SEAM

A MEASURELAB WHITEPAPER

Testing SEAM: what an intelligence model does to an AI agent meeting your data

A controlled head-to-head bench measuring what an intelligence model does to an AI agent meeting your data.

Bench run · 2026-04-29 Models · claude-sonnet-4-6, claude-haiku-4-5 Sample · 14 tasks, 39 paired trials

Abstract

Someone in your organisation has already wired an AI agent into a live data source. They pointed it at GA4, at the CRM, at a Google Sheet, at the data warehouse, or at all of those at once. They asked a question in plain language and got an answer in seconds. No pipeline. No ticket. No waiting. They will keep doing this. So will everyone else.

The market for governed agentic analytics is therefore not the firms that already route every question through a warehouse. It is everyone, because every team will pick the path of least friction, and the path of least friction is direct MCP access to whichever source holds the data. Governance has historically lived inside the warehouse. The behaviour we now need to govern lives outside it.

SEAM is Measurelab's response: an intelligence model that sits between an AI agent and the systems holding the data. It is not a semantic layer alone. It incorporates definitions, entity records, governance metadata, source priorities, telemetry, and the routing logic that resolves a question to the right place to answer it. What an agent gets when it asks SEAM a question is not raw connector access, but a governed reasoning path with the meaning of every term, the status of every metric, and the known gaps of every join already accounted for. It is the unavoidable governance layer the manifesto calls for, made operational.

To measure the difference an intelligence model makes, we built a controlled head-to-head bench across fourteen tasks spanning ten of Measurelab's modelled connections: BigQuery, GA4, GTM, Slack, Drive, GitHub, GCP, OpsHub, Harvest, and Atlassian-class data. The same model (`claude-sonnet-4-6`), same prompts, same independent LLM judge, same simulated user. The only manipulated variable was whether the agent's question was resolved through SEAM (the governed LLM-and-MCP path) or fell straight through to the raw connectors (the ungoverned version of the same path). Both profiles represent the low-friction route that employees gravitate toward in preference to the warehouse-and-dashboard alternative; the test is what governance costs, or saves, when added to that route.

That an agent backed by an intelligence model answers more correctly is not, on its own, a surprising finding. The more useful result is what it costs the ungoverned alternative to fail. Across 39 paired trials, the agent without SEAM burned 2.3 times the dollar cost, 2.3 times the input tokens, 1.5 times the wall-clock time, and 1.5 times the tool calls of the agent with it. The intelligence model saves more compute than it adds.

On the questions where institutional context matters most (those asking about defined-metric semantics) SEAM was correct 86% of the time against the ungoverned 21%, and surfaced the relevant governance, temporal, or known-gap caveat 93% of the time against 21%. Correctness overall came out at 95% versus 44%, with both pass-rate confidence intervals excluding zero.

The conclusion is not "buy a semantic layer". The conclusion is that ungoverned direct-to-source AI access is already happening inside organisations, that it is producing fluent, expensive, confident misinformation, and that the durable response is to make the intelligence model the unavoidable path rather than an optional one. SEAM is what that path looks like.

---

1. Methodology

1.1 THE TWO PROFILES

Two agent profiles were built on top of identical model and prompting infrastructure.

- **SEAM** is the governed agent. Every question is resolved against the intelligence model: metric definitions, entity records, governance status, temporal blocks (when a metric's meaning changed), known data gaps, telemetry of how metrics have been used, and the routing logic that decides which underlying source to call. The model spans warehouse and non-warehouse data alike.
- **Direct** is the ungoverned agent. It has the raw MCP connectors (BigQuery, GA4, Atlassian, Slack, Drive, GitHub, OpsHub and so on) wired straight in. It is the closest analogue we could build to what an analyst gets when they hook an LLM up to a source themselves.

Both ran on `claude-sonnet-4-6` at temperature 0 with identical seed prompts and identical follow-up budgets.

