THE FILTER OUTWEIGHS THE CHANNEL: DECOMPOSING WHY TOOL-USING AGENTS CORRUPT STRUCTURED RECORDS AT SCALE

Anonymous authors
Paper under double-blind review

ABSTRACT

When a tool-using agent answers a follow-up question over structured records, two jobs sit between the tool's output and the user: selecting the rows the user asked for, and delivering them. Deployed agents differ on both at once, so prior comparisons measured a joint effect. We unbundle them with four conditions over identical frozen payloads from a deployed travel agent: filtering done in-head by the model versus by a deterministic evaluator, crossed with delivery by payload (the model re-types the rows), by reference (the model names the rows and code expands them), or by a typed side channel (rows never pass through the model). The axes are far from symmetric. In flight search, taking the filter off the model recovers most of the collapse (10% to 78% turn fidelity at 100-row tables) and is the only axis with a detectable size dependence; delivery by payload costs about 23 points there, with no size dependence detectable at this resolution. Failure analysis shows why: in-head filtering is a fused interpretation-and-selection act whose set membership degrades with table size, while re-emission drops rows. A replication on the same deployed system's hotel search, a second record domain, transfers the filter-axis collapse (the filter-axis contrast is -23 points per unit of $\ln K$ on flights and -21 on hotels) and the typed channel's advantage, but the payload cost does not stay constant there: at 100 hotel rows re-emission falls to 46%, through missing emissions, empty tables and wrong row sets. The deployable rule follows: taking the filter off the model is necessary for size-immunity but not sufficient wherever re-emission itself breaks with table size or with the model family; delivery by reference helps a little; the typed side channel is the one arrangement that holds for every family and in both domains.

1 INTRODUCTION

Structured hallucination, in which an agent retrieves correct structured data through a tool call and then corrupts field values while formatting them, is not what hallucination detection is built to catch (Ji et al., 2023): validators see valid JSON, schema-parse metrics log success (Hoang, 2026). Prior work established the phenomenon **single-turn**; a deployed award-travel session runs five to twelve turns, each another synthesis pass over the same record set, and what the corruption scales with is this paper's question.

The stake is architectural: agent frameworks default to returning a tool's result into the model's context for the model to re-emit: every framework we surveyed does so, or requires wiring an opt-in bypass (LangChain, 2026; Vercel, 2026; CrewAI, 2026; Microsoft, 2026; OpenAI, 2026). The alternative (records as structured state, follow-ups answered by a deterministic tool, rows streamed on a typed side channel) exists (CopilotKit, 2026), adopted for rendering, context size and cost, with correctness noted but unmeasured (Anthropic, 2025), and the delivery distinction surfaces in practice only as a rendering bug (CopilotKit contributors, 2026); the corruption itself is documented single-turn (Yang et al., 2026) and over long agentic workflows (Laban et al., 2026). To our knowledge, no prior work joins the two: a controlled A/B on the same real record sets, with row- and field-level fidelity on **both** arms. We measure the corruption in a production multi-agent award-and-cash flight system built on Google's Agent Development Kit (ADK) as a fork of the `adk-samples` travel-concierge (Google, 2025): the fork-point architecture (full row lists into the planner's context, follow-ups in-head) is a

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

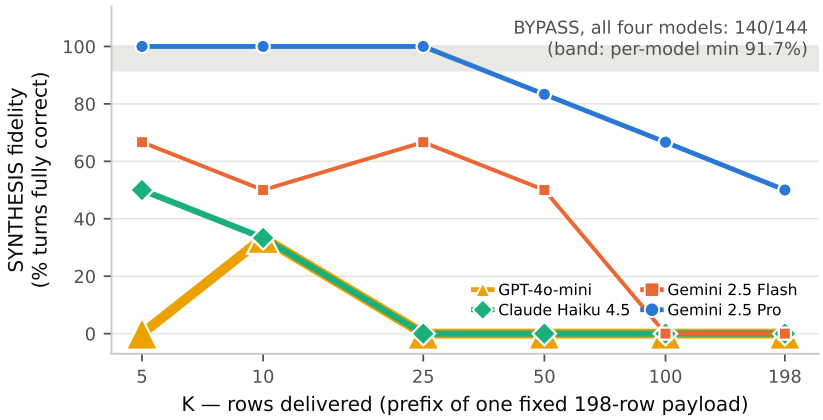


Figure 1: Four models, one content-controlled ladder: SYNTHESIS fidelity vs K-row prefixes of one frozen payload. **Capability buys threshold, architecture buys immunity**: collapse orders by capability (under 10 $\rightarrow \gtrsim$198 rows); the BYPASS band is size-flat for all four (140/144 runs). n=6 runs per (model, K) cell.

real system contract, not a strawman, and the deployed system is the same tree with the row path replaced: rows are filtered by a tool and delivered as state rather than re-typed; both run over the same frozen vendor payloads.

Contributions.

- **The size collapse, measured on both arms of a production system** (the flight study, restated).[1] Twenty dialogues over four routes vary delivered size (4→194 rows), a six-route ladder holds content constant via K-row prefixes (Figure 2), and a 3×3 Latin square separates size from position. SYNTHESIS collapses with size on every design while BYPASS is flat at 99.2% and 100% on the two size ladders (92.3% over the full corpus, where 22 of its 26 misses are wrong-argument failures). Strict per-turn field fidelity gives BYPASS 92.3% against SYNTHESIS 45.7%, a paired gap of +46.6pp that replicates across three model families: capability buys threshold, architecture buys immunity. Failures propagate (lag-1 hazard, value tracing) while BYPASS recomputes each turn from stored state; the pre-registered super-linearity test is uninformative because its null saturates (§5.6). SYNTHESIS's latency and cost scale with the re-emitted set while BYPASS's are flat in table size: 31.7 s against 4.9 s mean per turn over the corpus, at 8.7× the per-turn cost at the largest tables (§5.5).
- **A 2×2 that decomposes the gap.** Two new conditions separate the filter axis (in-head versus deterministic tool) from the delivery axis (by payload, the model re-types rows; by reference, it names row handles that code expands; by side channel, rows never pass through it): on flights the size-collapse is almost entirely in-head selection ($-$67pp at K=100, $-$23pp per $\ln K$) and delivery by payload costs about 23pp with no size dependence detectable at that resolution (§6).
- **A cross-domain replication that qualifies the rule.** On the same product's hotel search the filter-axis collapse transfers (in-head selection by reference falls from 79% to 23% across 10 to 100 rows while tool selection over the same tables holds at 90 to 94% from 25 rows on), but delivery by payload is size-dependent there (46% at 100 rows): taking the filter off the model is necessary, not sufficient, and only the typed channel holds in both tested domains (§7).
- **A methodology with computed ground truth, and its integrity report.** Expert-demonstrated dialogues replay against 408 frozen captures with ground truth computed by

[1] An earlier version of the flight study appeared in a non-archival workshop track (no proceedings). §3 to §5 substantially reproduce its instrument and results for self-containment; §6 and §7 are new.

the production filter evaluator; practising it surfaced eight instrument defects on flights and five on hotels, all fixed and reported, and the headline holds under both instrument versions (Appendix E).

2 Related Work

Constrained decoding guarantees shape, not content: validity reaches 100% while wrong-but-valid outputs dominate on small models (Ray, 2026), and Hoang (2026) measured the resulting structured hallucination single-turn in a production travel agent, introducing the field-fidelity metric and zero-out policy we adopt. The multi-turn tool-agent benchmarks we build on include τ-bench (Yao et al., 2024); a 39% single- to multi-turn drop with poor recovery is documented (Laban et al., 2025), self-conditioning is the closest mechanism match (Sinha et al., 2025), and "Beyond Task Completion" reports 27–78% of τ-bench successes as corrupt (Cao et al., 2026), procedural integrity on mock domains where ours is row-level data fidelity on real records indexed by delivered set size. CodeAct routes actions through code (Wang et al., 2024a); practitioner channels exist (LangGraph's custom stream mode, the Vercel AI SDK's data parts, AG-UI), but no framework we surveyed ships typed structured delivery as the default rendering path, and the documented motivations are rendering predictability, context size and cost, with correctness noted but not measured (CopilotKit, 2026). Turns within a dialogue are not independent and turn-level significance claims frequently fail cluster-robust correction (Schessl, 2026); every gap interval here resamples its cluster unit: dialogues, and routes in §6. The full survey, including BOULDER (Kartáč et al., 2026), is in Appendix G.

3 System and Conditions

3.1 The deployed system

The system under study (name withheld for anonymous review) is a multi-agent travel concierge on Google's ADK: a planning agent routes requests; sub-agents wrap the award (Seats.aero) and cash (SerpAPI) search pipelines, which write the canonicalised record set to session state. Two properties matter here.

1. **Search results do not enter the model's context.** The sub-agent's function response carries a count and the search parameters; records go to session state, and the client receives rows as typed events alongside, not through, the model's token stream.
2. **The model translates each follow-up's intent into arguments; a deterministic tool evaluates them.** The context tool evaluates filter arguments (vocabulary in Appendix C) over the stored record set. The remaining *translation* job (the full language-understanding problem of tool use, scope anaphora pinned by convention, §4.2) is where BYPASS's residual failures live (§5.4). The architecture changes the failure surface, not the difficulty: a mistranslated argument yields one wrong-but-real subset at any size, where re-emission puts every field of every row at stake on every turn.

We name the conditions for what happens to the records: in **SYNTHESIS** the record set passes through the model's generation and is re-emitted; in **BYPASS** records reach the interface as typed data and are never generated. The distinction is the delivery; model, sampling and instruction lineage are held constant. BYPASS is this system unmodified, driven over frozen payloads: its single model-side error source is translation, and records are never re-typed, so field values cannot mutate in transit.

3.2 SYNTHESIS: the counterfactual fork-point architecture

SYNTHESIS rebuilds the canonical ADK pattern of the `adk-samples` travel-concierge (Google, 2025), in which a search sub-agent returns a full structured row list into the planner's context, there is no context tool, and follow-ups are processed in-head, inside the mature product, so the row path, filter execution and delivery together, is what differs: the production tree deep-copied with exactly four deltas (a lineage-matched instruction variant, context tools removed, records injected mechanically into each sub-agent's summary so SYNTHESIS is not penalised for sub-agent transcription noise, and output schemas adjusted so the injected records survive; Appendix C). The planner emits each

turn's table through an output schema, parsed structurally; model, sampling, other sub-agents, tools and state writes are identical to production. Three consequences: translation error on the tool side shares one instruction lineage, so on that side the residual gap is attributable to the row path, which §6 then decomposes; mechanical delivery makes SYNTHESIS stronger than its fork-true form; and BYPASS's advantage is joint, synthesis bypass plus structured-state preservation, which §6 separates. Both conditions run Gemini 2.5 Flash at temperature 0.1 / top_p 0.5, inherited by lineage rather than chosen.

4 METHOD

Partway through the study our measurement surfaced defects in production and in the instrument itself (Appendix E); after fixing them we declared an **instrument freeze** and re-ran every experiment. Every headline number is post-fix; pre-fix numbers appear only in the labelled instrument-effect accounting (Appendix E) and the appendices.

4.1 FREEZE REAL, THEN REPLAY

Real searches ran through the production search tools, not raw vendor APIs (hub routing, deduplication and canonicalisation are part of the system under study), and everything they returned was frozen: 408 captures over 36 routes holding 9,809 records (largest payload 510 rows), four date bands and two cabins per route (Appendix C). Experiments serve the vendor boundary from these captures, never a live call.

4.2 DIALOGUE AUTHORING

Dialogue templates derive from demonstration sessions by an author who is an experienced award-travel practitioner; observed phrasings form the allowed vocabulary, and the 24 dialogues vary along four knobs over six arcs (Appendix C). Two rules keep turns scoreable rather than arguable: no heuristic constraints ("red-eye" is banned; "departing after 9pm" is allowed) and no undefined-aggregate superlatives (a superlative must name a total order; the second rule was adopted mid-study and retired such a turn in five dialogues, Appendix E). Thresholds and boundary probes are pinned so competing readings cannot both be defensible. A third convention pins down **scope semantics**: a new filter composes with the active scope, a refer-back restores the view where its referent was introduced, and a constraint revision re-executes the turn that set the constraint against that turn's own base; a turn meaning to reset must say so (worked examples in Appendix C).

4.3 THE ORACLE

Per-turn ground truth is **computed, not annotated**: each dialogue's frozen captures are replayed through the production search tools, leaving exactly the session state a real search leaves, and each filter turn's pinned arguments are passed to the production context tool's evaluator over that state; the returned envelope is the ground truth, normalised by the production canonicalisation layer at a fixed SHA for both conditions. A liveness verifier rejects empty or trivially-repeated ground truth before any model is called (Appendix C).

