FROM MEMORY TO AUTHORIZED ACTION: HOW STATE REPRESENTATIONS SHAPE LLM AGENT DECISIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

LLM agents must retain delayed user requests across interactions and execute them when the relevant conditions are met. However, retaining a request does not establish whether it is currently authorized for execution or should instead be deferred or closed. We introduce ReCASH-PM, an empirical framework for auditing the evidence that connects remembered requests to current action decisions in frozen language models. The framework separates state acquisition, request representation, and action selection to examine how authorization evidence is preserved and used. Across 512 human-evaluated dialogue prefixes from 128 dialogues, we evaluate two model families across three quantized configurations, comparing dialogue histories, summaries, and structured states alongside alternative decision rules. Our audit reveals that original state compilation improves aggregate decision accuracy by 10.55–24.02 percentage points over direct-history prompting while reducing authorized-action recall from 72.73–93.18% to 0–3.41%. Predicate checks distinguish whether acquired fields satisfy execution conditions from whether the model selects execution, while ablations with acquisition fixed reveal sensitivity to the state components presented. A revised interface separating active-request matching from closure evidence partially restores recall while increasing false execution. Higher aggregate decision accuracy therefore does not guarantee retention of authorized actions. ReCASH-PM jointly assesses authorized-action recall, false execution, erroneous closure, and state-acquisition cost, providing a practical basis for diagnosing how remembered evidence supports or suppresses action.

1 INTRODUCTION

Large language model (LLM) agents must retain some user requests across interactions and act when a later condition is met. In this setting, prospective memory means retaining an intention until a future time or event calls for action (Liu & Gabriel, 2026). Consider a report reminder scheduled for 13:00 and a taxi-booking reminder for 14:10. The user moves the report reminder to 14:10 but cancels the taxi reminder. At 14:10, the agent should send only the report reminder. Remembering both original requests is insufficient: it must determine which request remains active and is now authorized for execution.

Retrieval, summarization and explicit state tracking make past interactions available for later decisions. Availability does not establish present authority: an active request may not be due, and closure evidence must concern the nominated occurrence.

On the same dialogue prefixes, original state compilation raises aggregate accuracy while nearly eliminating decisions to execute reference-authorized requests. More correct non-execution decisions can conceal lost authorized actions, whose recovery can also increase false execution. We ask: **how do the acquisition, representation and use of request state change the balance between retained authorized actions, false execution and erroneous closure?**

We introduce ReCASH-PM, an empirical framework that makes these interfaces separately observable with frozen model weights (Figure 1). A writer reads the visible dialogue prefix and produces a fallible summary or state; a separate decision call selects an action for the nominated request. Direct-

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.

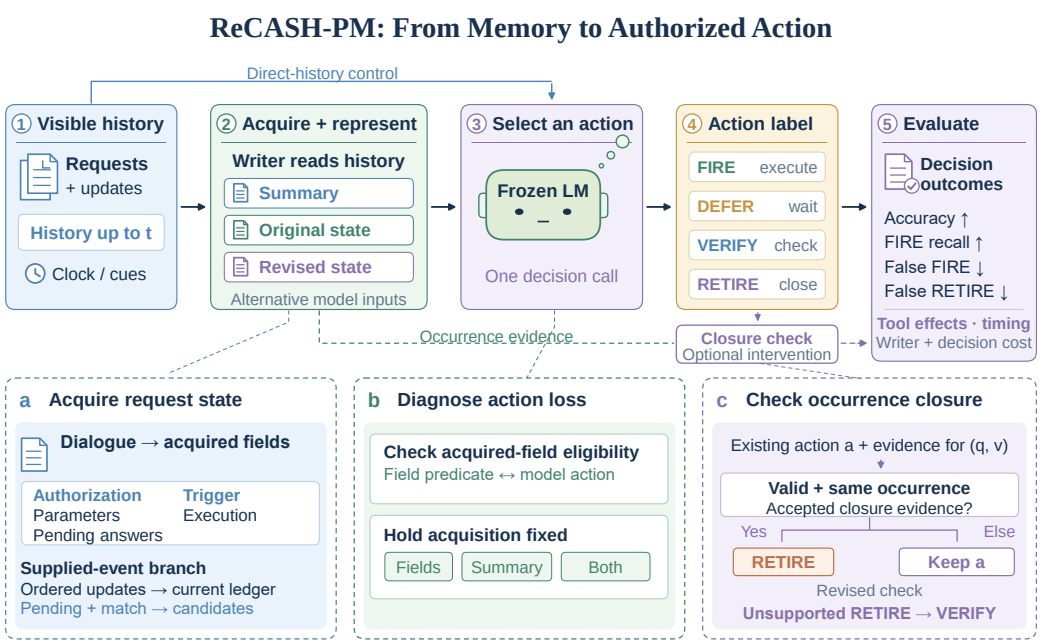


Figure 1: **ReCASH-PM follows request evidence from history to decisions.** Direct bypasses acquisition; other arms expose summaries or states. Insets show acquired fields, action-loss diagnosis and optional closure checks. Algorithm 1 specifies supplied-event updates and the original guard.

