PRE-REGISTERED PROTOCOL · PUBLISHED BEFORE THE RUN

Ungoverned chat vs. governed architecture on enterprise finance tasks

Both arms run the same frontier model. The variable under test is the architecture around it: who does the arithmetic, where the data comes from, how owner input is obtained, and whether every figure carries a source record.

This document contains the protocol, the dataset, the tasks, the metrics and the fairness rules. It contains no results. Results tables are published empty here and filled in after the run.

Dataset and generator published alongside this document · seed 20260907 · byte-identical on regeneration · myoctopus.ai

Contents

1	Why this is pre-registered
2	The two arms: the architecture is the variable
3	The dataset: NorthPeak Industrial
4	The five planted items
5	The three tasks
6	The metrics
7	Fairness rules
8	Result tables (to be filled after the run)
9	Caveat, and what this does not claim

What a reader can check

The dataset is synthetic and generated deterministically, so anyone can regenerate the exact inputs, re-run either arm, and compare their own numbers with the ones we publish. That is the whole point of publishing the protocol first.

© 2026 Octopus AI. Published before the run.

SECTION 1

Why this is pre-registered

A vendor benchmark published after the fact is worth very little. A protocol published before the run, with the data and the scoring key defined in advance, is worth something.

Two claims sit at the centre of how we describe our architecture: arithmetic runs as deterministic code rather than being produced by a language model, and the loop that gathers owner context is initiated, approval-gated and logged. Both are testable. Neither has been measured in public.

Pre-registration commits us to five things, and each one is a way for this to go badly for us:

- The tasks, the tolerances and the scoring key are fixed before either arm runs. We cannot move the target after seeing the numbers.
- The dataset and its generator are published, so the inputs are not ours to adjust.
- The comparison arm gets a genuine best effort, and its exact prompt is published.
- Every metric we defined gets reported — including the ones where the comparison arm wins.
- If the result contradicts the claim, the result is what gets published.

One correction to the original design

An earlier internal draft framed the two arms as a named vendor model versus our product. That would have measured the wrong variable: our own engine is a frontier model from the same class, so a model-versus-model framing tests the model, not the architecture. It would also have been a vendor hit piece, which readers discount on sight. Both arms therefore run the same model, and the argument is with ungoverned chat as an operating pattern in core finance — never with any model or vendor.

© 2026 Octopus AI. Published before the run.

SECTION 2

The two arms

Same model. Same analyst. Same day. Same dataset. Everything that differs is architecture.

Dimension	Arm A — ungoverned chat	Arm B — governed AFI layer
Model	Frontier LLM (identical build to Arm B)	Frontier LLM (identical build to Arm A)
Arithmetic	done by the model	done by deterministic code
Data access	analyst pastes what fits in context	queried from the data layer
Owner input	analyst chases by email	initiated in Teams, approval-gated, logged
Traceability	whatever the model cites	every figure carries a source record
Reviewer	analyst reads and accepts	named human approves before anything leaves

Arm B is the architecture, not a different brain. The model in Arm B has no arithmetic responsibility: it reads, routes, explains and drafts; the numbers come from the calculation engine.

What this framing deliberately refuses to test

Which model is better. That question is well covered elsewhere and it is not the question a CFO is asking. The CFO question is whether a number produced with AI assistance can be signed — which depends on where the arithmetic happened, whether the source is recorded, and whether a named human approved the output.

© 2026 Octopus AI. Published before the run.

SECTION 3

The dataset: NorthPeak Industrial

A fictional industrial group, generated deterministically from seed 20260907. Regeneration is byte-identical, which is what makes published results checkable.

Component	Detail
GL transactions	6,529 rows, October 2025 – December 2026
Budget	2,880 rows (3 entities × 8 cost centers × 8 accounts × 15 months)
FX	monthly rates per currency; rates move month to month
Cost centers	8, each with a named owner and mailbox
Entities	NP-US (USD), NP-UK (GBP), NP-IL (ILS)
Fiscal calendar	year ends 30 September, so fiscal Q1 FY2027 = October–December 2026

What each arm receives

Identical inputs: gl_transactions.csv, budget.csv, fx_rates.csv, cost_centers.csv and a short company briefing that states the fiscal calendar and the three entities. Arm A receives them as files an analyst may read, excerpt and paste. Arm B receives them loaded into the data layer.