4.4 METRIC

The primary metric is **per-turn field fidelity with strict count-mismatch zero-out**, the policy adopted from the single-turn study (Hoang, 2026); field-level scoring against source records descends from data-to-text evaluation (Wiseman et al., 2017). A post-submission audit found this metric can score a right-count-wrong-set turn as perfect; on the size ladder of §5.1, rescoring with an exact-set rule leaves BYPASS unchanged and lowers the in-head conditions by 4 to 7pp at mid and large K; the headline corpus is reported under the original rule. §6 and §7 use the exact-set rule: a turn is correct only if the records emitted, matched on the identity key, are exactly the ground-truth set with every scored field matching after canonicalisation; a parse failure or absent artifact fails, and an empty emitted table is correct only when the ground-truth set is empty; order is scored separately. Emitted records are aligned with ground-truth records on the identity key (`flight_number`, `date`, `source`); scored fields are those present and non-empty in the ground-truth record, drawn from

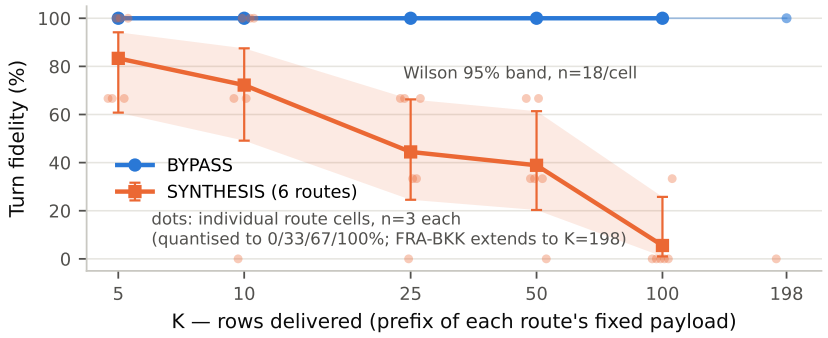


Figure 2: Turn fidelity vs K: the delivered table is the first K rows of each route's fixed payload, so size varies only by appending rows (six routes pooled). Light dots are per-route points at each K (n=3 runs each, hence quantised to 0/33/67/100%); the band is a Wilson 95% interval on the six routes pooled. Flash SYNTHESIS decays monotonically to 6% at K=100; BYPASS is 100% in all 93 runs.

a fixed fifteen-field set; comparison is exact after production canonicalisation (field list and key rationale in Appendix C). When the emitted count equals the ground-truth count, turn fidelity is the fraction of scored fields, over all aligned record pairs, whose emitted value matches the ground truth. If the counts differ, the turn's fidelity is instead **zero**, regardless of how many individual fields match. This is strict on purpose since the harm is downstream, i.e., the user acts on the table as shown. A silently missing award flight is a missed booking the user never learns existed.

A turn counts as **correct** iff it produced an artifact at all and its zero-out fidelity is 1.0. Turns where the deployed condition answered from the count-only envelope without materialising rows currently count as failures, so **BYPASS's published rate is a lower bound**. Partial-credit and order metrics are computed alongside every turn (Appendix C). Search turns establish context and are not scored; award-versus-cash comparison turns exercise a different production pipeline and are out of scope.

4.5 Runs, repeats, and the analysis layer

Each flight-corpus dialogue runs k=3 times per condition at production sampling as independent repeats, so results reproduce distributionally, not turn-for-turn. One corpus module decides what counts: network errors are holes a recovery re-run may fill, and a run with no scoreable output scores zeros on its expected filter turns (Appendix C).

Hypotheses. **H1a**: the gap appears at the first filter turn. **H1b**: synthesis failures are not independent across turns. **H1c**: synthesis fidelity degrades with the delivered record set's size. A pre-registered turn-count super-linearity conjecture was unsupported and is published as a diagnosed negative result (§5.6); the size instruments were pre-specified before their runs; a second family, **H2**, is motivation only (Appendix B). Verdicts: H1c supported (§5.1), H1a supported (§5.2), H1b supported (§5.6).

Scale. 408 captures / 9,809 records over 36 routes; 24 dialogues $\times$ k=3 runs $\times$ 2 conditions; 339 scored turns per condition carry the headline, on one model (Gemini 2.5 Flash; scale table: Table 9, Appendix C), with further families in §5.3 (Table 12, Appendix F).

5 Results

5.1 Delivered table size drives the synthesis collapse

The central result is a controlled manipulation: we vary only the size of the table delivered into the agent's context, holding route, cabin, filter vocabulary, turn wording and turn position fixed, and score the same filter turn.

Content-controlled (top-K), replicated over six routes. From each of six routes' fixed payloads (112–198 rows) we deliver the first K rows: a true prefix, so the small table is literally the opening of the large one and only the count differs. The same query ("only afternoon departures") is asked at every K; each cell pools 6 routes $\times$ 3 repeats (n=18; per-cell table: Table 7, Appendix A).

With only K varying, pooled Flash SYNTHESIS decays monotonically from 83% at K=5 to 6% at K=100 while BYPASS is perfect in all 93 runs (Figure 2); FRA-BKK's full 198-row payload, the one route run at K=198, also scores 0/3. The decay is graded, not a cliff: single routes at n=3 per cell quantise to 0/33/67/100%, but pooled over six routes the curve is a slope, and a Cochran–Armitage trend test with log-K scores rejects a flat profile ($z = -5.07$, two-sided $p = 3.9\times10^{-7}$; cluster-bootstrap robustness in Appendix F; FRA-BKK failure anatomy in Appendix A). No change of route, date, price distribution, or query is needed. Table size alone produces it.

Cross-route replication (date-varied). Independently, on four routes we vary delivered size by capture date (Figure 3, Table 6, Appendix A). SYNTHESIS falls from 75–100% on the smallest tables to near-zero beyond $\sim$74 rows in every route that reaches that size; BYPASS is flat at **99.2% (238/240)**, its two misses a single starved run (Appendix E). A control leg, a byte-identical 16-row table in all twenty dialogues, scores SYNTHESIS 77.5% and BYPASS 100% even inside dialogues whose treatment turn scored zero: the failure tracks the table delivered at that turn, not the depth of the dialogue.

The date-varied ladder confounds size with content; the top-K ladder does not, and they agree: the effect survives under a pure prefix manipulation. A size-by-position Latin square replicates the collapse at every position with no detectable position effect (Appendix E).

5.2 The gap at corpus scale

Over 24 authored dialogues, three repeats: **BYPASS 92.3% (313/339), SYNTHESIS 45.7% (155/339), gap +46.6pp** (dialogue-level bootstrap CI $[+35.7, +57.1]$). The gap is present at the first filter turn (not an accumulation effect) and never smaller than the per-route size ladders imply. BYPASS beats SYNTHESIS in 22 of 24 paired dialogues (exact sign test p=5.7×10^{-6}), 24 of 24 on GPT-4o-mini and 21 of 24 on Claude Haiku: 67 wins against 2 losses across three families.

5.3 The gap is not specific to one model family, and a tier upgrade only rents it back

With both conditions re-run over the **full 24-dialogue corpus** on two further families with the same agent stack (sampling caveat: Limitations; n=339 scored turns per cell, the headline's corpus), the paired gap holds across all three families: **+46.6pp** [+35.7, +57.1] (Gemini 2.5 Flash), **+49.6pp** [+39.2, +60.2] (GPT-4o-mini), **+55.8pp** [+41.1, +69.9] (Claude Haiku 4.5); per-condition cells with Wilson 95% CIs are in Appendix F (Table 12). The gap converges: +46.6 to +55.8pp across three families, every family's condition intervals disjoint. Family determines *who fails how*, not *how much the architecture costs*; BYPASS's floor (67.0–92.3% across the three families) tracks translation quality (Table 13), yet no family's BYPASS falls below any of the three families' SYNTHESIS; the exception is the Pro tier (72.0%, at 12$\times$ BYPASS-on-Flash's price).

The failure shapes are family-specific: Gemini corrupts (real rows, wrong set), GPT garbles (unparseable JSON), Claude omits (well-formed JSON that drops the `flights` array): one architecture, three failure shapes (Appendix F). On identical K-row prefixes of one frozen 198-row payload the collapse threshold orders by capability, under 10 rows for GPT-4o-mini, 10–25 for Claude Haiku, 50–100 for Gemini Flash, graceful only for Gemini Pro, while BYPASS is size-flat for all four (140/144; Figure 1).

A same-family **tier** upgrade only rents the gap back: over the full corpus (n=339 per cell) Gemini 2.5 Pro halves the gap rather than closing it (BYPASS 92.6% vs SYNTHESIS 72.0%, +20.6pp [+11.7, +31.4] against Flash's +46.6pp) at 5$\times$ Flash's per-turn cost and 1.6$\times$ its latency. BYPASS is tier-insensitive (92.3% Flash, 92.6% Pro), and BYPASS on the cheap production Flash beats SYNTHESIS on Pro at one twelfth the price (detail in Appendix F). Tier is the top rung of Figure 1's ordering: it raises the collapse threshold; it does not remove the mechanism.

5.4 What failure looks like

A grounding census over every emitting SYNTHESIS turn (1,306 turns across both instrument versions, 31 of them with an empty context) is categorical: fabrication tracks grounding absence, corruption tracks re-emission. The model never invents a table when the queried one was delivered. Post-fix, of 99 failed SYNTHESIS turns in the size ladder, 96 emit only real rows (partial fidelity $\approx$1.0): the set boundary is wrong, by truncation, by cascade on the model's own prior emission (a turn truncates a 74-row result to 55 and the next turn filters *that*), or by identifier splicing under load. Repairing the prompt removed most of the fabrication, which only ever occurred when no table had been delivered, but not the set corruption, which scales with delivered size and BYPASS cannot exhibit: its 26 failures are wrong-but-real filter arguments (22), right arguments over an under-delivered result (3), and one missing artifact (Appendices H, F).

5.5 Cost and latency, not only accuracy

Mean per-turn latency is BYPASS 4.9s against SYNTHESIS 31.7s; BYPASS's latency and cost are flat in table size because it re-emits nothing, while at the largest tables a synthesis turn averages 146 seconds and costs 8.7$\times$ BYPASS's per-turn cost for an 8%-fidelity answer. Dividing cost by accuracy, BYPASS produces a correct answer for about \$0.012 at every size while SYNTHESIS's cost per correct answer climbs to \$2.53 beyond 100 rows, roughly 200$\times$, an order-of-magnitude estimate that rests on the single correct SYNTHESIS turn in that band (Figure 4, Appendix F); a tier upgrade rents part of what the architecture gives outright, and the rent is high.

5.6 Errors cluster, only synthesis propagates values, and the depth test finds no super-linearity

After a failed turn SYNTHESIS fails again 71.9% of the time against 31.4% after a clean one, while BYPASS recovers 95.7%; value-tracing finds 53 of SYNTHESIS's 177 traceable failures propagated and 0 of BYPASS's 25, since BYPASS never filters its own emission. Our pre-registered super-linearity test finds none, for a structural reason: with a 56.9% first-turn failure rate the accumulation-only null saturates (Appendix C), which redirected the analysis to table size, which is monotone and does not saturate.

6 The Filter $\times$ Deliver Decomposition

§5 shows that SYNTHESIS collapses with delivered table size while BYPASS is size-flat, but the two architectures differ on two axes at once, so the +46.6pp gap is a joint effect. A filter turn has three stages, and only two of them differ across conditions. *Interpretation* of the user's turn is LLM work everywhere. *Filtering* decides which rows match: either the model does it in-head over the table in its context, or the deterministic `get_flight_context` tool does it over external state. *Delivery* gets the matching rows to the user: *by payload* (the model re-types every row), *by reference* (it names each row's *handle*, a short identifier such as `r7` stamped on the row at delivery, which code expands back to rows), or *by side channel* (rows never pass through the model). In SYNTHESIS the first two stages are one fused in-head act, reading the turn and selecting rows together, so we call that axis *in-head interpretation-and-selection*: we cannot observe a separate translation step there, and REFPICK measures the fused bundle by design; only the tool side has a separate translation act, common to every tool-side condition. Two new conditions fill the remaining corners. The experiment runs on the prefix ladder's ten-route pool, not the 24-dialogue corpus, so it decomposes the ladder's $K{=}100$ gap of 90pp rather than the +46.6pp headline, and it conditions on a delivered table (empty-table runs excluded under one rule, counts in Table 1), where the corpus scored such runs as failures.

6.1 Two new conditions: REEMIT and REFPICK

REEMIT is BYPASS's tree with delivery flipped to payload: the tool still filters over external state, but its rows are returned inline and the planner re-emits them in full. It differs from BYPASS only in delivery. **REFPICK** is SYNTHESIS's tree with delivery flipped to reference: each delivered row carries a visible handle, the planner emits only the handles of the matching rows under a delivery override, and scoring expands them in delivery order. REFPICK cannot fabricate a row or corrupt

Condition	Filter	Delivery	Exact-set fidelity (%) at K 5	10	25	50	100
SYNTHESIS	in-head	by payload	88	77	39	32	10 [4, 16]
REFPICK	in-head	by reference	97	87	52	33	34 [27, 40]
REEMIT	tool	by payload	90	99	98	87	78 [62, 91]
BYPASS	tool	by side channel	100	100	100	100	100 (all 50 runs)

Table 1: The four conditions and their prefix ladder (Flash; ten routes $\times$ five runs per cell, one scored filter turn per run; cells are means of route means; empty-table runs excluded under one rule, 38 REEMIT, 10 SYNTHESIS, 1 REFPICK, 0 BYPASS, counting them as failures would lower REEMIT and SYNTHESIS further; brackets: 95% route-clustered bootstrap intervals at K=100). Each condition removes one job from the model, and the two jobs are not symmetric.

a field, so its failures isolate exactly one thing, whether the model picks the right *set*. It differs from SYNTHESIS only in delivery, and its delivery is error-free by construction. Table 1 places the four conditions. Both run on the prefix ladder of §5.1 with the same frozen payloads, prompts lineage-matched to their parent, production sampling, and exact-set scoring: a turn is correct only if it emits exactly the answer set. Each pair sharing a scaffold isolates one axis: BYPASS−REFPICK is the filter axis with error-free delivery on both sides, REEMIT−SYNTHESIS the filter axis under payload delivery, REFPICK−SYNTHESIS and BYPASS−REEMIT the payload cost under in-head and tool selection. We report the four contrasts; averaging two estimates of an axis is a summary, not an identification.