1.2 THE TASK SET

Fourteen tasks were selected to cover the breadth of an analytics consultancy's information needs, deliberately spanning both warehouse and non-warehouse data sources:

CATEGORY	EXAMPLE TASK	SOURCE
Defined-metric semantics	"What was `billable_hours` measuring before 2026-04-01 vs after?"	OpsHub + intelligence model
Entity routing	"Which GTM container is live for our Client A work?"	GTM
Cross-system	"Is the GA4 to BigQuery export configured for Client B, and which dataset?"	GA4 + BigQuery
Implicit-entity	"How fresh is our GCP pricing data?"	GCP + intelligence model
Governance / known-gap	"Can I get hours-per-engagement for March 2026? How reliable is that join?"	Harvest + OpsHub
Negative case	"How many credits has Anthropic consumed against their retainer this month?"	(Anthropic is not a Measurelab client)

Each task was assigned a primary metric (correctness, caveat-capture, or hallucination-resistance) and either a reference answer or a rubric. The point of the spread is that "your data" is rarely all in one place. An intelligence model that only governs the warehouse only governs a fraction of the surface where ungoverned AI access is already happening.

1.3 THE JUDGE AND THE USER SIMULATOR

Every trial was scored by an independent judge, a separate `claude-sonnet-4-6` instance, returning a structured JSON triple `{correct, caveatCaptured, hallucinated}` plus one-sentence justifications, against either the reference answer or rubric. The `hallucinated` flag was reserved for fabricated entities (tables, joins, projects), not merely incorrect answers.

A user simulator (`claude-haiku-4-5` , temperature 0.5, capped at two follow-ups per side) acted as a Measurelab analyst pushing back on hand-wavy or unsatisfying replies. The simulator did not see the rubric or reference answer.

1.4 STATISTICS

Per-trial paired-bootstrap on Δ = (SEAM rate - Direct rate), 2000 iterations, 95% interval. Sample size: 14 tasks across 2 to 3 trials per side, 39 paired comparisons.

A separate batch ran the same 5-task subset under `claude-haiku-4-5-20251001` to test whether the SEAM advantage was model-specific. A third batch ran the SEAM profile alone for 5 trials per task on a 3-task subset to characterise consistency.

2. Results

2.1 HEADLINE PERFORMANCE

	SEAM (GOVERNED)	DIRECT (UNGOVERNED)	Δ	95% CI
Correctness	**94.9%** (37/39)	43.6% (17/39)	**+51 pp**	[+36, +67]
Caveat capture	**46.2%** (18/39)	7.7% (3/39)	**+39 pp**	[+23, +54]
Cost per pass of the bench	**$14.38**	$33.42	**-57%**	SEAM 2.3× cheaper
Wall-clock latency	476 s	734 s	-35%	SEAM 1.5× faster
Tool calls	79	121	-34%	SEAM 1.5× more direct

Both pass-rate confidence intervals exclude zero. The bench is small but the signal is unambiguous.