The scoring key (GROUND_TRUTH.md) is never exposed to either arm, to the analyst operating Arm A, or to any prompt. It is held by the scorer and published with the results.

> **Reproduce the inputs**
>
> Run python generate_dataset.py from the published archive. The output CSVs must hash identically to the ones shipped in the archive; the generator asserts the row counts and the consolidated totals as it builds, so a divergence fails loudly rather than silently.

© 2026 Octopus AI. Published before the run.

SECTION 4

The five planted items

Each item discriminates on a different failure mode. None of them is a trick: every one is an ordinary situation in a real close.

ID	Planted item	Why it discriminates
P1	$47,300 conference sponsorship booked to Engineering / Travel; belongs to Marketing / Programs	only surfaces by asking the owner
P2	$312,000 December Marketing overage that is a pulled-forward Q2 campaign	the GL shows the number; only the owner knows it is phasing
P3	FY2026 reclassification moved datacenter costs from Operations to Facilities	pure organizational memory; nothing in the current period explains it
P4	Monthly FX rates vary; a single average rate is materially wrong	tests whether the arm reads the rates file at all
P5	Fiscal year ends 30 September, so fiscal ≠ calendar	the wrong reading is off by $848,663.08

P1, P2 and P3 cannot be resolved from the data alone at any level of model capability, because the explanation does not exist in the data. They can only be resolved by obtaining it from the person who knows — which is why M4, the owner response rate, is a metric and not a footnote. P4 and P5 are resolvable from the data, and are there to test whether the arithmetic is done carefully every time or approximately most of the time.

© 2026 Octopus AI. Published before the run.

SECTION 5

The three tasks

Two deliverables a finance function actually produces every quarter, plus one repeat measurement.

T1 — Consolidation

"What is the consolidated operating expense for fiscal Q1 FY2027 across all three entities, in USD?" Exercises P4 and P5. Scored against the key to the cent, with the tolerance in Section 6.

T2 — Variance explanation

"Explain the top three material variances for fiscal Q1 FY2027 against budget." Exercises P1, P2 and P3. Scored on whether the variance is identified and whether the explanation is the true one rather than a plausible restatement of the number.

T3 — Reproducibility

T1 asked three times in three fresh sessions, with no memory carried between them. All three answers recorded. "Same inputs, same number, every time" is the literal determinism claim, so it gets measured directly rather than inferred.

> **Two design corrections made during build, recorded here for transparency**
>
> T2 originally specified December 2026, which put the P1 misbooking — a November item — outside the task's own scope. Scored over the full quarter, both owner-dependent items land inside the top five variances. And T3 did not exist in the first draft: the original design measured error rate but never consistency, which is the cheapest metric here and the most direct test of the claim.

© 2026 Octopus AI. Published before the run.

SECTION 6

The metrics

Five metrics, each attached to a specific claim. Every one gets reported, whichever way it falls.

ID	Metric	Definition	Claim under test
M1	Numeric error rate	share of reported figures outside a 0.5% tolerance against the key	the model never does the math
M2	Time to defensible pack	wall clock and analyst minutes, measured separately	capacity returned to the team
M3	Traceability rate	share of figures with a resolvable source record (file, row, transformation)	audit-grade lineage
M4	Owner response rate	share of owner questions answered, and elapsed time to answer	the initiating loop
M5	Reproducibility	identical T1 answers across three fresh sessions	determinism

Why M2 is two clocks

Ungoverned chat usually wins wall clock to a first answer, and we expect it to win here. It gives you a number in minutes with no setup. That is a real advantage and we will report it plainly.

What it loses is analyst minutes to a defensible answer: the chasing, the re-pasting, the reconciling, the re-checking, and the writing-up that a reviewer will accept. A single number that blends the two clocks would hide the finding, so we do not report one.

Scoring

- One scorer, not the operator, holds the key and scores both arms blind to arm label where possible.
- A figure counts as traceable only if the scorer can follow the citation to the exact source row and reproduce the transformation. “From the GL” is not a source record.
- For T2, an explanation counts only if it names the true cause. Restating the variance in words scores zero.
- M1 tolerance is 0.5%. It is set once, here, before the run.