6.2 The filter axis drives the size-collapse

Table 1 gives the ladder. We read each axis from its two paired contrasts at K=100 rather than from a factorial fit, because delivery has three levels (payload, reference, side channel) over four cells, so a two-level coding of delivery would collapse reference and side channel and make an interaction term uninterpretable. Filter, holding delivery fixed: REEMIT−SYNTHESIS = +68pp (both by payload) and BYPASS−REFPICK = +66pp (neither); their mean, −67pp for in-head selection, has a paired route-bootstrap interval of $[-74, -59]$ (p=.001). Delivery, holding the filter fixed: REFPICK−SYNTHESIS = +24pp and BYPASS−REEMIT = +22pp; their mean, −23pp for delivery by payload, has interval $[-35, -13]$ (p=.001). The size slopes, in pp per $\ln K$, separate the axes: filter $\times \ln K = -23$ $[-27, -18]$ (p=.001), against delivery $\times \ln K = -3.6$ $[-9.2, +2.2]$ (p=.22). The collapse with table size is almost entirely the filter axis. The delivery penalty shows no detectable slope at ten routes, but the interval does not establish flatness (REEMIT itself falls from 98% at K=25 to 78% at K=100): its size dependence, if any, is below this design's resolution.

The delivery axis replicates across families only in part: REEMIT on Claude Haiku 4.5 scores 76% at K=100 against Flash's 78%, but 0% on GPT-4o-mini (unparseable JSON, dropped fields); what every family shares is that BYPASS is flat at 100%.

All numbers are from one offline rescoring pass after an audit found two instrument defects (a false-perfect on right-count-wrong-set turns, an era-mixed stall exclusion) that push in opposite directions (Appendix F).

6.3 Mechanism: fused in-head interpretation-and-selection

REFPICK's failures show what in-head selection gets wrong at scale, because nothing else can go wrong there: every REFPICK failure at $K \geq 10$ carries zero invalid handles, duplicates or fabrications; what breaks is the set. An audited exemplar at K=50 emitted 17 handles for a 17-row answer, the right count, but three were not in the answer and three real rows were missed. This is imprecise set membership, a loose filter, not a counting slip; partial fidelity is about 1.0 for REFPICK by construction and says nothing about whether the set is right. REEMIT's failures are the delivery-stage analogue, rows dropped or duplicated from the emitted set (whether the tool-selected set was itself right is not separately audited); SYNTHESIS suffers both, and BYPASS touches the set with neither.

The result yields a deployable rule, stated family-conditionally. On Flash, the one family with a REFPICK cell, "hand the model a handle, never a payload" fixes delivery alone and buys about 23pp:

Condition	Filter	Delivery	K=10	25	54	100
SYNTHESIS	in-head	by payload	62	44	14	12 [6, 19]
REFPICK	in-head	by reference	79	59	38	23 [11, 36]
REEMIT	tool	by payload	73	69	60	46 [32, 60]
BYPASS	tool	by side channel	83	90	92	94 [86, 99]

Table 2: Hotel replication, exact-set fidelity (%) by K (Flash; 14 dialogues, dialogue as cluster, equal dialogue weights; k=5 for BYPASS and REFPICK, k=3 for REEMIT and SYNTHESIS; non-emitting turns fail; escape repair in Appendix F.1; brackets: 95% dialogue-clustered bootstrap at K=100).

REFPICK still sits at 34% at K=100 and still collapses with size, because the model is still selecting in its head, whereas REEMIT, with a deterministic filter and the model still serialising, holds 78%. There the load-bearing half for the size story is taking the filter off the model. On GPT-4o-mini delivery by payload fails outright, so nothing short of the typed channel holds there. What every family shares is that BYPASS, with neither job on the model, is size-immune; that is what it gets right that a reference-only fix does not.

7 Cross-Domain Replication: Hotel Search

We replicate the design on the same product's cash hotel search, a second vertical with its own tool, prompt rules and record schema, frozen independently; a first pass surfaced five instrument defects, fixed before the reported runs (Appendix F.1). Eight cities (Appendix F.1) were captured through the production search tool at one stay each, 167 to 180 rows; 14 authored dialogues (89 turns, 73 scored) span five arcs. Ground truth is the frozen tool's own output on the capped table (unpriced listings excluded and reported; a limit never cuts a tie). Differences from flights: a hotel row is about 340 tokens against 240; $K \in \{10, 25, 54, 100\}$, 54 being the production page cap; scoring is exact-set throughout. The deployed prompt encodes the same one-row and tie rules. The four conditions are the flight builders with the hotel tool and prompts substituted.

Table 2: the filter-axis result transfers. REFPICK falls from 79% at ten rows to 23% at a hundred, as on flights (87 to 34), while BYPASS, the same tool over the same tables, rises from 83% to 94%; BYPASS−REFPICK, delivery error-free on both sides, is $+71$pp $[+57, +83]$ at K=100, and filter $\times \ln K$ is -21 $[-29, -13]$. Two things differ from flights. First, hotel BYPASS is not a flat 100%: 83% at K=10, where ten-row tables empty under the filter chains and the agent sometimes answers without a tool call (no artifact, a failure under the metric), and 94% at K=100, where the misses are translation errors that cascade through later turns ("kid-friendly" sent as "pool"; a scope lost across a re-expansion); neither kind grows with K. At K=25 two runs of one dialogue looped to the runtime's 500-call cap, and every filter turn of those runs scores zero. Second, on hotels delivery by payload is not size-flat. REEMIT holds 60% at K=54 and falls to 46% at K=100 (BYPASS−REEMIT $+48$pp $[+34, +62]$; delivery $\times \ln K$ -8 $[-12, -3]$): of 219 filter turns at K=100, 29 emit no table, 28 emit an empty table where the truth is not, and 66 emit the wrong row set. So at 100 hotel rows both jobs are load-bearing: taking the filter off the model is necessary, as on flights, but here it is not sufficient.

8 Conclusion

In a deployed multi-turn agent, re-emitting a tool's records through the model's context each turn costs 46.6pp of per-turn field fidelity against carrying them as state on a typed side channel; the load variable is the size of the re-emitted set. The decomposition attributes the collapse to in-head selection and a further cost to payload delivery, not detectably size-dependent on flights and size-dependent on hotels: taking the filter off the model is necessary, not sufficient, and only the typed channel holds in both domains. Practitioners adopted it for rendering; it was also, unknowingly, a correctness decision.

Limitations. Six routes and one cabin bound the size manipulation; results come from one deployment with ground truth sharing its evaluator; the hotel replication uses deep-inventory cities (Appendices I, E.1, B).

AI USE STATEMENT

In this work, we used generative AI tools (Claude, Anthropic; and, as the system under study, Gemini 2.5 Flash and the other model families named in §5.3) for the following tasks with required disclosure: drafting and revising the paper's prose, implementing the replay harness and analysis code, executing experiment runs, and retrieval and discovery of related work. We have not used generative AI tools to generate data or ground truth: every ground truth is computed by the production evaluator over frozen real vendor payloads, and every dialogue script was authored or approved by the authors. Additionally, we used generative AI tools for research ideation under the authors' direction. We have reviewed all AI-assisted work: the study design (the conditions, the size and prefix manipulations, the dialogue scripts, the fidelity metric and its zero-out policy, and every ground-truth semantics decision) was specified and adjudicated by the authors, who personally audited the exhibits, per-turn tables and failure cases reported here; AI-generated code was verified by test suites and by replaying frozen captures byte-for-byte; AI-drafted claims were checked against committed result files, from which the analysis layer reproduces every published number; and AI-retrieved references were verified against the cited sources. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The study involves no human subjects and no production user traffic: every dialogue is an authored script, and no personal data appears in captures or traces. Vendor payloads were obtained through the production system's own search tools under its vendor agreements, are stored frozen for replay, and are not redistributed, since vendor terms preclude releasing them. The authors are affiliated with the company that operates the deployed system under study, a conflict of interest we state here; the system's name is withheld for anonymous review and given at camera-ready. The paper's practical recommendation concerns software architecture and carries no foreseeable harmful application.

REPRODUCIBILITY STATEMENT

The camera-ready artifact releases the dialogue templates with their knob specifications and the scope-semantics rulebook (§4.2, Appendix C); the replay harness, including the four-delta SYNTHESIS construction and the REEMIT and REFPICK builders (Appendix C) and the model-override layer implementing sampling parity and the top_p ablation (Appendix E.1); the analysis layer, which reproduces every published number from committed result files; capture manifests recording the producing git SHA, request parameters and payload content hashes; per-run traces with vendor payload contents stripped or hashed (tool calls, arguments, token usage, per-turn scores; payload identity is derivable from the dialogue specifications and capture manifests, and no content hash is in the traces themselves); and the internal pre-registration text with its commit hashes. The frozen payloads themselves cannot be released under vendor terms, so reproduction on our exact data requires the original payloads; the method replicates on a fresh capture or on another deployment's own data with the harness alone. Runs are unseeded repeats at production sampling on a non-bit-deterministic serving stack, so results reproduce distributionally rather than turn-for-turn (Limitations).

REFERENCES

ADK contributors. ADK issue #3969: Output schema consistently ignored by LLM agent with tools. GitHub issue, opened 18 December 2025, 2025. URL https://github.com/google/adk-python/issues/3969.

ADK contributors. ADK issue #5926: AgentTool with input_schema and LiteLLM/Claude skips inner agent tool calls. GitHub issue, opened 1 June 2026, 2026. URL https://github.com/google/adk-python/issues/5926.

Anthropic. Code execution with MCP: Building more efficient agents. Engineering blog post, 4 November 2025, 2025. URL https://www.anthropic.com/engineering/code-execution-with-mcp.

Luca Beurer-Kellner, Beat Buesser, Ana-Maria Creţu, Edoardo Debenedetti, Daniel Dobos, Daniel Fabian, Marc Fischer, David Froelicher, Kathrin Grosse, Daniel Naeff, Ezinwanne Ozoani, Andrew

Paverd, Florian Tramèr, and Václav Volhejn. Design patterns for securing LLM agents against prompt injections, 2025. URL https://arxiv.org/abs/2506.08837.

Hongliu Cao, Ilias Driouich, and Eoin Thomas. Beyond task completion: Revealing corrupt success in LLM agents through procedure-aware evaluation, 2026. URL https://arxiv.org/abs/2603.03116. Unreviewed preprint; τ-bench mocks with a simulated user.

CopilotKit. Controlled generative UI; AG-UI protocol specification. Online documentation, accessed 28 July 2026, 2026. URL https://docs.copilotkit.ai. AG-UI specification: https://docs.ag-ui.com.

CopilotKit contributors. CopilotKit issue #4935: useFrontendTool render args uses LLM-generated tool call args instead of AG-UI event data. GitHub issue, opened 20 May 2026, 2026. URL https://github.com/CopilotKit/CopilotKit/issues/4935.

CrewAI. CrewAI documentation. Online documentation, accessed 28 July 2026, 2026. URL https://docs.crewai.com.

Yixin Dong, Charlie F. Ruan, Yaxing Cai, Ruihang Lai, Ziyi Xu, Yilong Zhao, and Tianqi Chen. XGrammar: Flexible and efficient structured generation engine for large language models, 2024. URL https://arxiv.org/abs/2411.15100.

Laïla Elkoussy and Julien Perez. AgentLTL: A trace-verification framework for measuring, enforcing, and training procedural compliance in tool-using LLM agents, 2026. URL https://arxiv.org/abs/2607.02599.

Cosimo Galeone, Minsu Park, Giuseppe Ettorre, and Daniele Ligorio. When correct isn't usable: Improving structured output reliability in small language models, 2026. URL https://arxiv.org/abs/2605.02363.

Saibo Geng, Hudson Cooper, Michał Moskal, Samuel Jenkins, Julian Berman, Nathan Ranchin, Robert West, Eric Horvitz, and Harsha Nori. JSONSchemaBench: A rigorous benchmark of structured outputs for language models, 2025. URL https://arxiv.org/abs/2501.10868. Cited for schema-compliance measurement only.

Google. ADK samples: travel-concierge. GitHub repository google/adk-samples (renamed google/adk-recipes), 2025. URL https://github.com/google/adk-recipes.

Google. A2UI v0.9. Open specification, GitHub repository, 2026. URL https://github.com/google/A2UI.

Shengyue Guan, Jindong Wang, Jiang Bian, Bin Zhu, Jian-guang Lou, and Haoyi Xiong. Evaluating LLM-based agents for multi-turn conversations: A survey, 2025. URL https://arxiv.org/abs/2503.22458.

Bhaskar Gurram. Auditing automated evaluation, error propagation, and runtime mitigation in tool-using language agents, 2026. URL https://arxiv.org/abs/2604.16706. Unreviewed preprint; injected parameter errors on real execution traces.

Tung Thanh Hoang. Structured hallucination in tool-using agents: Measuring and mitigating LLM synthesis corruption in production. In *Second Workshop on Agents in the Wild: Safety, Security, and Beyond*, 2026. URL https://openreview.net/forum?id=CaJudGHxoR. AIWILD @ ICML 2026.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Delong Chen, Wenliang Dai, Ho Shu Chan, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 2023. URL https://arxiv.org/abs/2202.03629. arXiv:2202.03629.

Ivan Kartáč, Mateusz Lango, and Ondřej Dušek. Reasoning gets harder for LLMs inside a dialogue. In *Proceedings of ACL 2026*, pp. 12263–12303, 2026. URL https://aclanthology.org/2026.acl-long.560.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. LLMs get lost in multi-turn conversation, 2025. URL https://arxiv.org/abs/2505.06120.

Philippe Laban, Tobias Schnabel, and Jennifer Neville. LLMs corrupt your documents when you delegate, 2026. URL https://arxiv.org/abs/2604.15597.

LangChain. LangGraph documentation: custom stream mode. Online documentation, accessed 28 July 2026, 2026. URL https://langchain-ai.github.io/langgraph/concepts/streaming/.

Microsoft. Microsoft agent framework documentation: workflow events. Online documentation, accessed 28 July 2026, 2026. URL https://learn.microsoft.com/en-us/agent-framework/.

OpenAI. OpenAI responses API documentation. Online documentation, accessed 28 July 2026, 2026. URL https://platform.openai.com/docs/api-reference/responses.