history prompting provides the comparison without acquisition. A revised state separates matching the active request from matching evidence for its closure. Additional controls expose different information views of one acquired state, check its execution predicates, or apply an occurrence-specific closure check to the output. The predicates describe the acquired fields, rather than serving as ground-truth authorization.

Across 512 human-evaluated prefixes from 128 dialogues, we test two model families in three quantized configurations. Original compilation raises aggregate accuracy by 10.55–24.02 percentage points over Direct while reducing authorized-action recall from 72.73–93.18% to 0–3.41%. Most omitted actions become deferrals. Revised state recovers some recall relative to Original, with increased false execution. These results yield three contributions: action-sensitive measurement, interface comparisons distinguishing field eligibility from selection, and evaluation of native action references, tool effects, timing and acquisition cost.

2 Related Work

Remembering information. ICAE, Activation Beacon, LongLLMLingua and LLMLingua-2 compress context through memory slots, key/value activations or prompts (Ge et al., 2024; Zhang et al., 2025; Jiang et al., 2024; Pan et al., 2024). LoCoMo and LongMemEval test conversational recall, temporal relations and knowledge updates (Maharana et al., 2024; Wu et al., 2025). These evaluations leave open whether an acquired request authorizes an action now.

Delayed intentions and changing state. PM-Bench tests delayed execution; PIS stores typed intentions; StateMem tracks supersession (Liu & Gabriel, 2026; Zhao & Wu, 2026; Fan et al., 2026). EAL-Bench examines authorization errors in memory, and ToolSandbox evaluates state changes after tool use (Cerruti et al., 2026; Lu et al., 2025). Their settings motivate distinguishing an active request from closure, but do not isolate acquisition, representation and selection on the same dialogue prefixes.

Locating failures. MemTrace traces memory-operation errors, while CausalFlow studies counterfactual repair (Deng et al., 2026; Bonagiri et al., 2026). Here we compare field eligibility with model

selection and vary the information shown after one acquisition; Section 3 formalizes this observable diagnosis.

3 From changing requests to auditable decisions

We measure action retention (RQ1), compare field eligibility with model selection (RQ2), and evaluate native action references and execution (RQ3).

3.1 Define the current request and decision

A natural-panel example comprises a visible dialogue prefix and nominated request occurrence. The model sees only utterances and tool records up to that cut, excluding future dialogue and reference labels. FIRE executes now; DEFER awaits a known condition or pending answer; VERIFY seeks material evidence or checks pending execution; RETIRE closes a completed, cancelled or superseded occurrence. The panel scores these labels; RQ3 measures execution effects.

Identity links the update to the report reminder, version distinguishes its revision, and the effective trigger becomes 14:10 (Figure 2). The taxi reminder remains inactive despite the matching time. Cancellation persists until explicitly authorized reactivation. Each new occurrence, including recurrence after completion, has a distinct identity.

For supplied events, we specify these updates before selecting pending requests whose effective triggers match the present observation:

$$L_t = U(M, E_{\le t}), \qquad B_t = \{r \in L_t : \text{pending}(r) \wedge [\text{time}(r) = \tau_t \vee \text{cue}(r) \in C_t]\}. \quad (1)$$

Here U updates metadata M with ordered visible events $E_{\le t}$. Ledger L_t retains inactive records to distinguish cancellation from absence; B_t matches pending records to clock τ_t or cues C_t. This specifies supplied-event semantics; natural dialogue requires a writer to acquire state from text.