© 2026 Octopus AI. Published before the run.

SECTION 7

Fairness rules

Non-negotiable, and they are the point. A strawman comparison would invalidate everything downstream of it.

Rule	What it means in practice
Pre-register	this protocol is published before either arm runs; the version and date are on the cover
Best-effort Arm A	a competent analyst, a well-written prompt, unlimited attempts within the time budget, and the exact prompt published with the results
Publish the inputs	dataset and generator published, so anyone can re-run both arms
Report the losses	where ungoverned chat wins — setup time, cost to a first answer — it is reported as a win
Same analyst, same day	one operator runs both arms in a single session to remove skill and day-effect variance
No key leakage	the scoring key never enters a prompt, a pasted excerpt or the analyst's workspace

The published Arm A prompt matters more than it looks. The usual objection to a benchmark like this is that the comparison arm was set up to fail. Publishing the prompt verbatim lets a reader improve on it and re-run, which is the only durable answer to that objection.

© 2026 Octopus AI. Published before the run.

SECTION 8

Result tables

Published empty. These are the exact tables that will be filled after the run — the shape is fixed now so the reporting cannot be reshaped around a favourable result.

M1, M3, M4, M5 — by arm

Metric	Arm A — ungoverned chat	Arm B — governed AFI layer
M1 Numeric error rate (>0.5%)	—	—
M3 Traceability rate	—	—
M4 Owner response rate	—	—
M4 Median time to owner answer	—	—
M5 Identical answers across 3 runs	—	—

M2 — two clocks, by task

Task	Arm A wall clock	Arm A analyst minutes	Arm B wall clock	Arm B analyst minutes
T1 Consolidation	—	—	—	—
T2 Variance explanation	—	—	—	—
To a reviewer-accepted pack	—	—	—	—

© 2026 Octopus AI. Published before the run.

T3 — three fresh sessions, T1 answer recorded verbatim

Run	Arm A answer	Arm B answer
Run 1	—	—
Run 2	—	—
Run 3	—	—

Answers recorded exactly as returned, including any hedging or range. Any deviation between the three runs of an arm is a miss for that arm on M5.

Planted items — surfaced or missed

Item	Arm A	Arm B
P1 Misbooked sponsorship	—	—
P2 Pulled-forward campaign	—	—
P3 Prior-year reclassification	—	—
P4 Monthly FX vs. average rate	—	—
P5 Fiscal vs. calendar quarter	—	—

The scoring key exists and is sealed

The consolidated fiscal Q1 FY2027 figure, the calendar-quarter figure that a fiscal-calendar error produces, and the ranked variance list are all fixed by the generator and held in GROUND_TRUTH.md. They are published with the results, not before, and they were not derived from any arm's output.

© 2026 Octopus AI. Published before the run.

SECTION 9

Caveat, and what this does not claim

Stated up front, because it is the first objection a sharp reader raises — and because it costs nothing to say.

This is synthetic data with deliberately planted items. It demonstrates architectural behaviour under known conditions. It is not a claim about performance on any specific customer's books, and it is not a claim about any model's capability.

Specifically, this benchmark does not claim:

- that one model is better than another — both arms run the same model, by design
- that ungoverned chat is not useful — it is useful, and it wins the metrics it wins
- that these numbers would reproduce on a real ledger, which has messier data and no planted items
- that a single run settles anything; n=1 per task, and we will say so in the results

What it does establish, either way

Whether the difference between an ungoverned and a governed architecture is measurable at all on tasks a finance function performs every quarter — and if it is, on which metrics, by how much, and where the trade-offs run in the other direction.

> ### Get the dataset, or challenge the protocol
>
> The dataset, the generator and this protocol are published together at myoctopus.ai. If you think a task, a tolerance or the Arm A setup is unfair, tell us before the run and we will amend the protocol in public and version it. myoctopus.ai/benchmark-protocol

About Octopus AI — Octopus AI is the Artificial Financial Intelligence layer for enterprise finance: arithmetic runs as deterministic code, every outbound action is approval-gated by a named human, and every figure traces back to its source. SOC 2 Type II, ISO 27001, GDPR and CCPA aligned. myoctopus.ai

© 2026 Octopus AI. Published before the run.