Shishir G. Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. Gorilla: Large language model connected with massive APIs. In *NeurIPS 2024*, 2024. URL https://arxiv.org/abs/2305.15334. arXiv:2305.15334.

Stephan Rabanser, Sayash Kapoor, Peter Kirgis, Kangheng Liu, Saiteja Utpala, and Arvind Narayanan. Towards a science of AI agent reliability. In *ICML 2026*, 2026. URL https://arxiv.org/abs/2602.16666. arXiv:2602.16666.

Jaideep Ray. The constraint tax: Measuring validity-correctness tradeoffs in structured outputs for small language models, 2026. URL https://arxiv.org/abs/2605.26128.

Vikas Reddy, Sumanth Reddy Challaram, and Abhishek Basu. Reason less, verify more: Deterministic gates recover a silent policy-violation failure mode in tool-using LLM agents, 2026. URL https://arxiv.org/abs/2607.07405. Venue self-declared and unverified.

Maximilian Schall and Gerard de Melo. The hidden cost of structure: How constrained decoding affects language model performance. In *Proceedings of RANLP 2025*, pp. 1074–1084, 2025. doi: 10.26615/978-954-452-098-4-124. URL https://aclanthology.org/2025.ranlp-1.124.

Ferdinand M. Schessl. The autocorrelation blind spot: Why 42% of turn-level findings in LLM conversation analysis may be spurious, 2026. URL https://arxiv.org/abs/2604.14414.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools, 2023. URL https://arxiv.org/abs/2302.04761.

Akshit Sinha, Arvindh Arun, Shashwat Goel, Steffen Staab, and Jonas Geiping. The illusion of diminishing returns: Measuring long-horizon execution in LLMs, 2025. URL https://arxiv.org/abs/2509.09677.

Harsh Soni. ToolFailBench: Diagnosing tool-use failures in LLM agents. In *Workshop on Agents in the Wild (AIWILD) and Workshop on Failure Modes of Agentic AI (FAGEN) @ ICML 2026*, 2026. URL https://arxiv.org/abs/2607.04686. arXiv:2607.04686.

Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung-yi Lee, and Yun-Nung Chen. Let me speak freely? A study on the impact of format restrictions on performance of large language models. In *EMNLP 2024 Industry Track*, 2024. URL https://arxiv.org/abs/2408.02442. arXiv:2408.02442.

Vercel. Vercel AI SDK documentation: Data parts. Online documentation, accessed 28 July 2026, 2026. URL https://ai-sdk.dev/docs/ai-sdk-ui/streaming-data.

Xingyao Wang, Yangyi Chen, Lifan Yuan, Yizhe Zhang, Yunzhu Li, Hao Peng, and Heng Ji. Executable code actions elicit better LLM agents. In *ICML 2024*, 2024a. URL https://arxiv.org/abs/2402.01030. arXiv:2402.01030.

Xingyao Wang, Zihan Wang, Jiateng Liu, Yangyi Chen, Lifan Yuan, Hao Peng, and Heng Ji. MINT: Evaluating LLMs in multi-turn interaction with tools and language feedback. In *ICLR 2024*, 2024b. URL https://arxiv.org/abs/2309.10691. arXiv:2309.10691.

Brandon T. Willard and Rémi Louf. Efficient guided generation for large language models, 2023. URL https://arxiv.org/abs/2307.09702.

Sam Wiseman, Stuart Shieber, and Alexander Rush. Challenges in data-to-document generation. In *Proceedings of EMNLP 2017*, 2017. doi: 10.18653/v1/D17-1239.

Weimin Xiong, Ke Wang, Yifan Song, Hanchao Liu, Sai Zhou, Wei Peng, and Sujian Li. More vulnerable than you think: On the stability of tool-integrated LLM agents, 2025. URL https://arxiv.org/abs/2506.21967.

Yuqing Yang, Qi Zhu, Zhen Han, Boran Han, Zhengyuan Shen, Shuai Wang, Vassilis N. Ioannidis, and Huzefa Rangwala. When LLMs read tables carelessly: Measuring and reducing data referencing errors. In *Proceedings of ACL 2026*, 2026. URL https://arxiv.org/abs/2606.32029. arXiv:2606.32029.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models, 2023. URL https://arxiv.org/abs/2210.03629.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains, 2024. URL https://arxiv.org/abs/2406.12045.

Yuqi Zeng, Qixiang Deng, Yulei Wan, Ruiquan Jiang, Xiaoqing Zheng, and Xuanjing Huang. Rethinking agentic RAG: Toward LLM-driven logical retrieval beyond embeddings, 2026. URL https://arxiv.org/abs/2605.27123.

Dongsheng Zhu, Xuchen Ma, Yucheng Shen, Xiang Li, Yukun Zhao, Shuaiqiang Wang, Lingyong Yan, and Dawei Yin. When tools fail: Benchmarking dynamic replanning and anomaly recovery in LLM agents (ToolMaze), 2026. URL https://arxiv.org/abs/2606.05806.

Kunlun Zhu, Zijia Liu, Bingxuan Li, Muxin Tian, Yingxuan Yang, Jiaxun Zhang, Pengrui Han, Qipeng Xie, Fuyang Cui, Weijia Zhang, Xiaoteng Ma, Xiaodong Yu, Gowtham Ramesh, Jialian Wu, Zicheng Liu, Pan Lu, James Zou, and Jiaxuan You. Where LLM agents fail and how they can learn from failures, 2025. URL https://arxiv.org/abs/2509.25370.

A THE SIZE-MANIPULATION DIALOGUES AND PER-ROUTE LADDERS

The manipulation should be checkable rather than taken on trust, so the dialogue set is reproduced here. The tables below describe the original d25–d30 six-dialogue design, which sits within the larger 20-dialogue ladder of §5.1. Every dialogue has the same six-turn shape (search, two scored filter turns, second search, two more scored filter turns) with **turns 2, 3, 5 and 6 scored and byte-identical across all six dialogues** (Table 4); the only textual difference is the date in the two search turns, the manipulation instrument, because the capture date determines how large the delivered table is (Table 3). Each dialogue contains one **treatment leg** (size varies) and one **control leg** (always the same `close_in` capture: 25 delivered rows, 16 after the seat filter); the `_early` / `_late` suffix counterbalances which leg comes first in the dialogue.

Turn 2/5 applies `min_seats=2`; turn 3/6 adds the published afternoon bucket (12:00–17:59) and composes with the previous turn. Ground truth is computed by the production filter evaluator over the frozen payload, never annotated (Table 5).

The control leg is identical in all six dialogues (16 $\rightarrow$ 5 rows): synthesis runs whose 194-row treatment turn earned zero credit (one via an unscoreable emission) still average 90% (9/10 control turns) on the byte-identical 16-row control leg of the same dialogues, which is what separates delivered-table load from context accumulation. Every scored turn passed an oracle liveness gate before any model was run (empty or trivially-repeated ground truth is rejected at authoring time), and the authored `expected_gt_rows` is asserted against the evaluator's `total_matched` at run time, so payload drift cannot silently rescale the treatment variable.

dialogue	treatment size	leg order	first search date	second search date
d25_fra_bkk_size_s_early	9 rows	treatment first	2027-06-24 (release_edge)	2026-08-14 (close_in)
d26_fra_bkk_size_m_early	74 rows	treatment first	2026-10-14 (near)	2026-08-14 (close_in)
d27_fra_bkk_size_l_early	194 rows	treatment first	2027-04-14 (far)	2026-08-14 (close_in)
d28_fra_bkk_size_s_late	9 rows	control first	2026-08-14 (close_in)	2027-06-24 (release_edge)
d29_fra_bkk_size_m_late	74 rows	control first	2026-08-14 (close_in)	2026-10-14 (near)
d30_fra_bkk_size_l_late	194 rows	control first	2026-08-14 (close_in)	2027-04-14 (far)

Table 3: The six size-manipulation cells. Sizes are §5.1's design labels (post-seat-filter counts); the delivered payloads are 17, 103 and 198 rows (control 25).

t	speaker turn	role	ground truth
1	"Find award flights from FRA to BKK on 2027-06-24, economy class"	search (establishes context)	—
2	"At least 2 seats please"	**scored** · fresh filter	9 rows
3	"Only afternoon departures"	**scored** · composes with previous	3 rows
4	"Find award flights from FRA to BKK on 2026-08-14, economy class"	search (establishes context)	—
5	"Same here — at least 2 seats"	**scored** · fresh filter	16 rows
6	"Afternoon departures only, please"	**scored** · composes with previous	5 rows

Table 4: The dialogue script, shown for d25 (treatment first); the other five differ only in the two dates; their ground-truth counts differ accordingly (Table 5).

A.1 The per-route date-varied ladders (§5.1)

Each route is a single-route, single-cabin ladder varying delivered size by capture date; Figure 3 plots these cells (Table 6).

A.2 The top-K ladder cells

The five pooled cells of the content-controlled (top-K) ladder (§5.1); these are exactly Figure 2's plotted points (Table 7).

Failure anatomy on the content-controlled (top-K) ladder, FRA-BKK. Three failures were residual starvation (the search sub-agent reported success but its inner search never delivered rows, Appendix E; one in the full-payload K=198 cell, outside this table's K range), and the mid-K failures are genuine truncations whose self-reported counts exceed their emitted rows by one (claimed 12, emitted 11; claimed 19, emitted 18). The K=100 collapse is starvation-free: all three FRA-BKK runs received their 100 rows and truncated their answers (20/30/34 rows emitted against a ground truth of 42).

B The Demonstration-versus-Instruction Pilot (H2)

During earlier engineering work, seeding a session with an *answer-shaped* model turn ("Found 40 award options …" followed by a rendered table) made the agent pattern-complete the next filter request without calling the filter tool, fabricating a count (40→10, 7→2, 2→1 across three probes). A local A/B on one route, cabin and date (5 runs per variant): an answer-shaped seed with an explicitly strengthened system-prompt prohibition produced 0/5 grounded turns; a tool-trajectory seed produced 5/5 under either prompt. n **is small, uncounterbalanced, one route and one model: not a controlled experiment**, motivating only the hypothesis that in-context demonstration governs grounding more strongly than an on-point prompt prohibition. The pre-registered design, a 2×2 of seed shape $\times$ prompt strength over $\sim$30 stratified seeded contexts $\times$ 5 runs per cell, is future work.

C Method and Analysis Detail

C.1 Scope semantics, the metric's definition, and the hypotheses, in full

A third convention pins down **scope semantics**, since every turn's ground truth depends on what a follow-up operates over: a new filter **composes** with the active scope ("only morning departures"

dialogue	treatment leg	control leg
d25 (S, early)	9 → 3	16 → 5
d26 (M, early)	74 → 11	16 → 5
d27 (L, early)	194 → 87	16 → 5
d28 (S, late)	9 → 3	16 → 5
d29 (M, late)	74 → 11	16 → 5
d30 (L, late)	194 → 87	16 → 5

Table 5: Ground-truth row counts per cell.

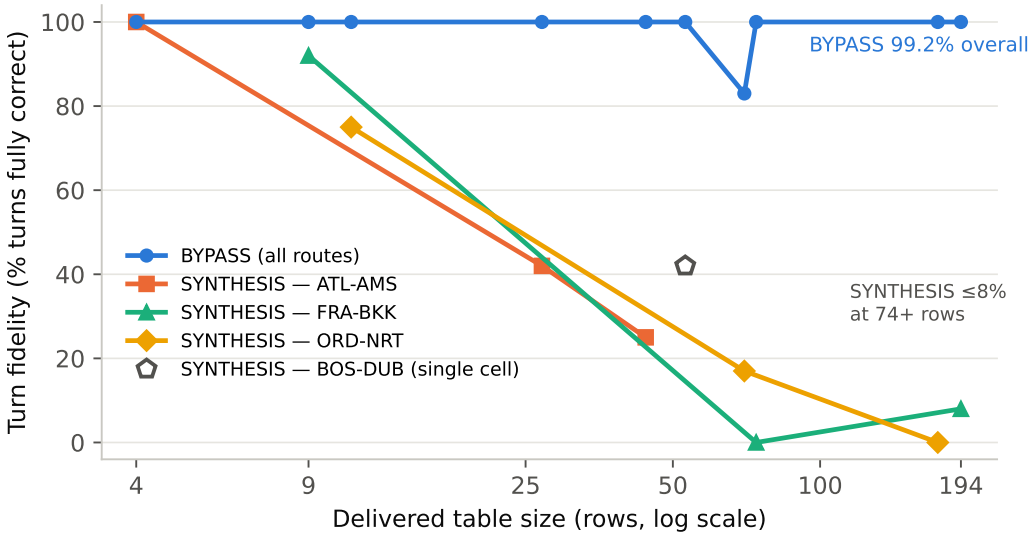


Figure 3: Turn fidelity vs table size. SYNTHESIS falls from 75–100% on the smallest tables to ≤8% beyond 74 rows in every route that reaches that size (one 0%→8% wobble at FRA-BKK's largest cell); BYPASS is 100% everywhere except a single dip (70 rows, 83%), one starved run (Appendix E); 99.2% overall.

narrows the current table); a refer-back **restores** the view where its referent was introduced ("show all the morning ones again" returns to the table that named "the morning ones", dropping filters and limits applied after it); and a constraint revision **re-executes** the turn that set the constraint against that turn's own base ("actually, up to 87,499 points is fine" re-derives the cap from the full table, not the narrowed one; the deliberately odd cap is a pinned boundary probe, one point below the payload's real 87,500 value). A turn meaning to reset must name its base explicitly ("across ALL results…").

The primary metric's policy is adopted from the single-turn study (Hoang, 2026); field-level scoring against source records descends from data-to-text evaluation (Wiseman et al., 2017).[2]

Partial-credit fidelity (the field-match fraction without the zero-out: **98.2% for BYPASS vs 82.7% for SYNTHESIS**) and the ordered-turn axis (order-exact: **94.2% vs 51.8%**) are computed alongside every turn and reported in full in Appendix C.