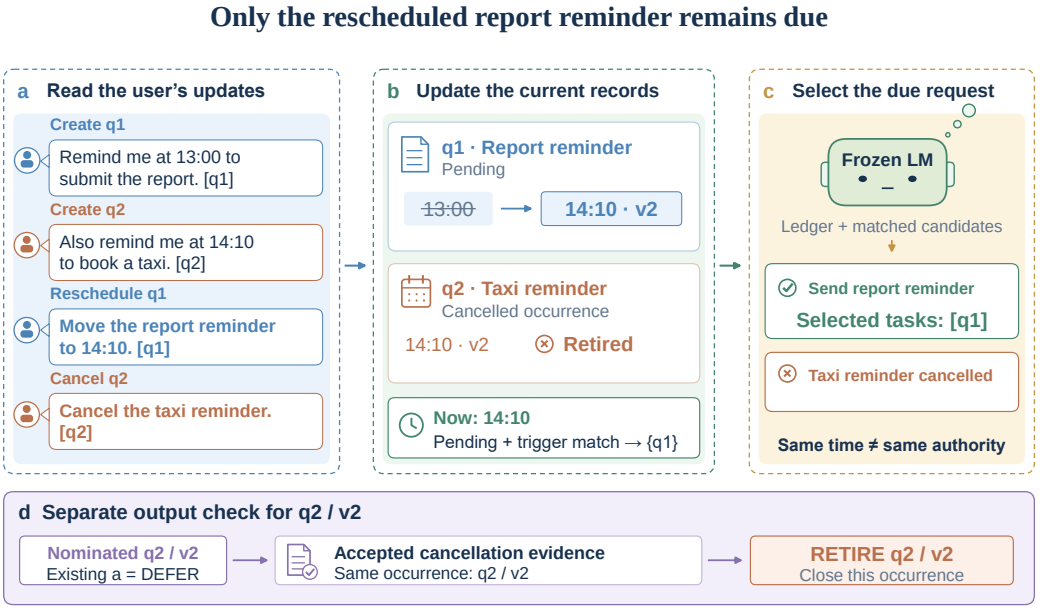


Figure 2: **Only the rescheduled report reminder remains due.** At 14:10, the report reminder moved from 13:00 remains active; the taxi reminder is cancelled. The lower panel uses same-occurrence cancellation evidence to change DEFER to RETIRE. This is a semantic illustration, not a measured parser result; execution means sending the reminder.

Algorithm 1 Current-request compilation and occurrence-bound closure. The procedures intervene at distinct interfaces.

Require: Metadata M, ordered visible events $E_{\leq t}$, clock τ_t, cues C_t
Ensure: Current ledger and trigger matches; occurrence-bound correction of an existing decision

procedure COMPILE($M, E_{\leq t}, \tau_t, C_t$)
  Initialize L_t from M; distinguish pending from not-yet-created records
  for all visible steps $s \leq t$ in order **do**
    Activate newly created records only if their status is not-yet-created
    for all updates e at s in recorded order **do**
      $r \leftarrow L_t[e.\text{id}]$
      Cancellation retires r; rescheduling or a cue override changes its trigger
      Only explicitly authorized reactivation restores pending status
      Increment the version; record the update type
    end for
  end for
  Compute B_t using Equation 1; sort and annotate trigger matches
  return (L_t, B_t) ▷ Expose current records and candidates
end procedure
procedure CLOSUREGUARD(a, i, ℓ_t)
  if $\neg V(a)$ **then return** a
  if $C(i, \ell_t)$ **then return** RETIRE
  return a ▷ Keep the decision when closure lacks support
end procedure

3.2 ACQUIRE STATE AND SEPARATE ACTION EVIDENCE FROM CLOSURE

Direct supplies history; Summary supplies acquired prose; Original and Revised supply their respective compiled states. The writer sees only the visible prefix and can omit evidence, misinterpret updates or produce invalid output. These comparisons evaluate acquisition together with its resulting representation.

Revised separates *active-occurrence match*, which identifies the nominated commitment after explicit parameter revisions, from *terminal-evidence match*, which checks completion, cancellation or replacement evidence for that occurrence. The report reminder can therefore match before first execution; the taxi cancellation closes only its own occurrence. A first execution needs no completion receipt, but absent closure evidence does not itself authorize action.

Revised records authorization, trigger status, required parameters, execution state and pending answers. An acquired-state rule evaluates predicates over those fields. Fields-only and summary-only share that acquisition but expose different views; ordinary rereading instead supplies Direct's tentative decision and the visible history. Comparing these outputs tests field eligibility, information exposure and rereading without assuming correct acquisition.

3.3 AUDIT CLOSURE AT THE OUTPUT INTERFACE

An input representation changes the evidence offered to the model. An output check instead constrains a decision already made. For a nominated occurrence $i = (q, v)$, the original closure guard changes a valid decision to RETIRE when accepted evidence supports closing that occurrence:

$$G(a, i, \ell_t) = \begin{cases} \text{RETIRE}, & V(a) \wedge C(i, \ell_t), \\ a, & \text{otherwise.} \end{cases} \tag{2}$$

Here a is the decision, ℓ_t the accepted state, V checks validity, and C checks occurrence-specific terminal or inactive-version evidence. The original guard only adds retirement; it cannot create FIRE or validate an existing RETIRE. The revised check also maps unsupported RETIRE to VERIFY. These interventions test closure separately from recovering execution. Acceptance operates on acquired fields and citations, whose semantic correctness is a separate diagnostic question.