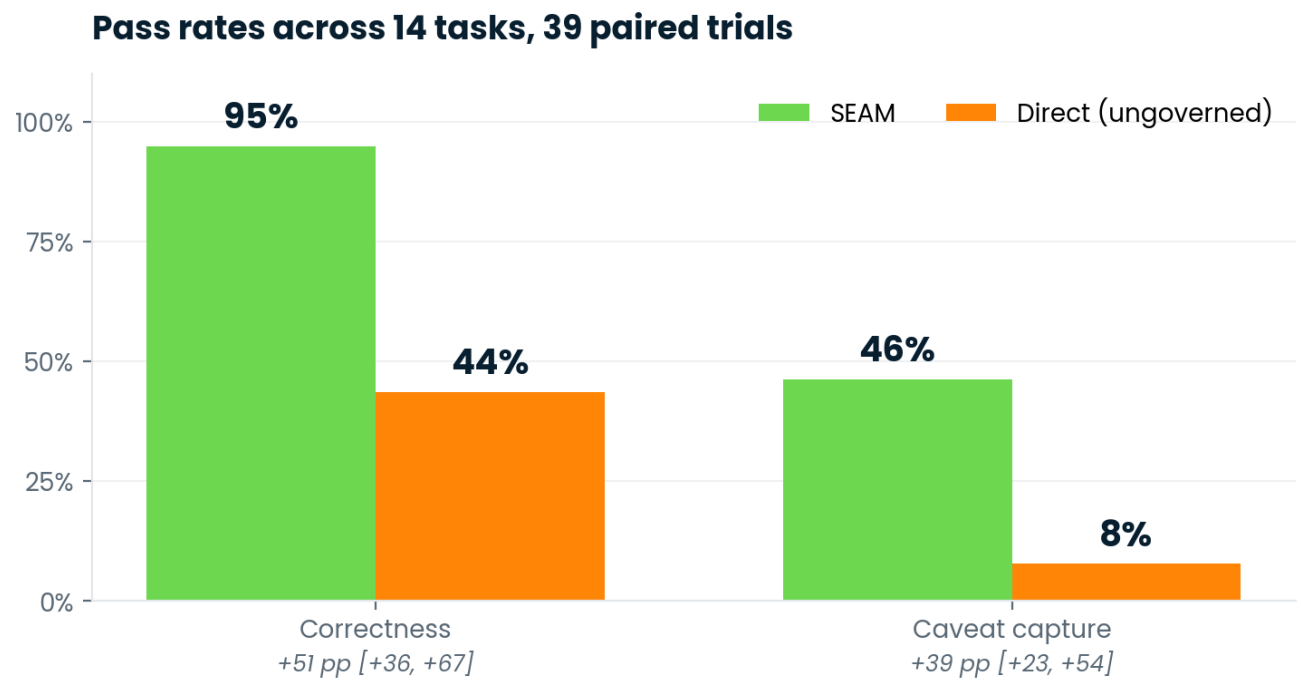

Figure 1. Pass rates across 14 tasks, 39 paired trials per side.