Hypotheses. **H1a**: the fidelity gap appears at the first filter turn. **H1b**: synthesis failures are not independent across turns; history predicts failure. **H1c**: synthesis fidelity degrades with the size of the delivered record set. A turn-count **super-linearity** conjecture (failures compound faster with depth than an independence null) was pre-registered but unsupported, published as a diagnosed negative result (§5.6). The size instruments (§5.1, §5.3, Appendix E) were designed after the main

[2]A post-submission instrument audit found that this metric can score a turn as perfect when the emitted set has the right count but the wrong membership, since field fidelity is computed over aligned pairs only. Exact-set rescoring of the size ladder (§5.1) leaves BYPASS unchanged and lowers the in-head conditions by 4 to 7pp; the headline corpus retains its original scoring.

route	rows → SYNTHESIS acc.	BYPASS
FRA-BKK	$9 \to 92\% \cdot 74 \to 0\% \cdot 194 \to 8\%$	100%
ORD-NRT	$11 \to 75\% \cdot 70 \to 17\% \cdot 174 \to 0\%$	100%*
ATL-AMS	$4 \to 100\% \cdot 27 \to 42\% \cdot 44 \to 25\%$	100%
BOS-DUB	$53 \to 42\%$	100%

Table 6: Date-varied ladders. *ORD-NRT's BYPASS column excludes one starved run (10/12 at the 70-row cell, the single dip visible in Figure 3); with it, 99.2% overall (238/240 scored turns). Figure 4's n row shares this accounting: the starved run's two turns count in the $\leq$16 band at their delivered size of 0 rows, and two SYNTHESIS turns (d27 sweep k2 and d33 sweep k3, the 194- and 174-row cells; delivered payloads 198 and 190 rows) emitted nothing scoreable and drop from the $>$100 count.

K delivered	5	10	25	50	100
SYNTHESIS	83% (15/18)	72% (13/18)	44% (8/18)	39% (7/18)	**6%** (1/18)
BYPASS	100% (18/18)	100% (18/18)	100% (18/18)	100% (18/18)	100% (18/18)

Table 7: Content-controlled prefix ladder pooled over six routes. FRA-BKK's single-route full-payload cell (K=198, n=3: SYNTHESIS 0/3, BYPASS 3/3) is not pooled here; it completes §5.1's 93-run BYPASS total.

corpus's first-turn rates were observed, each pre-specified (cells, byte-identical scored turns, expected ground-truth counts) before its runs. A second family, **H2** (demonstration-versus-instruction), was stated as motivation and not measured (Limitations; Appendix B). Verdicts: **H1c supported** as a controlled manipulation (§5.1); **H1a supported** (§5.2); **H1b supported** (§5.6).

C.2 The undefined-aggregate rule, worked

"Which day is cheapest?", asked after a date-*range* search whose result spans several days, names no total order: ranking days requires a day-level aggregate over per-flight prices (the day's minimum fare? its mean? award points or cash?) that the utterance never picks, so several answers are defensible and the turn is unscoreable by construction. The §4.2 rule retired exactly such a turn, which had shipped in five dialogues; its replacement is an unambiguous window-spanning filter turn ("Morning departures across all the days"; Appendix E).

Thresholds are authored as payload-pinned percentiles with a lint asserting 20–80% survival; boundary probes are pinned strictly between two adjacent real values, so strict-versus-inclusive readings cannot both be defensible.

C.3 The oracle's read path

Because the deployed system serves large result sets by reference (rows in state, envelope carrying a pointer), the oracle reads the full served row set through the same state path the client's stream uses, rather than trusting the envelope's inline rows. The liveness verifier of §4.3 asserts, before any model is called, that no turn's ground truth is empty or a trivial repeat of the previous turn's, on every turn of every dialogue. Three properties keep the oracle independent of the arm that shares its code. First, it executes the authored arguments, pinned in the dialogue spec, never the model's: BYPASS's own translations are scored against it and fail when the divergence changes the selected rows (22 wrong-argument failures; Appendix F.10), so the deployed arm is not correct by construction. Second, the authored `expected_gt_rows`, derived by hand from each frozen payload, is asserted against the evaluator's `total_matched` on every scored turn of every dialogue, across all turn families (Appendix C.11). Third, the evaluator is challenged rather than trusted: row-level hand audits cover all 99 ladder failures and both exhibits (§H.1, Appendix D), and the Appendix E challenge found three ground-truth defects, fixed with both arms re-scored. We use the production evaluator because it implements exactly what the oracle needs, a deterministic filter over frozen state under the §4.2 semantics; a second implementation built only for scoring would add disagreements needing their

Component	SYNTHESIS	BYPASS
Translate (NL → filter args)	model *(identical)*	model *(identical)*
Record storage	model context, re-emitted every turn	session state, written once
Follow-up filter	in-context, by the model	deterministic evaluator
Delivery to UI	parsed from model output	typed side channel

Table 8: What differs between the conditions. The headline gap measures the joint effect of the last three rows; §6 unbundles the filter and delivery factors with two further conditions.

own adjudication without removing the shared-specification risk. An independent re-implementation of the §4.2 semantics is the check this design still lacks (Limitations).

C.4 Partial credit and order, in full

Partial-credit fidelity over aligned pairs, post-fix: **98.2% for BYPASS and 82.7% for SYNTHESIS** (mean over scored turns with an emitted artifact, n=338 and n=333; pre-fix: 93.3% and 72.8%), a 15.5-point gap against the zero-out headline's 46.6.

Turns flagged `ordered` (limit-N superlatives and explicit re-sort requests) are scored on emitted sequence as **order-exact** (the emitted sequence equals the ground-truth sequence) and **order agreement** (the fraction of positions holding the correct record). On the post-fix corpus's 86 (BYPASS) and 85 (SYNTHESIS) ordered turns: order-exact 94.2% vs 51.8%, mean order agreement 95.9% vs 58.2% (pre-fix: 93.7% vs 61.2% and 94.2% vs 64.0%).

C.5 The analysis layer's instruments

The corpus module's inclusion rules: three independent sweeps per condition (the k=3 repeats); the range/flex arc taken only from its corrected re-run (Appendix E), the defective original excluded in *both* conditions; recovery re-runs may fill a dialogue lost to a network error; a run producing no scoreable output is scored as zeros on its expected filter turns.

Beyond the aggregates named in §4.5, the analysis layer computes per-position marginal hazard; cumulative failure versus the accumulation-only null $1-(1-p)^n$; and a per-dialogue **counterfactual ratio**: observed failures per dialogue divided by what an independent-failures null with the same per-turn rate predicts: 1.0 means failures land as if independent, above 1 means they cluster (0.95× for SYNTHESIS; Appendix C.12).

Value-tracing labels. A failed turn is *traceable* when a comparable row set exists for it (SYNTHESIS: the model-emitted table; BYPASS: the typed side-channel artifact; turns with no scoreable artifact are grounding failures, excluded). A traceable failure is `propagated` when its emitted rows are consistent with filtering a prior turn's emission rather than the ground truth (set equality or subset over the identity keys of prior emissions), and `independent` when the error is fresh. When both explanations fit, the labeler resolves to `propagated`, so the propagated share is an upper bound; for BYPASS no model-emitted set exists, current or prior, so its zero is structural rather than a survived test.

C.6 The depth test's degenerate null

With SYNTHESIS's 56.9% first-turn failure rate, the accumulation-only null $1-(1-p)^n$ saturates at 81% by position 2 and 92% by position 3, leaving at most 8–19pp of measurable headroom above the null: super-linearity of any plausible magnitude is undetectable in this regime. SYNTHESIS's marginal hazard is flat with depth (57/58/60/39/55% at positions 1–5); BYPASS's position-1 hazard is 0.0% on the main 24-dialogue corpus: all 26 of its failed turns (§5.4) sit at later positions.

Frozen captures / records	408 / 9,809
Routes	36
Dialogues	24
Runs per condition	72 (24 dialogues $\times$ k=3)
Scored turns per condition	339
of which turn-matched	264
Conditions	2
Models	1 (Gemini 2.5 Flash)

Table 9: Main-study scale. Capture, record, and route totals are the full frozen corpus, which also serves the size and prefix instruments (Appendix A); further families replicate at the same 339-turn scale (§5.3, Appendix F.6). The headline is the full 339 scored turns per condition; the 264 turn-matched turns (identical turn sets in both instrument versions' corpora) isolate the instrument effect (Appendix E).

C.7 THE SYNTHESIS CONSTRUCTION'S CONSEQUENCES, IN FULL

- **Translation error is common-mode.** The same planner, with the same instruction lineage, decides what the user asked for in both conditions, so the residual gap is attributable to delivery.
- **Mechanical delivery makes SYNTHESIS stronger than its fork-true form.** In the real fork-point system the sub-agent's records pass through an LLM emission before reaching the planner; we remove that step, so our SYNTHESIS is a charitable version of the architecture it stands for.
- **BYPASS's advantage is joint, and we name it.** BYPASS benefits jointly from synthesis bypass (records never re-typed) and structured-state preservation (each follow-up recomputes from the original set). §6 separates them; the deployed system does both, and no deployed pattern we know of does one without the other, so the comparison is the pair, not a single-factor ablation.

C.8 CORPUS COMPOSITION AND CAPTURE PROVENANCE

The 408 captures decompose as 270 fixed-date award, 18 award date-range, and 120 cash. The four date bands are **close-in** ($\approx$3 weeks out), **near** ($\approx$3 months), **far** ($\approx$9 months), and **release-edge** ($\approx$11 months, the edge of the booking window where award seats are first released). Captures are restricted at capture time to non-stop and one-stop itineraries with the pre-filter count recorded in each manifest, so the restriction is never a silent truncation; each capture records the git SHA that produced it, the request, and per-call vendor traffic.

C.9 SCALE

Table 9 reproduces the study's scale.

C.10 THE FOUR DELTAS, IN FULL

The production agent tree is deep-copied in memory with exactly four deltas: (1) the planner's instruction replaced by a variant walked block-by-block against production, removing only blocks that presuppose the context tool and client-side rendering; this variant, with the sampling inherited unchanged (§3.2), is what the abstract's "lineage-matched" names: the same production lineage in both arms, differing only in delivery-entailed blocks; (2) the deterministic context tools removed from the planner's toolset; (3) each flight-search sub-agent given a callback replacing its LLM-generated summary with `{total_found, search_params, records}` read verbatim from state, so the row list is delivered *mechanically*; (4) those clones dropping their output schema, which would otherwise strip the injected records. Neither condition sets an output-token cap: the only generation parameters configured in either arm are temperature 0.1 and top_p 0.5 (§3.2; Claude inherits temperature only, Appendix E); provider defaults apply identically across arms within each family. Each delta is entailed by the swap: in-context delivery cannot coexist with a deterministic context tool, with instruction blocks that presuppose it, or with output schemas that strip the injected

rows; the introduction's "row path replaced" quantifies over this one variable, implemented as these four deltas.

C.11 Dialogue inventory

Six arcs cover narrow-and-pick, award/cash comparison, release-edge with truthful empty results, range/flex, multi-table refer-back, and compound power-user filters; turn families span the filter vocabulary: time buckets, day-rollover arrival filters, seat and program constraints, superlatives with sort and limit, negated predicates, refer-back, and cumulative narrowing. The four variation knobs are payload swap, order permutation, follow-up addition or omission, and phrasing alternates. Dialogue identifiers encode route and design cell (`d02_lhr_fco_close`, `d26_fra_bkk_size_m_early`; Appendix A).

C.12 Operational detail moved from the body

The fifteen scored fields are flight number, airline, booking source, mileage cost, taxes, cabin, date, origin, destination, departure and arrival timestamps, duration, stops, remaining seats, and cash price. `source` is included in the identity key because the same flight legitimately appears once per booking program. The much smaller partial-credit gap is the expected signature: SYNTHESIS frequently emits a *nearly*-right set, which is still a wrong answer to a filter question. Turns flagged `ordered` (limit-N superlatives and explicit re-sorts) are scored on emitted sequence, reported as their own axis and never folded into fidelity; mean order agreement is 95.9% vs 58.2% (Appendix C.4).

Every run persists a readable trace and a complete raw record (tool calls, model texts, token usage, final state, per-turn scores). The analysis layer computes dialogue-level bootstrap CIs (2000 resamples, dialogue as the unit, following Schessl, 2026) plus per-position and lag-1 hazards, cumulative failure versus the accumulation null, accuracy by table size, and runs a value-tracing labeler that marks each traceable failure `independent` or `propagated` (definitions in Appendix C.5).

§5.6's cut detail, with Appendix C.5's test construction: SYNTHESIS's excess over the accumulation-only null is negative at every position (-14.8 to -22.3pp at positions 2–5) and its failures per dialogue land at $0.95\times$ the per-dialogue independence prediction (Appendix C.5); lag-1 samples are $n{=}146/121$ (SYNTHESIS, 71.9% vs 31.4%) and $n{=}23/244$ (BYPASS, 4.3% vs 10.2%; the nominal inversion, a lower hazard after a failure, is a single-failure cell, 1/23, consistent with independence rather than any protective effect).

D The Exhibits and the Fabrication Census in Full

D.1 The census decomposition

Of 1,275 context-populated emitting turns (both corpora, production model), 1,245 emit only delivered rows; the 30 exceptions decompose into identifier-level corruption of real rows (25 turns: connecting segments of multi-leg itineraries dropped or recombined, or a single garbled code, inside an otherwise-real table; e.g., a real "…, BR62" tail swapped to "BR68", one run splicing 24 of its 103 emitted rows) and five turns in two pre-fix multi-search dialogues where the *queried* award table itself had never been delivered (the context held only cash rows) and the model produced two-row placeholder tables (`UA123`, `JL789`, `KL456`) for the missing set. Every from-nothing fabrication in either corpus sits over a missing queried table. All exhibits are schema-valid SYNTHESIS turns, and the census finds no third variant in which a delivered table is present and a fabricated one replaces it.

