

PREPARED FOR NORTHWIND EVALS (FICTIONAL) • RUN 29 SEP 2026

NAMED IN

4 of 27

answers. The most-named competitor appeared in 19 of 27.

When buyers ask AI assistants about LLM evaluation, Northwind Evals is rarely named.

We asked three buyer-style questions in ChatGPT, Claude and Perplexity, three times each, and read every answer by hand. This page shows what we saw.

01 THE PROMPTS

P1	What are the best tools for evaluating LLM applications in production?
P2	How do I set up regression tests for a RAG pipeline before each release?
P3	Which LLM evaluation platforms are SOC 2 compliant and can be self-hosted?

02 WHAT EACH RUN RETURNED

	CHATGPT			CLAUDE			PERPLEXITY			NAMED
	R1	R2	R3	R1	R2	R3	R1	R2	R3	
P1	☐	☐	☐	☐	☒	☐	☒	☒	☐	3 / 9
P2	☐	☐	☐	☐	☐	☐	☐	☒	☐	1 / 9
P3	☐	☐	☐	☐	☐	☐	☐	☐	☐	0 / 9

Filled = Northwind named Outline = not named. R1-R3 = three separate runs of the same prompt.

03 SHARE OF VOICE – TIMES NAMED, OF 27 ANSWERS

Lanternbench		19/27	41%
Tallyforge		14/27	38%
Orrery AI		9/27	20%
Northwind Evals		4/27	9%

Share of voice = each brand’s mentions as a share of all 46 mentions of these four brands. Competitor names are fictional.

04 WHERE THE ANSWERS GOT THEIR FACTS (58 CITATIONS)

DOMAIN	WHAT IT IS	CITES
lanternbench.com	Competitor docs and blog	11
g2.com	Review / category pages	7
reddit.com	Community threads	6
medium.com	Third-party “best tools” lists	5
github.com	Repos and awesome-lists	4
northwindevals.com	Your site	1
Other (17 domains)	Long tail	24

05 WHAT STOOD OUT

- 01 **Category questions favour whoever owns the comparison content.**

In P1, every answer that named a competitor cited a “best tools” list or a competitor’s own comparison page. Northwind has neither.
- 02 **You were named only where the assistant searched the web.**

All four mentions came from answers that cited live web results (Perplexity, and Claude with search). No answer from ChatGPT named Northwind in these runs.
- 03 **Requirements questions (P3) went to competitors with public security pages.**

Your SOC 2 report and self-hosting option exist but are not on a page the assistants cited.

Read this as a sample, not a score. AI answers vary from run to run and change as models update. Three prompts are far too few to measure overall visibility. Runs were made on 29 Sep 2026 in fresh sessions, logged out where possible, default model, location set to the United States. We report what we observed; we do not guarantee any change in how assistants answer.

Next step. The AI Visibility Audit runs 20–30 buyer prompts across ChatGPT, Claude, Perplexity and Gemini, plus Google AI Overviews, three runs each. It covers competitor share of voice, the full cited-source map, technical readiness and a dated 30-day fix list. Reply to this email to book a 20-minute walkthrough.