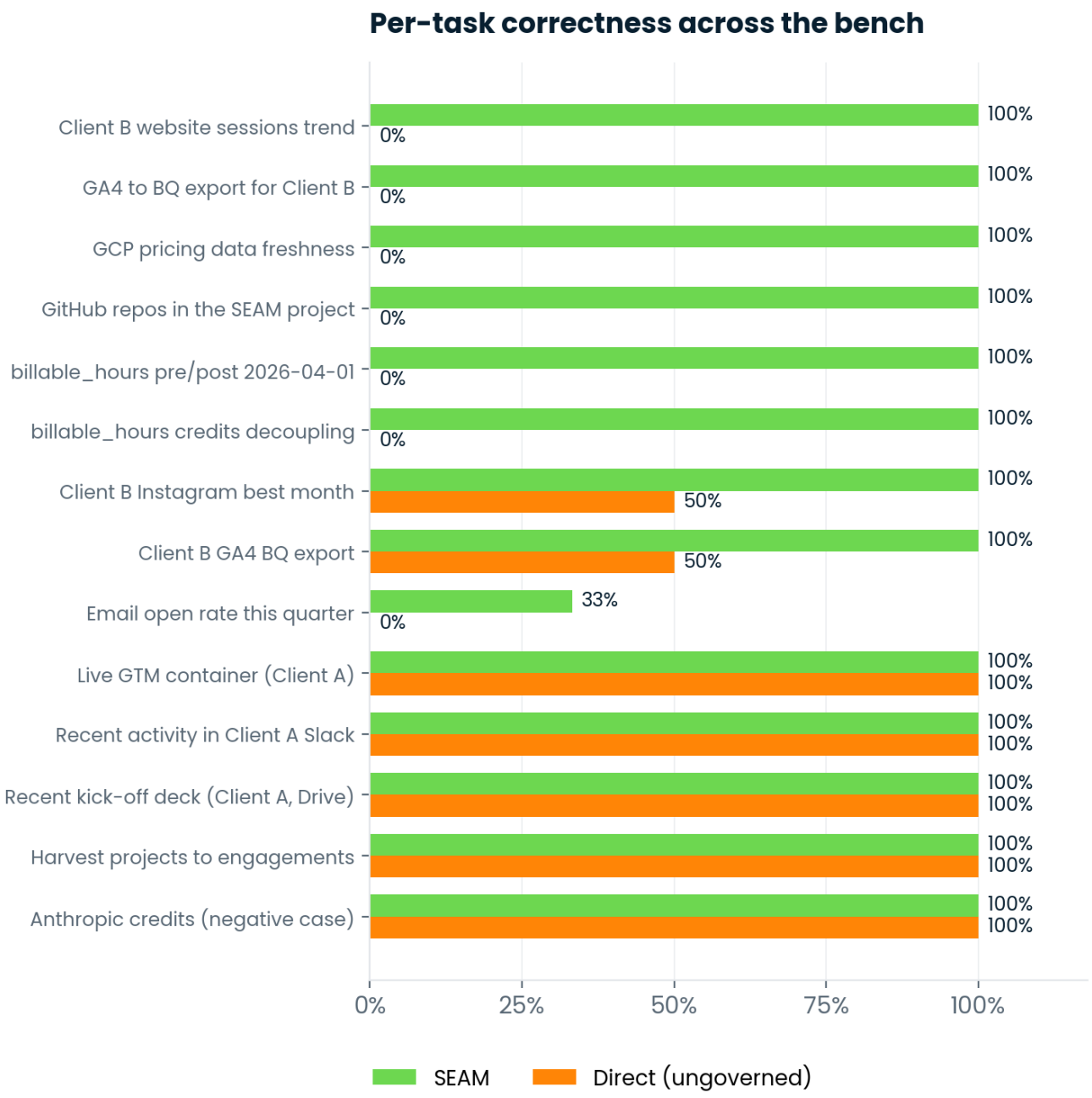

Figure 2. Per-task correctness, sorted by SEAM lead. SEAM matched or beat the ungoverned profile on every task.

2.2 THE HEADLINE FINDING: PERFORMANCE ON DEFINED METRICS

The strongest separation appears on questions whose answer lives in the firm's intelligence model: definitions, governance status, temporal blocks, known gaps. These are the questions where institutional context matters most. They are also the questions where the ungoverned default is most dangerous, because the answers it produces sound exactly as confident as the right ones.

On the **five defined-metric tasks** (n=14 paired trials):

	SEAM (GOVERNED)	DIRECT (UNGOVERNED)	Δ	95% CI
Correctness	**85.7%** (12/14)	21.4% (3/14)	**+64 pp**	[+36, +86]
Caveat capture	**92.9%** (13/14)	21.4% (3/14)	**+71 pp**	[+43, +93]