D.2 Exhibit 1: fabrication under grounding absence (the plausible face)

Dialogue `d02_lhr_fco_close`, sweep k1. The frozen payload is LHR→FCO on 2026-08-14, economy: **48 real records** whose only mileage values are {7500, 14000, 15000, 22500, 25500, 28200, 47500} and whose booking sources are {aeroplan, alaska, delta, flyingblue, qantas, virginatlantic}. The *readiness gate* is a deterministic pre-search check that all required slots are filled; its "Ready" verdict licenses the search dispatch. In this run it returned "Ready", but the search sub-agent was never invoked (the pre-fix starvation event, Appendix E), so the context held no award records at any

t	user turn	active scope	GT	emitted
1	"Find award flights from LHR to FCO…"	*search: should deliver 48 rows; never fires*	*not scored*	—
2	"Show me ones less than 10,800 points"	$<$10,800 pts	27	2
3	"Only morning departures please"	+ morning (composes)	19	2
4	"Of those, at least 2 seats"	+ $\geq$2 seats (composes)	17	2
5	"Give me the cheapest one"	cheapest of the 17	1 (of 17)	1
6	"Show all the morning ones again"	back to t3's full morning table	19	2
7	"And the 3 earliest of those"	3 earliest of those 19	3 (of 19)	2

Table 10: Exhibit 1 (`d02_lhr_fco_close`, sweep k1): a starved SYNTHESIS run fabricates a two-row table and then filters the fiction. GT is the oracle's expected emission per turn; "emitted" is what the model displayed.

turn. Asked at turn 2 for "ones less than 10,800 points" (ground truth 27 rows, all at exactly 7,500), SYNTHESIS emitted (identity fields shown; the records' other fields elided):

```
{"flights": [
 {"flight_number": "BA234",
  "mileage_cost": "10000",
  "source": "avios"},
 {"flight_number": "VS123",
  "mileage_cost": "9500",
  "source": "virginatlantic"}],
 "summary": "I found 2 award flights for less than
  10,800 points."}
```

None of it exists in the payload: not the flight numbers, not the mileage values, not the source `avios`; `schema_parsed: true`. The fabrications are *plausible*: under the user's cap, on carriers that credibly fly London→Rome, with textbook flight numbers: nothing distinguishes the reply from a genuine two-row result. And the fiction propagates: the remaining turns filter, sort and re-emit the invented rows. In Table 10 (and Table 11), each row's **GT is the correct table *after executing* that turn's request** under the scope shown, computed by the oracle from the frozen payload and independent of what the run did.

Fidelity is zero for the rest of the dialogue because the model is filtering fiction; on the same starved dialogue, BYPASS emits zero rows, a visible failure.

D.3 Exhibit 2: fabrication under starvation

Dialogue `d23_jfk_lhr_relax_close`, sweep k2. At turn 1 the readiness gate returned "Ready" but the award search sub-agent was never invoked, so **no records ever entered the context**: the frozen payload this dialogue should have delivered holds 74 rows (the same no-search event appears in BYPASS runs; see Appendix E). Asked at turn 2 for flights under 30k points, SYNTHESIS made zero tool calls and emitted two placeholder-shaped records (`UA123 / 25000 / united_miles`, `DL789 / 28000 / delta_skymiles`): round point values, flight numbers absent from the frozen payload, non-production source strings. Asked at turn 4 to widen the cap ("Actually, up to 87,499 points is fine!"), the model does not return to any source: it **generates three additional fictional rows** (`BA456 / 35000 / avios`, `VS101 / 40000`, `AA200 / 80000`), one flight per major carrier, **to fill the widened window** (Table 11).

Under the same starvation, BYPASS emits zero rows (visibly empty, so the user retries) where SYNTHESIS emits a complete, plausible, schema-valid table that is entirely false.

E Integrity Detail

Practising the method surfaced defects in our own instrument, all reported: three ground-truth defects in one arc (the only arc synthesis ever won; BYPASS 100% against SYNTHESIS 48% after correction), a pre-fix starvation event whose exclusion widened the pre-fix gap and narrows the post-fix one by 1.7pp, one unexplained -6.0pp SYNTHESIS residual across instrument versions (the

t	user turn	active scope	GT	emitted
1	"Find award flights from JFK to LHR…"	*search: should deliver 74 rows; never fires*	*not scored*	—
2	"less than 30k points"	<30k pts (over the full 74)	23	2 (fabricated)
3	"morning departures"	+ morning (composes)	6	2 (the same fiction re-emitted)
4	"Actually, up to 87,499 points is fine!"	cap **relaxed** to ≤87,499 over the full table (morning no longer active)	44	5 (more fiction)
5	"cheapest one"	cheapest of the 44	1 (of 44)	1

Table 11: Exhibit 2 (`d23_jfk_lhr_relax_close`, sweep k2): starved SYNTHESIS fabricates placeholder rows, then fabricates more to fill a relaxed cap. The t4 view grows past t2's 23 because the user is backing off the budget set at t2: "up to 87,499 points is fine" asks about everything under the new cap, not just the morning subset they had drilled into (the scope semantics of §4.2). Contrast t3, which narrows, and Exhibit 1's t6, which returns to an earlier view.

gap exceeds +33pp under either), and a 3×3 size-by-position Latin square that finds size decisive (p=3×10^{-10}) and no position effect (p=0.36). Narratives, accounting and regressions follow.

E.1 Measurement integrity, body summary as submitted

The range/flex arc initially scored BYPASS 52% against SYNTHESIS 59%, the only arc synthesis won; challenged rather than published, it dissolved into three defects in our own ground truth and reads BYPASS 100% against SYNTHESIS 48% after correction, which is why the oracle's liveness verifier exists. Pre-fix, the search sub-agent sometimes never fired (21.2% of SYNTHESIS search turns, 5.2% of BYPASS's); repairs took the rates to 9.7% and 0.0%, affected turns count as failures, and excluding them widened the pre-fix gap and narrows the post-fix one by 1.7pp. Fixing the instrument moved the headline on the comparable subset from +33.7 to +49.6pp; a −6.0pp SYNTHESIS residual exceeds the sweep-to-sweep spread and is flagged, not absorbed, and under either version the gap exceeds +33pp. A 3×3 size-by-position Latin square finds size at fixed slot decisive (χ^2=39.4, p=3×10^{-10}) and no detectable position main effect (p=0.36), BYPASS at 106/108 (Appendix E).

Measurement integrity, in full

We found three defects in our own ground truth, and report them The range/flex arc initially scored **BYPASS 14/27 (52%) vs SYNTHESIS 16/27 (59%)**, the only arc synthesis won. Challenged rather than published, it dissolved into three instrument defects (full narratives in Appendix E); after correction: **BYPASS 25/25 (100%), SYNTHESIS 12/25 (48%)**, a 59-point swing (from −7pp to +52pp; turn accounting in Appendix C). The §4.3 liveness verifier exists because of this, the corrected arc replaces the defective one in *both* conditions, and the same arc captured a user-facing production bug, fixed (Appendix E).

The starvation event: diagnosed and largely removed In pre-fix runs, the search sub-agent sometimes never fired despite a ready verdict, failing the dialogue wholesale (21.2% of SYNTHESIS search turns, 5.2% of BYPASS's). The mechanisms differ by condition (Appendix E); repairs took the post-fix rates on the main 24-dialogue corpus to **0.0% (BYPASS) and 9.7% (SYNTHESIS)** (starved runs persist on other legs; §5.1), and affected turns are reported as failures, not excluded. Excluding starved runs *widens* the pre-fix gap (re-scored: +35.3 → +37.5pp): the confound suppressed the effect rather than manufacturing it. Post-fix, excluding the 11 of 72 starved SYNTHESIS runs (BYPASS loses none) lifts SYNTHESIS to 47.4% (gap +44.9pp): the headline does not rest on residual starvation.

Instrument fixes: one move unexplained Fixing the instrument moved the headline on the comparable subset (the 19 dialogues whose turn sets are identical under both instrument versions; 264 turns per condition) from +33.7pp to +49.6pp. The BYPASS side is accounted for (removing starvation and the fabrication-teaching prompt lifted it from 81.4% to 91.3%); the SYNTHESIS side is not: two fixes expected to *raise* it by roughly 4pp preceded a fall from 47.7% to 41.7%, a −6.0pp

residual exceeding the pre-fix corpus's own sweep-to-sweep (repeat-to-repeat) spread, which we flag as unexplained rather than absorb (full accounting in Appendix E). We declared the freeze before the final runs and report both readings; under either instrument version the gap exceeds +33pp.

A factorial separates size from position In the 24-dialogue corpus, table size and turn position are correlated, so observational stratification alone cannot separate a size account from a first-turn account. The size ladders (§5.1) counterbalance position, and a **3×3 size-by-position Latin square** settles it: size at fixed position (slot) is decisive (92→33→0% at slot 1; χ^2=39.4, p=3×10^{-10}, surviving cluster bootstrap in 2000 of 2000 resamples) while no position main effect is detectable (χ^2=0.8, p=0.36); BYPASS crosses the same square at 106/108. Regression, design and the secondary texture are in Appendix E.

The three ground-truth defects, in full (1) An unscoreable turn: "Which day is cheapest?" has no defined aggregate (Appendix C) and was retired under the §4.2 rule, replaced with an unambiguous window-spanning filter turn ("Morning departures across all the days"). (2) Stale downstream ground truth: after the replacement, later turns' ground truth was not re-composed, penalising the deployed condition for being right: one turn demanded 64 rows (date only) while BYPASS's 9 (morning ∩ date) was correct under our own scope convention. The arc's scored-turn count drops from 27 to 25 in the correction: for the two payloads with zero morning departures (d16, d17) the replacement turn is itself the seats filter, so each dialogue's later standalone "Of those, at least 2 seats" turn became a no-op duplicate and was removed in the re-composition; both conditions had answered both turns correctly, so the drop favours neither side. (3) A hardcoded day pick (range-start + 2) on one payload selected a day with zero morning departures, unscoring every downstream turn.

The production bug the same arc captured: a date-window argument the context tool did not accept was silently dropped by the framework, so a user asking for a week received one day (2 rows instead of 6); post-fix the same turns return 467 / 6 / 6 rows, exact. A third production fix from the same study: duplicate cash itineraries had been reaching users (61 rows across 21 captures; now zero). The two ground-truth semantics corrections are §4.2's refer-back and constraint-revision scope rulings.

The starvation mechanisms For BYPASS, the production prompt's worked examples taught the model to answer instead of searching; the fabrication fix (§H.1) repaired it, taking the rate on affected dialogues from 14.2% to 1.4% at the time of the production fix (on the post-fix main corpus the BYPASS rate is 0.0%; Appendix E). For SYNTHESIS, `output_schema` biases the planner toward emitting a schema-valid "searching…" object in place of the tool call, an ADK-level behaviour that an ordering instruction mostly but not fully counteracts; the residual post-fix 9.7% is schema pressure winning over the instruction. The schema-versus-tool-call tension is documented in the framework's own tracker: declared output schemas are unreliably honoured when tools are present (ADK contributors, 2025), and an input-schema variant suppresses an inner agent's tool calls entirely on the LiteLLM/Claude path (ADK contributors, 2026).

The position question: the three instruments' design The size ladders (§5.1) counterbalance position: each size band appears at both an early and a late position. An observational regression over all 527 scored SYNTHESIS turns of the main 24-dialogue corpus and the size ladders (logistic, correct ∼ $\log_2$(size)+position) puts the size term at χ^2=83.3 (p=7×10^{-20}) with position controlled, leaving a small residual position term (χ^2=6.1, p=0.013) that observational data cannot adjudicate (accumulated context grows with position by construction), and that the Latin square was designed to settle. The square: two routes, three delivered sizes (9/11, 70/74, 174/194 rows) each appearing at each dialogue slot exactly once per route, byte-identical scored turns in every slot, oracle-verified identical ground truth for a size at every slot.

The square's cell pattern hints at one real secondary texture: small-table accuracy drifts down when large tables precede it in context (92/75/58% across slots) while large-table accuracy does not worsen, a mild accumulated-context tax, far from accounting for the size effect.

The headline movement, accounted in full (Appendix E) The BYPASS side: removing starvation (0.0% post-fix) and the fabrication-teaching prompt lifted BYPASS from 81.4% to 91.3% (+9.9pp). The SYNTHESIS side: two of the eight fixes (timestamp canonicalisation and tie-break re-authoring) removed penalties only SYNTHESIS could incur and were expected to *raise* it by roughly 4pp, yet

SYNTHESIS fell from 47.7% to 41.7% (-6.0pp). Per-dialogue movement is large in both directions ($+41.7$ to -46.7pp on individual dialogues), consistent with SYNTHESIS's high run-to-run variance, but the aggregate move exceeds the pre-fix corpus's own sweep-to-sweep spread and we do not have a mechanism for it.

Sampling-parity ablation (Limitations) Claude runs with top_p unset because Anthropic's API rejects temperature and top_p together. A Gemini ablation (the only family where the combination is settable) measured clearing top_p at -15.7pp for BYPASS and -5.9pp for SYNTHESIS; if the direction transfers across families, the parity break biases against BYPASS and the Claude gap is conservative.

F Results Detail

F.1 Hotel instrument notes

Cities and dialogues. The dialogues filter on price, rating, location, star class, review count, amenities and name. Cities: London, New York, Paris, Singapore, Tokyo, Bangkok, Nashville and Orlando, one stay each (three nights, two adults; Orlando five nights with two children), all deep-inventory markets with no empty searches (a quarter of flight captures are empty), so vendor scarcity is untested here.

