

The AI Grounding Test

Reliable AI. Ground every agent in meaning you already govern.

Run the checklist against your own environment, not against a vendor's feature list. Work through it with the two or three people who would actually have to answer, and time the answers where noted. The time it takes to answer is usually the finding.

Score at the end of each one. The score tells you whether the problem is a workflow annoyance or an architecture gap, and the two need different budgets.

An agent queries your warehouse and returns a number. The number is plausible. It's also different from the number in the BI report, and the difference surfaces in a meeting rather than in a test.

The model inferred what “ARR” meant from column names. It will infer something slightly different tomorrow. Prompt tuning moves the error around without removing the cause.

Section A: What your agents actually read

- ☐ Name one AI or text-to-SQL surface in production or pilot: [FILL IN]
- ☐ What does it read to understand your data? Raw schema / hand-written prompt context / a governed definition store / other: [FILL IN]
- ☐ Has an agent-produced number ever disagreed with the BI report for the same metric? Yes / No
- ☐ Can you show, for one agent answer, the exact definition it resolved against? Yes / No
- ☐ Can you trace that answer back to source columns and the code that produced them? Yes / No

Section B: Time the answers

- ☐ Time to reconstruct the provenance of a single agent output, query back to source: [FILL IN]
- ☐ Time to answer “which governed definition did the agent use here”: [FILL IN]
- ☐ Number of AI initiatives currently blocked or slowed on data trust: [FILL IN]

Section C: Requirements to test on any vendor call

- ☐ Does the platform serve agents the same governed definitions that BI tools consume, from one repository, so an agent and a dashboard cannot disagree by construction? MetaKarta serves compiled context, meaning the definition was governed and compiled before the agent ever asked. Ask what the alternative serves, and whether the agent is inferring meaning at query time.
- ☐ Is that context reachable over MCP without custom integration work? MetaKarta MCP exposes two toolsets, Metadata Management Tools and Semantic Hub Tools, from the full platform.
- ☐ Does every response return the metadata evidence that produced it?
- ☐ Is there a place to validate the context an agent will receive before the agent runs in production? Context Sandbox is where MetaKarta does this. Test the claim yourself, on your data, with the grounding visible.
- ☐ Can a business user ask a question in natural language and get the same answer the BI report gives? Ask MetaKarta runs against the same compiled definitions.
- ☐ Is column-level lineage available for the AI provenance chain, computed from the actual code rather than inferred from query logs?
- ☐ Scope the claim before you believe it. Which parts ship today, and which parts are roadmap? MetaKarta's shipped state covers structured AI consumption through Snowflake Cortex Analyst and Databricks Genie, MetaKarta MCP, and Context Sandbox validation.

Score

1 point each for: raw schema or hand-written prompt as the context source, a yes on the disagreement question, a no on showing the resolved definition, a no on tracing to source, over 30 minutes on the provenance timer.

0 to 1. Your grounding holds for now. Retest when the second agent ships.

2 to 3. The agent is guessing at meaning, and the guesses are inconsistent across runs. Expect the failure to surface publicly before it surfaces in testing.

4 to 5. There's no provenance chain to defend an output. Any AI governance mandate that arrives will land on an estate that cannot answer it.

Next step

Take one agent answer that was wrong and walk the chain backward on paper. Wherever the trail goes cold is the gap a vendor has to close in front of you, on your data.

Get in touch.

Request a demo to learn more about MetaKarta and how it delivers compiled context to AI agents.

[Request Demo](#)