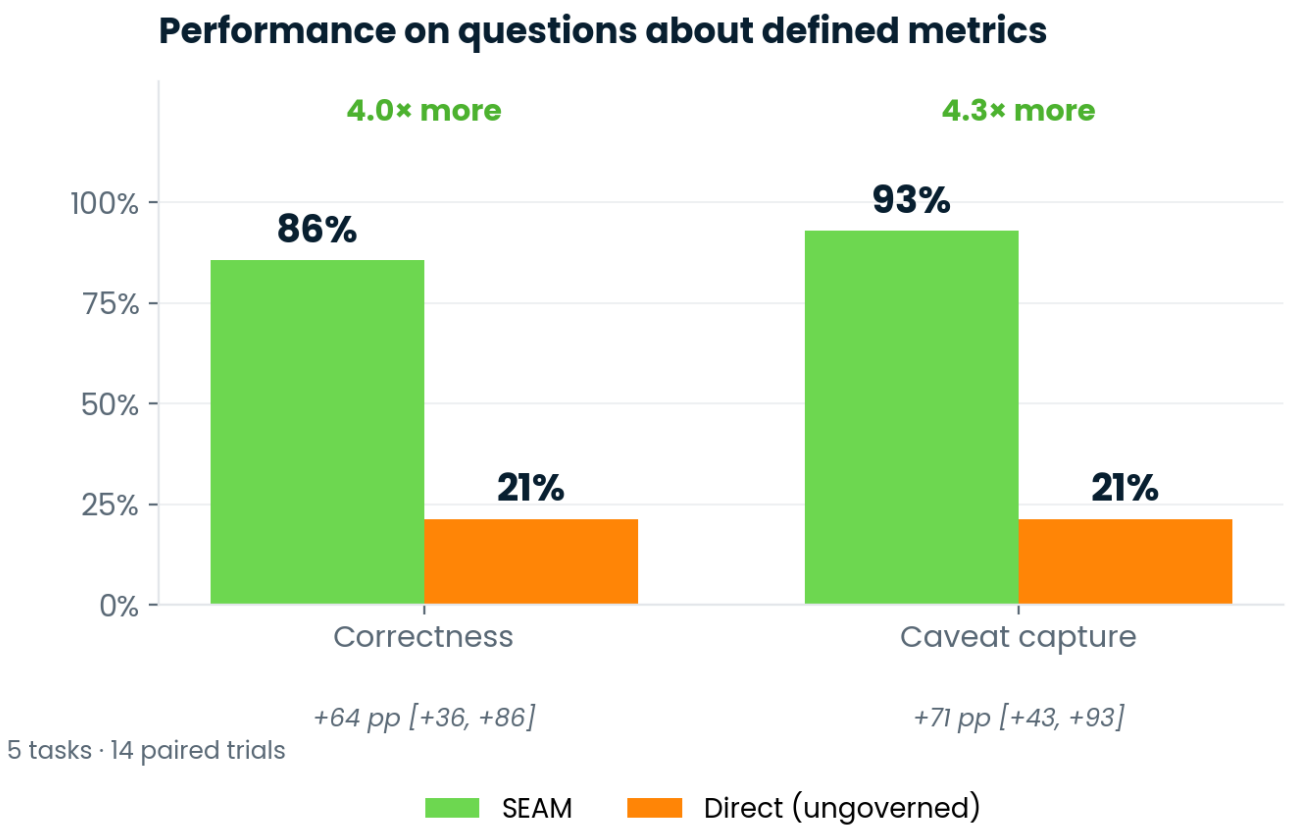

Figure 3. On questions about defined metrics, SEAM was four times more likely to be correct and roughly the same again more likely to surface the relevant caveat.

These are the largest deltas in the bench. On a question like *"When did* `billable_hours` *stop including credit equivalents?"*, a question with a single canonical answer in the intelligence model, SEAM resolved it in eight seconds and quoted the temporal change verbatim, including the caveat for trend analysis. The ungoverned agent, after 109 seconds and 21 tool calls, concluded that the metric *"is not a defined field anywhere in your systems"* and constructed a competing fictional change date.

That second answer is the part of this work that should land hardest. It is not a hallucination in the strict sense. No table was invented; no join was fabricated. It is something else: a confident, well-articulated reconstruction of a definition the agent did not know, presented to the analyst with no signal that they should not act on it. This is the failure mode of ungoverned access at scale.

2.3 THE EFFICIENCY STORY

The conventional procurement assumption is that an intelligence model adds cost: more context, more tokens, more dollars per question. The bench found the opposite. Every operational metric we tracked moved in SEAM's favour:

	SEAM (GOVERNED)	DIRECT (UNGOVERNED)	DIRECT ÷ SEAM
Cost per pass of the bench	**$14.38**	**$33.42**	**2.3×**
Input tokens	4.7 M	11.0 M	2.3×
Wall-clock latency	476 s	734 s	1.5×
Tool calls	79	121	1.5×
Output tokens	18.3 K	21.0 K	1.1×