Escape repair. Hotel names carry apostrophes, and a re-emitting model sometimes writes them as `\'`, which is not a JSON escape; a repair pass removes the invalid backslash before parsing and is applied to every condition alike. It changed no scored turn in the reported runs.

Dropped units. Seven dialogue-by-repetition units failed before scoring on infrastructure (a token-endpoint read timeout, a server-side cancel) and were re-run once under the same pin and merged; two BYPASS units at $K{=}25$ (New York, repetitions 1 and 3) hit the runtime's 500-call limit on the dialogue's drop-a-constraint turn and are scored zero, not excluded. A dialogue keeps its weight when a repetition is missing; only its own estimate is less precise.

Scoring and inference. Record identity is the vendor's property token, falling back to the canonicalised name when the token is absent; the scored fields are name, nightly price, overall rating, hotel class and review count. A capped hotel table can empty under a filter chain, which the flight liveness check forbids and the hotel design allows at small K; such a turn is correct only on an empty table, and a prose answer without a table fails under the same no-artifact rule as everywhere else. Slopes are per-dialogue least-squares fits of dialogue-mean fidelity on $\ln K$ over the four K; a condition's slope is the mean over dialogues; an axis contrast averages two conditions' slopes; intervals and the paired contrasts come from one bootstrap over dialogues (2000 resamples, the same resampled dialogues on every condition and K). Dialogues that share a city share its capture; clustering is by dialogue, not city.

Instrument defects surfaced by a first pass. As on flights, running the hotel matrix once surfaced instrument defects, all fixed before the reported runs: three translation rules missing from the production hotel prompt (dropping one of several stacked constraints was re-filtered within the current view rather than from the full table; "again" was not treated as a re-expansion; a city named in a follow-up was not referred back to its own search), which made BYPASS misses cascade through later turns; a re-emission instruction that let the model shorten the record schema, emitting rows without the vendor token and nightly price; and a parser that rejected a non-JSON apostrophe escape. Each fix was made in the instrument, the full matrix was re-run under it, and the reported ladder is that run. The in-head arms, which none of these fixes touch, did not move between the passes; the run matrix, repetitions, cluster unit, ladder and dialogues were fixed before either.

Corpus difference. The hotel dialogues include one move the flight corpus lacks: dropping one of several stacked constraints ("Never mind the restaurant"), which must be answered by re-filtering the full table rather than narrowing the current view.

F.2 Hotel disclosure rates

On the hotel turns where the frozen tool excluded listings for missing data, a binary sentence-level disclosure score reads: BYPASS 96% and REEMIT 87%, both quoting the tool's exclusion note;

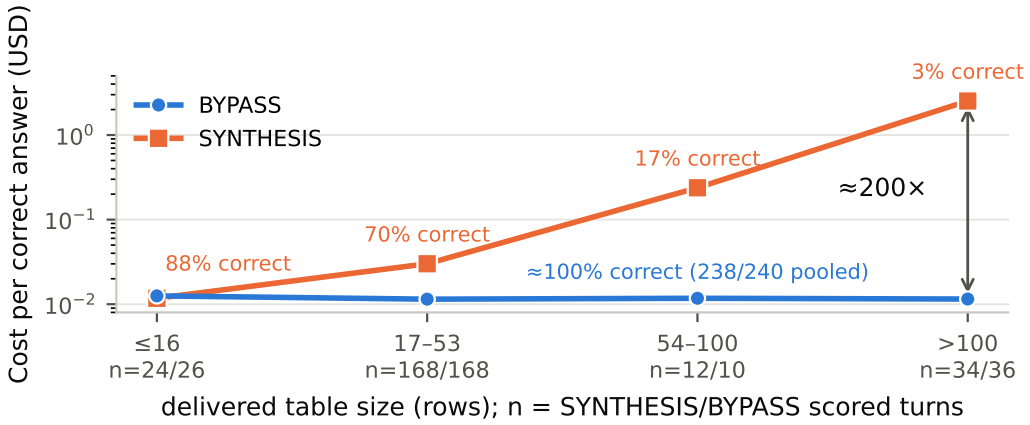


Figure 4: Cost per correct answer by delivered-size band (size-ladder sweeps; mean turn cost ÷ accuracy, log scale). BYPASS is flat at ≈\$0.012; SYNTHESIS climbs to \$2.53 beyond 100 rows, ≈200×. Point labels give SYNTHESIS's per-band accuracy (one correct SYNTHESIS turn in the >100 band; the multiple is an order-of-magnitude estimate). Unequal n: the starved ORD-NRT run's two BYPASS turns (0 of 72 rows delivered) count in ≤16, its only misses; two >100-row SYNTHESIS turns emitted nothing scoreable (Appendix A).

SYNTHESIS 11%, which must count the exclusions itself; REFPICK 0%, whose delivery override asks for handles and a summary and never mentions exclusions, a limitation of the metric for that condition as much as of the condition.

F.3 The two instrument defects behind the 2×2 rescoring

Every number above is from a single offline rescoring pass over persisted traces after an adversarial audit found two instrument defects, both corrected without re-running a model. The first scored a right-count-wrong-set turn as perfect (it scored field fidelity over only the intersecting rows once the counts matched), which inflated the in-head conditions by 4 to 7pp at mid and large K; exact-set scoring removes it, and the headline corpus was not rescored and retains its original scoring. The second excluded stalled cells only where a later guard had flagged them, deflating REEMIT by 5 to 12pp at the K where stalls clustered; one uniform rule now excludes a filter turn only when its delivered table was empty. The two corrections push in opposite directions and leave the qualitative 2×2 unchanged.

F.4 Family, cost and clustering paragraphs, in full

The failure shapes are family-specific. **Gemini corrupts**: real rows, wrong set (§5.4). **GPT garbles**: unparseable JSON mid-table, type-broken fields, dropped required fields. **Claude omits**, concentrated on larger tables: well-formed JSON that drops the required `flights` array while the prose summary counts the delivered table correctly (an accurate count attached to a missing table). One mechanism, three masks (specifics, per-family latencies, and a small-table pilot artifact in Appendix F).

The taxonomy has a load axis, and it is the paper's thesis in one picture. Re-run on all four models with identical K-row prefixes of the same frozen 198-row payload, the content-controlled ladder (§5.1) shows the **collapse threshold ordering by model capability**: under 10 rows for GPT-4o-mini, 10–25 for Claude Haiku, 50–100 for Gemini Flash, while Gemini Pro alone degrades gracefully (perfect through 25 rows, 83% at 50, 67% at 100, and 50% at the full 198), and BYPASS is size-flat for all four models (140/144 across the ladder; Figure 1). Capability buys threshold; architecture buys size-immunity. Cells pool six repeats.

model	BYPASS	SYNTHESIS	gap
Gemini 2.5 Flash	92.3% [89.0, 94.7]	45.7% [40.5, 51.0]	**+46.6pp** [+35.7, +57.1]
Gemini 2.5 Pro	92.6% [89.3, 95.0]	72.0% [67.0, 76.5]	**+20.6pp** [+11.7, +31.4]
GPT-4o-mini	67.0% [61.8, 71.8]	17.4% [13.7, 21.8]	**+49.6pp** [+39.2, +60.2]
Claude Haiku 4.5	82.9% [78.5, 86.5]	27.1% [22.7, 32.1]	**+55.8pp** [+41.1, +69.9]

Table 12: Cross-family replication at full-matrix scale (n=339 scored turns per cell, k1–3; Wilson 95% CIs on cells, dialogue-level bootstrap CIs on gaps).

F.5 Cost and latency, not only accuracy

On the full 24-dialogue grid, mean per-turn latency is **BYPASS 4.9s vs SYNTHESIS 31.7s**, a 6.5× penalty. BYPASS's latency and cost are flat in table size because it re-emits nothing, while SYNTHESIS's scale with the set it must reproduce (Figure 5, Appendix F). At the largest tables a synthesis turn averages 146 seconds (194 at worst) and costs 8.7× BYPASS's flat per-turn cost, for an 8%-fidelity answer. Dividing cost by accuracy (Figure 4): BYPASS produces a *correct* answer for ≈$0.012 at every table size, while SYNTHESIS's cost per correct answer climbs to **$2.53 beyond 100 rows, roughly 200×**; tables that size are precisely the regime a search product exists for. Model tier does not escape this arithmetic (§5.3; Appendix F): a model upgrade rents part of what the architecture gives outright, and the rent is high.

Two independent instruments show synthesis failures are not independent draws, and bypass failures are. Lag-1 conditioning: after a failed turn, SYNTHESIS fails again 71.9% of the time against 31.4% after a clean turn (a **recovery rate of 28.1%**), while BYPASS recovers **95.7%**. Value-tracing (a deliberately liberal upper bound; Appendix C.5): **53 of SYNTHESIS's 177 traceable failures (30%) are propagated; 0 of BYPASS's 25 are**: for BYPASS no model-emitted set exists to trace, so propagation is precluded and the zero reflects that no counterexample arose. The mechanism is architectural: BYPASS recomputes each turn from preserved state; SYNTHESIS filters its own poisoned output, and its residual failures are isolated translation slips.

Our pre-registered turn-count super-linearity test finds none. The reason is structural: with SYNTHESIS's 56.9% first-turn failure rate the accumulation-only null saturates, leaving almost no measurable headroom above it; the test is degenerate in its own high-error regime (construction, excess-by-position and sample sizes in Appendix C). This redirected the analysis to table size (§5.1), which is monotone, controlled, and does not saturate.

F.6 Cross-family full matrix (§5.3)

Table 12 gives the per-condition cells behind §5.3's paired gaps (n=339 scored turns per cell: all 24 dialogues; Wilson 95% CIs on cells, dialogue-level bootstrap CIs on gaps; the cell intervals pool turns and are descriptive; inferential gap claims rest on the dialogue-level CIs). Extending the repeats from three to five sweeps for the two non-Gemini families (GPT-4o-mini and Claude Haiku 4.5; n=565 per cell) moves no cell by more than 1.1pp and no gap by more than 1.4pp.

F.7 Trend-test robustness (§5.1)

Because repeats within a route are correlated, the decreasing top-K trend was also verified under route-level cluster bootstrap: it survives, appearing in 2000 of 2000 resamples, i.e., p below the bootstrap's resolution floor ($p < 5\times10^{-4}$).

F.8 Family failure shapes, specifics (§5.3)

GPT's garbling includes type-broken fields ("21h 55m" where the schema wants an integer) and responses that drop required fields (a `flights` array with no `summary`, the mirror image of Claude's omission), including 13 runs with no scoreable output at all, four of them because the re-emitted tables inflated the transcript past the model's 128k context window; BYPASS, whose context never carries the table, has no such run. Claude's omission mode appeared on thirty-five runs: twenty-nine with an intact prose summary counting the delivered table correctly ("Found 48

model	BYPASS turns correct	argument divergence	of which harmless	failed turns from wrong args
GPT-4o-mini	67.0%	45.6% (150/329)	41% (61/150)	79% (89/112)
Claude Haiku 4.5	82.9%	32.0% (103/322)	63% (65/103)	66% (38/58)
Gemini 2.5 Flash	92.3%	27.2% (92/338)	76% (70/92)	85% (22/26)
Gemini 2.5 Pro	92.6%	31.2% (105/336)	82% (86/105)	76% (19/25)

Table 13: Cross-family translation error over scored BYPASS turns (n=339 per model; rows ordered by ascending BYPASS accuracy). Divergence = applied filter arguments differ from the authored spec; denominators are turns with trace-recoverable applied arguments (the shortfall of 10/17/1/3 turns is all failures). Harmless = divergent arguments that still select exactly the right rows, so divergent minus harmless = the wrong-argument failures. Last column: wrong-argument failures over all failed turns; the remainder are 13/3/3/3 failures with right arguments plus the 10/17/1/3 unrecoverable, closing each total (112/58/26/25).

award flights…") and six emitting an empty object. Omission is even cheap (4.0 s per Claude synthesis turn; not emitting the table is fast), where garbling is not (16.0 s per GPT synthesis turn). A three-dialogue pilot had suggested Claude's synthesis was markedly stronger (64.7% there); that was an artifact of the pilot's small tables, which rarely trigger the omission mode. GPT's ladder zero at K=5 (0/6) against 33% at K=10 is a real non-monotonicity of its garble-dominated mode. A power note on size-flat BYPASS: a zero-failure cell excludes, with 80% probability, only a true per-turn decay of at least 8.5pp at n=18 (prefix-ladder cells) and 23.5pp at n=6 (family-ladder cells); the exclusion of a large-size decay rests on pooling across the four instruments (93/93 prefix ladder, 140/144 family ladder, 238/240 date-varied ladder, 106/108 Latin square), where every BYPASS miss is starvation-shaped (empty delivery, no rows reaching state) and none is size-graded.

F.9 The tier's residual failures and per-answer costs (§5.3, §5.5)

Gemini 2.5 Pro's full-corpus cells are in the cross-family matrix (Appendix F.6); a Pro synthesis turn averages 35.5 s against Flash's 22.8 s (scored filter turns, the same turn class Figure 5 plots; §5.5's 31.7 s is the all-turns mean; 1.6×) and $0.141 against Flash's $0.028 (5×), twelve times BYPASS-on-Flash's $0.012. Residual synthesis failures carry the size signature of §5.1 (in the tier pilot runs: near-miss truncation on 40–77-row tables, 77 delivered → 74/76 emitted; collapse at 138 → 0 and 76 emitted).

On the content-controlled ladder Pro's correct full-payload answers cost $0.67 each (3 of 6 correct at a mean $0.336 per turn) against BYPASS's flat $0.012. A pre-fix probe on five failure-selected dialogues agreed with the tier comparison's direction: Flash-BYPASS 60.0% vs Pro-SYNTHESIS 52.0%.

