Does GLM really fail at artifact-writing tools? — predeclared design

Question. GLM-5.2 and GLM-5.3 score SYN 0 on the two golden workflows whose deliverable is a long report. Is that a tool-calling capability deficit, or an artifact of the output-token budget the harness gave them?

Predeclared **before** any run in this study executed. Arms, scorer and pass criteria are fixed here so the analysis cannot be fitted afterwards (CLAUDE.md, *A/B evidence standard for agent-behaviour features*).

Hypothesis

H1. The failure is a budget artifact, not a capability deficit. The artifact-writing call is the single largest output emission in a golden workflow — its argument *is* the deliverable — so a global `max_tokens` cap truncates that one call preferentially over every other tool call, whose arguments are a handful of ids and numbers.

H0. GLM cannot reliably emit `write_report_artifact` regardless of budget.

Prior evidence (already owned, not re-run)

- Both GLM ids carry `protocol: anthropic` in `config/agent_channels.yaml` (pinned 2026-07-29, `5db78b3`, for an unrelated empty-tool-call-id defect), so both were subject to langchain-anthropic's `_FALLBACK_MAX_OUTPUT_TOKENS` = **4096** in every historical run.
- **Measured body sizes** of 354 successful `write_report_artifact` calls: median ≈ 1,595 tokens, p75 ≈ 2,526, p95 ≈ 3,851, max ≈ 5,365. The body **alone** exceeds 3,000 tokens in 16% of cases, leaving under 1,096 for reasoning and call syntax inside a 4096 budget.
- **Natural experiment, Run #114 → #115** (glm-5.3, same 5 workflows, 1 trial, same day, budget the only change):

workflow	#114 (4096)	#115 (uncapped)
risk–manager–control–day	SYN 0/4, obj 82.1	SYN 4/4, obj 100.0
high–board–portfolio–review–day	SYN 0/5, obj 68.6	SYN 4/5, obj 77.1
trader–rfq–booking–day	SYN 5/5, obj 65.1	SYN 5/5, obj 96.8
risk–limit–breach–day	SYN 3/3, obj 100.0	SYN 3/3, obj 100.0
ops–settlement–day	SYN 2/2, obj 90.9	SYN 2/2, obj 90.9

`completion_tokens` maxed at **exactly 4096** on all five #114 threads and reached **12,314** in #115. #114 made **zero** artifact-tool calls on the two failing workflows; #115 called it on all five.

This is suggestive but is n=1 per cell, unrandomised, and covers **only glm-5.3**. GLM-5.2 has no post-fix run at all. Hence this study.

Arms

2 models × 2 workflows × 2 trials × 2 budgets = **16 matches**, as two runs.

- **Models** — `glm-5.2`, `glm-5.3` (both anthropic-protocol).
- **Workflows** — `high-board-portfolio-review-day` and `risk-manager-control-day`: the two where the failure appears, and the two with the largest deliverables. The three workflows GLM already passes are deliberately excluded — they are the within-model control that the effect is size-dependent rather than model-wide.
- **Budgets** — `4096` (reproduces the historical condition) and `32768` (the post-fix default).

Budget is a separate RUN, not a separate arm. It is a process-level setting (`OPEN_OTC_AGENT_MAX_OUTPUT_TOKENS`) with no column on `arena_run`, so it cannot enter the contestant key the way `reasoning_effort` did in migration 0058.

- The two runs **must never be merged.** `merge_runs` folds by `(workflow_id, model_id, reasoning_effort)`; budget is not in that key, so a merge would average two operating regimes into one row.
- Runs execute **strictly sequentially**. Golden workflows resolve books by NAME and the purges are name-scoped, so two concurrent arena runs on one database contaminate each other's fixtures silently.
- No agent code is modified. Both arms use the existing env-var seam.

Scorer (fixed in advance)

Primary outcome, per trial, binary:

> **artifact produced** — the transcript contains a `write_report_artifact` call at the graded step AND a resulting `kind="text"` artifact.

Secondary, per trial: SYN `passed/total`, `objective_score`, `tool_calls`, and — from the trace DB — the count of LLM spans with `completion_tokens` at the arm's budget (the truncation signature).

Pass criteria

H1 is supported iff, across the 8 trials per arm:

1. arm B (32768) produces the artifact in **≥ 75%** of trials, AND
2. arm A (4096) produces it in **≤ 25%** of trials, AND
3. arm A shows truncated spans (`completion_tokens` == 4096) on the failing trials while arm B shows none at its own budget.

H1 is refuted if arm B's artifact rate is under 50% — that would show the budget is not what gates the call.

Anything between is reported as **inconclusive**, not rounded toward H1.

Reporting commitments

Per-trial raw results are retained (`artifacts/arena/<run>/…`, per-trial transcripts). Regressions and costs are reported alongside wins: a budget lift that fixes SYN but inflates tool calls or token spend is reported as such. A model that fails in arm B is reported as a genuine capability finding.