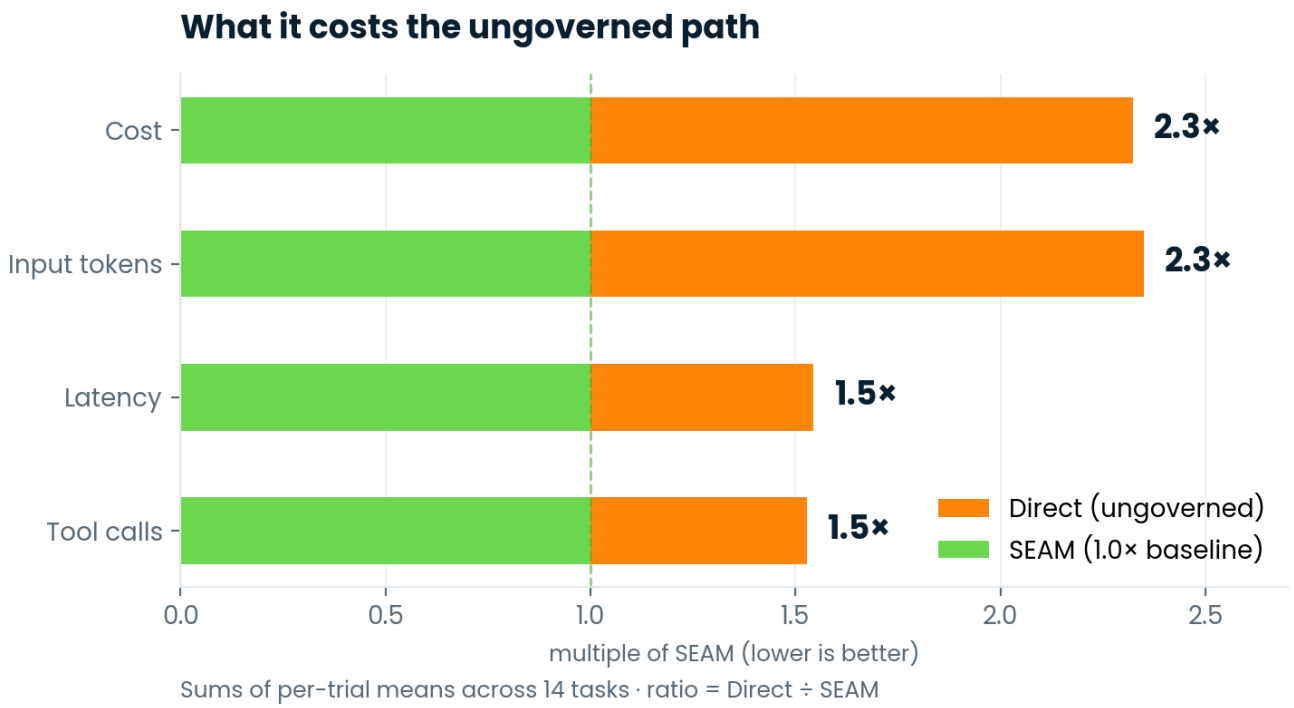

Figure 4. What the ungoverned path costs, expressed as a multiple of SEAM.

The ungoverned agent burns roughly two-and-a-half times the input tokens and dollars of the governed one, runs roughly 50% longer, and makes 50% more tool calls per question. The mechanism is straightforward. Without an intelligence model, the agent *explores* the data: opening tables, sampling rows, hunting for the right join, often retrying. With one, it resolves the question and goes straight to a targeted query.

The pattern sharpens further on questions whose answer lives in the model. Across the three deeply-modelled BigQuery tasks ("which dataset holds Client B's GA4 export", "which month

had the biggest Client B Instagram net growth", "when did `billable_hours` stop including credits"), the ungoverned agent ran the same questions at:

- 18.2× the cost
- 18.3× the input tokens
- 13.1× the tool calls
- 9.6× the wall-clock latency

The same workload, taking nearly two-and-a-half minutes per question on the ungoverned path against eight seconds on SEAM, at 18 times the cost, with worse answers.

Two operational implications follow. First, the cost story flips the conventional procurement question. At any reasonable agentic-analytics workload, an intelligence model pays for itself in token spend before it pays for itself in answer quality. Second, the bench answers the most common objection to governing the LLM-and-MCP path. The fear, when proposing to bring governance to direct-to-source AI access, is that governance will reintroduce the friction that pushed users away from warehouse-and-dashboard analytics in the first place: more steps, more waiting, more central-team involvement. The bench shows that fear is misplaced. Adding SEAM to the LLM path does not slow the agent down. It speeds it up, with less than half the input tokens and 50% fewer tool calls. Within the low-friction route that employees already prefer, governance costs nothing on the friction dimension. The governed version can be the path of least resistance.