F.10 Failure anatomy (§5.4)

Beyond truncation, SYNTHESIS over-emits by carrying a filter it should have dropped. On the deployed Gemini 2.5 Flash, BYPASS's 22 wrong-argument failures are translation and scope errors; corpus-wide, its applied arguments differ from the authored spec on 27.2% of argument-recoverable scored turns (92/338 of 339), but 70 of those 92 argument sets (76%) still select exactly the right rows: argument-level noise absorbed by the deterministic evaluator, where synthesis would have re-typed every field.

Wrong-argument failures order exactly inversely with the four BYPASS accuracies (Table 13), including the Pro/Flash flip at the top, and the harmless share rises with BYPASS accuracy (41% to 82%): BYPASS's floor is translation quality (how well a model maps each follow-up's intent into filter arguments, §3.1), measured.

G Extended related work and system detail

G.1 Extended related work

The body's §2 keeps the closest works; the full positioning:

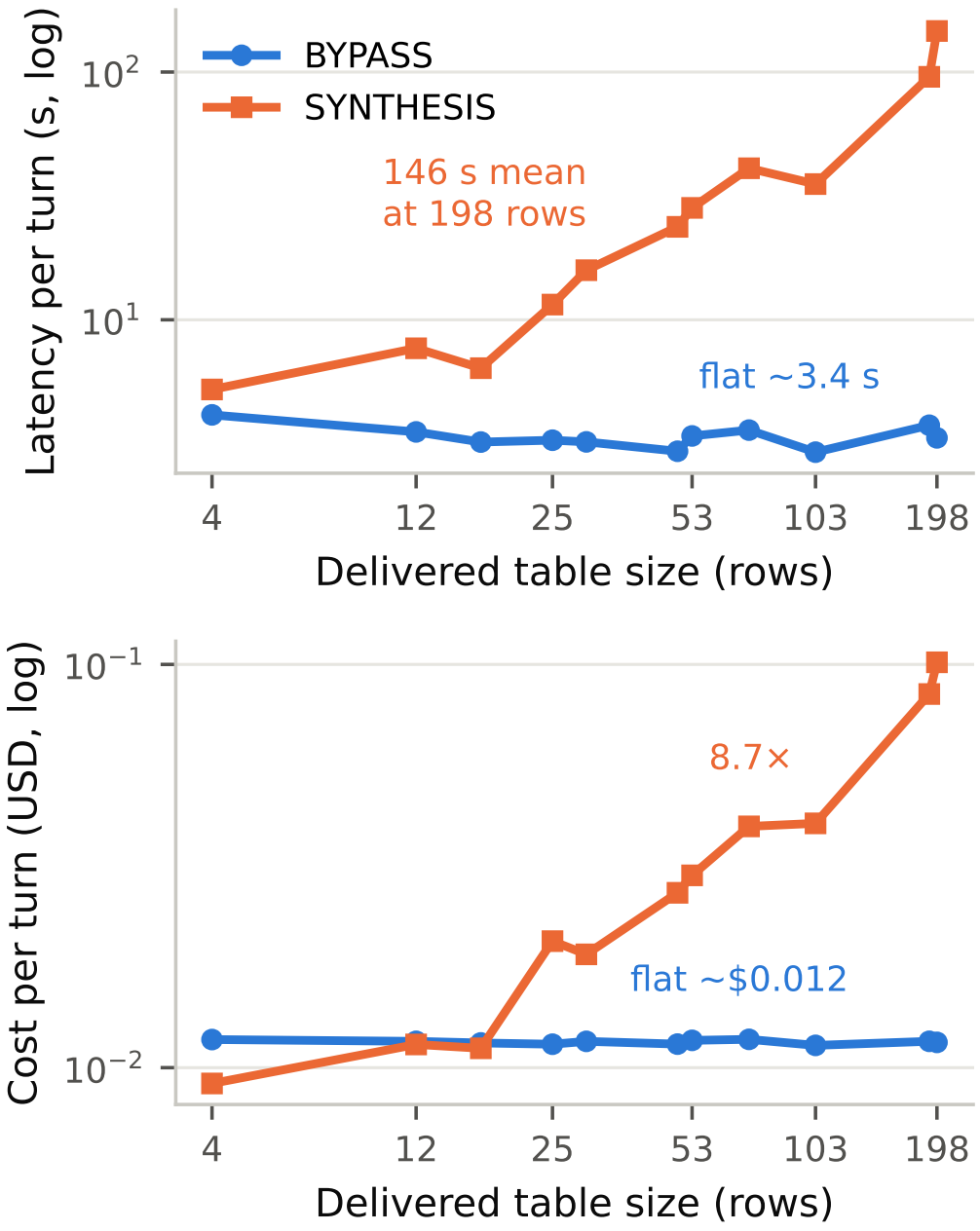


Figure 5: Per-turn latency (top) and cost (bottom) against the actual delivered table size, over the size ladder's scored filter turns. SYNTHESIS scales from ∼5s to a 146s mean turn and, at the largest size, 8.7× BYPASS's per-turn cost; BYPASS is flat at ∼3.4s and ∼$0.012 at every size. At the smallest tables SYNTHESIS is marginally cheaper, crossing above BYPASS by 25 rows. (§5.5's 4.9s/31.7s are all-turns means over the full grid; this plot's means cover the ladder's scored filter turns only. §5.1's design labels are post-seat-filter counts; this figure uses the true delivered sizes.)

Structured output: validity is not correctness. Constrained decoding guarantees shape, not content: format restriction degrades reasoning (Tam et al., 2024), generation shifts onto lower-confidence paths (Schall & de Melo, 2025), validity hits 100% while accuracy falls and wrong-but-valid outputs dominate (Ray, 2026), task accuracy coexists with zero format accuracy (Galeone et al., 2026); grammar-constrained infrastructure (Willard & Louf, 2023; Dong et al., 2024) and

schema-compliance benchmarks (Geng et al., 2025) show the shape guarantee is real. Hoang (2026) measured the resulting structured hallucination in a production travel agent single-turn, introducing the field-fidelity metric and zero-out policy we adopt. This paper compares two delivery architectures on a new dataset in a multi-turn design (results are not compared to that study's), showing the failure survives into a *multi-turn agentic pipeline over real vendor records*, where corrupted output re-enters the context and is filtered again.

Multi-turn agent evaluation and error compounding. Tool-use work addressed invocation (Yao et al., 2023; Schick et al., 2023; Patil et al., 2024); our subject is what happens to the returned data afterwards. τ-bench is the reference multi-turn tool-agent benchmark (Yao et al., 2024), MINT showed that single-turn performance does not reliably predict multi-turn (Wang et al., 2024b), a 39% average single- to multi-turn drop with poor recovery is documented (Laban et al., 2025), and self-conditioning, in which past errors raise later error rates, is the closest mechanism match (Sinha et al., 2025); failure taxonomies (Zhu et al., 2025), stability analyses (Xiong et al., 2025), reliability framing (Rabanser et al., 2026), a Tool-Skip category matching our secondary observation (Soni, 2026), and injected corruption producing over-trust and reduced recovery (Zhu et al., 2026) situate the problem. Closest are three unreviewed preprints: AgentProp-Bench injects parameter errors into real execution traces and audits evaluators and propagation, not fidelity to a delivered record set (Gurram, 2026); "Beyond Task Completion" reports 27–78% of τ-bench successes as *corrupt* (Cao et al., 2026): procedural integrity on τ-bench's mock domains (airline and retail), where ours is field- and row-level data fidelity; "Reason Less, Verify More" finds, on a budget agent, 78% of failures are silent wrong-state and gates the *write* side (Reddy et al., 2026), where ours bypasses the *read* side. All index degradation by turn count or injected fault on simulated or injected data; ours is indexed by delivered record-set size, on real data, and the corruption is self-inflicted.

Architectures that keep data out of the model. CodeAct routes actions through executable code (Wang et al., 2024a); "heavy brain, precise hands" designs pair LLM planning with a deterministic backend (Zeng et al., 2026); trace-level compliance verification adds a deterministic gate (Elkoussy & Perez, 2026); the prompt-injection literature converges on context-minimization (Beurer-Kellner et al., 2025). Our mechanism is in this family; its practitioner form exists in LangGraph's custom stream mode, the Vercel AI SDK's data parts, and the AG-UI protocol, but **no framework we surveyed ships typed structured delivery as the default end-to-end rendering path**, and the documented motivations are rendering predictability, context size and cost, with correctness noted but not measured (CopilotKit, 2026). Survey method: for each framework (LangGraph, Vercel AI SDK, CrewAI, Microsoft Agent Framework, OpenAI Responses API, CopilotKit/AG-UI; documentation accessed 28 July 2026) we read the current public streaming and rendering documentation and asked one question: does structured tool output reach the client renderer without re-entering the model's context on the documented default path? Default means the quickstart path, not an available opt-in.

Evaluation methodology. Turns within a dialogue are not independent, and turn-level significance claims frequently fail cluster-robust correction (Schessl, 2026); every gap interval here therefore resamples *dialogues*, not turns, (see Guan et al. (2025) for the evaluation landscape).

Reasoning inside dialogue. Closest in setting is BOULDER (Kartáč et al., 2026): reasoning over travel-task travel domains, isolated and embedded in dialogue, with a substantial isolated-to-dialogue drop attributed largely to multi-turn framing. BOULDER measures whether the model *computes correctly over records it was given*; we measure whether the record set *survives delivery*; it has no synthesis-bypassing arm and no row- or field-level fidelity metric, and its ablations cannot separate the assistant's dual generation role from longer context; the generation burden can only be removed architecturally.

G.2 Deployed-system detail

Award inventory is seats bookable with loyalty points, priced per booking program, so the same flight legitimately appears once per program; *cash* is ordinary priced tickets. The search sub-agents perform hub routing, deduplication and canonicalisation before writing the record set to session state.

H QUALITATIVE FAILURE ANALYSIS: FULL CENSUS AND EXHIBITS

H.1 AFTER THE PROMPT FIX: THE FABRICATION IS GONE, THE CORRUPTION IS NOT

On the frozen instrument we audited every failed SYNTHESIS turn in the size ladder, 99 of them. **96 of the 99 carry partial-credit fidelity $\approx$1.0: every row the model emits is real.** Fabricated tables were never a data-present phenomenon; post-fix, fabrication persists exactly where residual starvation persists (19 empty-context turns, mean 1.4 rows) and nowhere else. What did not vanish is the set error. The dominant modes:

- **Truncation.** Asked to show a 74-row result, the model emits 55 of the 74 (real rows, correctly formatted, silently incomplete; `d26_fra_bkk_size_m_early` t2: GT 74, emitted 55, partial 1.0).
- **Cascade on its own emission.** The next turn then filters the emitted 55 rather than the true 74: asked for the afternoon subset (GT 11), it returns 45: real rows over a corrupted base and a corrupted scope (same dialogue, t3). This is Exhibit 1's propagation mechanism (Appendix D), on real rows instead of invented ones.
- **Identifier splicing under load.** In 25 of 943 post-fix populated emitting turns an identifier is corrupted inside an otherwise-real row set, concentrated in large-table runs (specifics in Appendix F.8): field-level corruption that count-based checks cannot see.

H.2 THE PRE-FIX RUNS PIN THE ATTRIBUTION (EXHIBITS 1–2, APPENDIX D)

The obvious deflationary reading, that synthesis merely had a bad prompt, is answered by comparing runs before and after the prompt repair. Pre-fix, the production prompt itself taught fabrication (Appendix E): in one starved run SYNTHESIS answered a points filter over an undelivered payload with a plausible schema-valid table, then filtered its own fiction for five more turns; constrained decoding guaranteed the shape of the lie. BYPASS's failure on the same starvation was visible emptiness: its table comes from state, and empty state renders empty (turn by turn in Appendix D, Tables 10–11). Repairing the prompt **removed the fabrication but not the failure**: fabrication tracked grounding absence under both prompts, while set corruption survived every prompt repair, scales with the delivered table (§5.1), and BYPASS cannot exhibit it. Attributing the gap to bad prompting is backwards: the better the prompt got, the more purely the residual gap measured the architecture.

I EXTENDED LIMITATIONS

The size manipulation has bounded coverage. The content-controlled ladder covers six routes but one cabin, with per-route ceilings of 112 to 198 rows against the 510-row payloads we hold; a true 200-plus band through range captures was declined because it would exchange the position confound for a query-shape confound at exactly the cell carrying the effect. Whether the decay moves with route complexity, cabin or record width is open, and the cell resolution localises the shape only coarsely: the pooled decay is monotone from 83% at $K{=}5$ to 6% at $K{=}100$, with no single knee we can name.

Scope conventions are common-mode. The §4.2 rulebook was fixed before runs and applied identically to both arms, so a defective convention shifts both conditions' levels rather than the between-arm contrast (the Appendix E corrections were applied to both arms).

Checks that do not share the evaluator code path. A shared-evaluator defect would depress SYNTHESIS alone; the checks that do not share that code path are the authored expected-row assertions (Appendix A), the pre-run liveness verifier (§4.3), and the hand audits of all 99 ladder failures against the frozen payloads (§H.1); the independence detail is in Appendix C.3. The size-conditional curves (Tables 6 and 7) let readers reweight the marginal under any deployment's size distribution; vendor terms preclude releasing the payloads.

Statistical resolution. CIs are wide (dialogue-level bootstrap: BYPASS 88.3–95.9, SYNTHESIS 34.8–57.2; gap [+35.7, +57.1]) because $n{=}24$ dialogues is the resampling unit, and strata are small: the largest-size $\times$ deepest-position stratum holds 7 turns, and the two deepest turn positions are reached by only 8 and 3 dialogues respectively.

Other open items. An outcome category for count-only-envelope turns would raise BYPASS's measured rate honestly; a greedy-decoding ablation and a temperature sweep were designed and not run. The corpus-level marginal (+46.6pp) depends on the authored size mix; readers can reweight the size-conditional curves under any deployment's size distribution.