	Category discovery
A2	LLM observability platforms for startups	Category discovery
A3	Which tools do AI teams use to test prompts before release?	Category discovery
A4	Open-source vs hosted LLM evaluation tools	Category discovery
A5	Best platform to monitor RAG quality in production	Category discovery
A6	What should I use to benchmark my chatbot?	Category discovery
B1	How do I catch hallucination regressions before a release?	Problem / how-to
B2	How to set up evals for a RAG pipeline in CI	Problem / how-to
B3	How to compare two prompt versions on the same test set	Problem / how-to
B4	How do I build a golden dataset for LLM testing?	Problem / how-to
B5	How to measure answer faithfulness automatically	Problem / how-to
B6	How to run LLM evals on every pull request	Problem / how-to
C1	Lanternbench alternatives	Comparison / alternatives
C2	Tallyforge vs Orrery AI	Comparison / alternatives
C3	Is Lanternbench worth it for a small team?	Comparison / alternatives
C4	Cheaper alternative to Tallyforge	Comparison / alternatives
C5	Orrery AI vs Lanternbench for RAG	Comparison / alternatives
C6	Northwind Evals vs Lanternbench	Comparison / alternatives
D1	Self-hosted LLM evaluation tool with SOC 2	Requirements / compliance
D2	LLM eval platform with EU data residency	Requirements / compliance
D3	Which LLM testing tools keep data on-premises?	Requirements / compliance
D4	HIPAA-ready LLM evaluation tools	Requirements / compliance
D5	LLM evaluation tool that does not train on our data	Requirements / compliance
D6	Air-gapped LLM evaluation options	Requirements / compliance

Raw answers, run logs and citation exports are delivered with the audit as a spreadsheet. In this sample, all data is illustrative.