2.4 PERFORMANCE ON ENTITY ROUTING

On the eight **entity-routing tasks** (which dataset, which container, which channel, which file, n=22 paired trials):

	SEAM (GOVERNED)	DIRECT (UNGOVERNED)	Δ	95% CI
Correctness	**100%** (22/22)	50% (11/22)	+50 pp	[+32, +73]
Caveat capture	22.7% (5/22)	0% (0/22)	+23 pp	[+5, +41]

SEAM was correct on every entity-routing trial. The ungoverned agent was correct half the time. The mechanism is the same as in the defined-metric case: when the question is "where does Client B's GA4 export live?", an agent backed by an intelligence model resolves the entity and returns the dataset; an agent without one explores BigQuery looking for clues, frequently finds the wrong project, and sometimes gives up.

Crucially, several of these tasks are not warehouse-resident. They live in GTM, Slack, Drive, and GitHub. An intelligence model that only governs the warehouse cannot govern these. SEAM's model spans them because the unit of governance is the *meaning* of the entity, not the place it is stored.

2.5 CROSS-MODEL ROBUSTNESS

We re-ran a 5-task subset under `claude-haiku-4-5-20251001` to test model dependence. SEAM's advantage held, though it compressed:

	SEAM (HAIKU)	DIRECT (HAIKU)	Δ	95% CI
Correctness	67%	40%	+27 pp	[+7, +53]
Caveat capture	13%	0%	+13 pp	[0, +33]

An intelligence model benefits a stronger model more than a weaker one on this task set, probably because exploiting rich structured context demands more from the model's reasoning. A practical takeaway: when deploying SEAM, pair it with the strongest model the use case can justify.

2.6 CONSISTENCY

To distinguish capability from luck, we ran SEAM for five trials each on three of the harder tasks. The pattern was uniform consistency where it mattered:

- The `billable_hours` temporal change. 5/5 correct, 5/5 caveat captured, latency CV 30%.
- The Anthropic credits negative case. 4/5 correct refusals, with one trial diverging into a longer exploratory path before declining.
- The email open rate. 0/5 correct, 3/5 caveat. The failure is *deterministic*: SEAM currently routes to a non-canonical upstream table for that metric, which makes it a fixable governance issue rather than a stochastic instability.

Where SEAM succeeds it succeeds reliably. Where it fails it fails predictably. Both properties are what an organisation would want from a governed system. The ungoverned alternative offers neither.

2.7 HALLUCINATION

Both profiles scored 0% on the bench's strict hallucination metric, defined as any fabricated entity (table, project, join). The result was expected, given that at temperature 0 with judge oversight neither agent invented entities outright. We report the strict floor for completeness; given both sides scored zero, we have not foregrounded it among the headline metrics.

We note for transparency that the strict definition flatters the ungoverned agent. The bench did observe at least one case where it produced a plausible-sounding alternative history about a metric definition change that did not occur. SEAM did not. Future iterations of this bench should split the strict and soft cases.

3. Limitations and what this bench can claim

This bench is small. Fourteen tasks across two to three trials per side gives 39 paired comparisons on the main run, fewer on the subgroups. The size is a deliberate constraint rather than an oversight: each trial is expensive to run end-to-end (agent plus user simulator plus judge across multiple turns of conversation), so working within a fixed compute budget meant choosing breadth across categories over depth within them. The implications are worth being explicit about.

We report confidence intervals, not p-values. A reader expecting a frequentist hypothesis test will not find one. The bootstrap intervals describe the range of plausible Δ values consistent with the data we have. Where they exclude zero (correctness, caveat capture, defined-metric subgroups) we treat that as a strong directional signal. We do not claim that a null hypothesis has been rejected at any conventional α level. The multiple-testing surface across primary metrics, subgroups, and model variants is wide enough that uncorrected p-values would be misleading and properly corrected ones would be hard to interpret at this sample size.

The population of inference is the bench itself. A statistical test conventionally infers from a sample to a population that the sample was drawn from. The fourteen tasks here are a curated set spanning ten connection types, not a random draw from "all possible analytics questions". The right reading of the results is: on tasks of this shape, in a Measurelab-style operating context, SEAM produced these effects with this uncertainty. Generalising to other firms, other sectors, or other AI usage patterns requires either replicating the bench in those contexts or arguing on theoretical grounds that the underlying mechanism (institutional context as a substitute for exploratory tool use) is the same.

Headline effects are robust to the small sample. Some subgroup effects are not. A correctness delta of +51 pp with a CI of [+36, +67] on 39 paired trials is large enough that the small-n caveat does not undermine it. The interval would tighten with more trials but the direction and magnitude are clear. Other intervals sit closer to zero and should be treated with more care. The Haiku caveat-capture interval [0, +33], in particular, touches zero and is consistent with no effect; we report it because it is what the data shows, but we would not cite it as evidence of a Haiku caveat-capture lift without further trials.

Live-data drift, judge family, and the strict hallucination definition are additional constraints. Several tasks ask for current numbers (sessions, freshness) that drift over weeks, so reproducibility is point-in-time. The judge is itself a Sonnet-family LLM, which reduces but does not eliminate self-preference bias; cross-family judging (Opus or a non-Anthropic model) would tighten this. The strict hallucination metric counts only fabricated entities, not confidently-wrong narratives, which we observed at least once on the ungoverned side.

The next bench is the right answer to most of these. Doubling trials per task to six brings the headline n to 84 paired comparisons, which tightens every interval materially. Expanding from

fourteen to twenty-five or thirty tasks reduces the curated-bench critique. Cross-family judging closes the self-preference gap. None of these are necessary to publish what we have. They are the next moves in a programme of work rather than a one-off bench.

---

4. Conclusion

The argument for an intelligence model is usually made on quality grounds. This bench supports that case (SEAM was four times more likely to be correct on defined-metric questions, and roughly the same again more likely to surface the relevant caveat) but it is not the argument the bench makes loudest.

The argument it makes loudest is about *behaviour*. People in your organisation already point AI agents at live data sources, and they will keep doing so for the same reason any of us pick any tool: because it is right there, because it is faster than waiting for a central team to model the answer in a warehouse and surface it through a dashboard, because it works often enough. The instinctive worry, when proposing to govern that behaviour, is that governance will reintroduce the very friction users are trying to escape. This bench tested both versions of the low-friction LLM-and-MCP path under controlled conditions and found that worry unfounded. Adding SEAM to the path does not slow it down. It speeds it up. The premise that governance trades speed for correctness is, on this evidence, inverted: governance trades nothing, and improves correctness.

For organisations weighing how to respond to the spread of agentic data access, the practical implication is straightforward. The instinctive responses (restrict tool use, mandate warehouse routing, build documentation that hopes to be read) all fail because they fight behaviour rather than govern it. The response that holds is to make the governed path the path of least resistance, and to make it the only path that reaches the data. Definitions become infrastructure rather than documentation. Meaning becomes consistent across warehouse and non-warehouse alike. Every answer carries its reasoning, inspectable by the organisation if not the user. SEAM is what the operational form of that response looks like.

The alternative is the same conversation about why nobody's numbers agree, except now the conversation happens ten times a day, conducted between agents that sound articulate while being wrong.

Acknowledgements and reproducibility

The bench, manifest definitions, verbatim agent and judge prompts, all task rubrics and reference answers, and full trial-level traces are preserved alongside the run bundle. The methodology is reproducible at a point in time. Live-data tasks (current sessions, current freshness) will return different specific values on later runs, but the structural finding (that SEAM materially improves correctness, caveat capture, and operational efficiency) should reproduce wherever a comparable intelligence model is in place.

For methodology details, raw bundles, or to discuss how this maps to your own AI-analytics tooling, contact Measurelab. The principles behind this work are set out in the SEAM Manifesto.

Bench run 2026-04-29. Models: `claude-sonnet-4-6` (agents and judge), `claude-haiku-4-5-20251001` (user simulator and cross-model batch). Statistical method: paired-bootstrap, 2000 iterations, 95% intervals. Sample: 14 tasks, 39 paired trials on Sonnet, 15 on Haiku, 15 on consistency.