3.4 EXPERIMENTAL SETUP

Data and observation units. The primary panel reuses 512 human-evaluated prefixes from 128 dialogues: 64 SGD and 64 Taskmaster-2 dialogues, with four cuts each (Rastogi et al., 2020; Byrne

et al., 2020). Half of each source's dialogues are hash-ordered random candidates and half satisfy fixed cancellation, revision or waiting patterns. This defines the evaluated cohort, not natural class prevalence. References comprise 88 FIRE, 107 DEFER, 90 VERIFY and 227 RETIRE. Independent evaluation adds 600 cuts from 200 unused ABCD test dialogues with native next-step/action annotations (Chen et al., 2021). Execution checks cover 36 EAL cases and 264 requests, 95 local-tool ToolSandbox scenarios, and the complete seven-day PM-Bench episode. Each retains its own endpoint (Appendices A–F); Section 5.4 specifies the separate supplied-state panel.

Models and controls. We evaluate the Gemma 3 12B and Llama 3.1 8B model families (Gemma Team, 2025; Grattafiori et al., 2024). The recorded local checkpoints are Gemma Q4, Llama Q4 and Llama Q8, with fixed weights and requested seed 908. Q4/Q8 differ in recorded checkpoint source as well as precision, so their comparison is descriptive rather than an isolated quantization treatment. The natural panel contains 13 conditions: Direct, Summary, Original, Revised, two views of revised state, ordinary rereading, two inspired memory controls, two acquired-state rules and two closure checks. PIS-inspired and StateMem-inspired arms use documented local observation interfaces. No result-dependent retry is used. External tasks retain their native scorers under the local role serialization, dispatch and simulator adaptations in Appendix A.

Metrics and comparisons. Natural-panel accuracy uses all 512 prefixes and counts invalid outputs as errors. Authorized-action recall is FIRE on the 88 FIRE references; false execution is FIRE on the 424 non-FIRE references; false closure is RETIRE on the 285 non-RETIRE references. All rates are percentages. The predeclared primary contrast is Revised versus Direct, evaluated jointly by retained action and error rates. Original versus Direct characterizes the motivating discrepancy. Paired intervals use 10,000 resamples of the 128 dialogue clusters, seed 9182026, without multiplicity correction. External metrics use their native denominators. ToolSandbox resamples whole scenarios within four base families; PM-Bench contributes one whole episode per controller. Acquisition-inclusive cost charges both writer and decision calls.

4 RQ1: DISTINGUISH CORRECT DECISIONS FROM RETAINED ACTION

For Gemma Q4 on the same 512 prefixes, Original raises accuracy from 29.69% to 52.73%, yet selects FIRE on only one of the 88 FIRE references, compared with 64 for Direct (Tables 1 and 2). Llama Q4 and Q8 likewise gain accuracy while correct FIRE decisions fall from 77 to zero and from 82 to three. Correct non-execution decisions offset the lost FIRE decisions, so aggregate accuracy conceals nearly complete action loss.

Table 1: **Accuracy gains and action retention favor different representations.** All conditions use the same 512 prefixes. Recall uses 88 FIRE references; F-FIRE (false execution) uses 424 non-FIRE references; F-RET (false closure) uses 285 non-RETIRE references. All values are percentages; bold marks the best value within each model and metric.

Model	Input	Acc. ↑	Recall ↑	F-FIRE ↓	F-RET ↓
Gemma Q4	Direct	29.69	**72.73**	66.75	**0.35**
	Summary	42.58	50.00	37.97	1.05
	Original compiled	**52.73**	1.14	**3.54**	9.82
	Revised compiled	30.08	13.64	17.92	1.05
Llama Q4	Direct	25.59	**87.50**	81.60	0.70
	Summary	26.95	29.55	26.89	**0.00**
	Original compiled	**36.13**	0.00	**0.24**	4.56
	Revised compiled	24.41	38.64	35.38	**0.00**
Llama Q8	Direct	21.09	**93.18**	91.04	0.35
	Summary	26.76	55.68	52.36	0.35
	Original compiled	**45.12**	3.41	**4.72**	3.86
	Revised compiled	25.00	50.00	44.58	**0.00**

Revised changes this balance. For Gemma, correct FIRE decisions rise from one to 12, while false FIRE increases from 3.54% to 17.92% on the other 424 prefixes. Across the three configurations,

revised recall reaches 13.64%, 38.64% and 50.00%; false execution increases relative to Original in each. Against the predeclared Direct comparison, Revised reduces false execution but retains fewer authorized actions, so it does not meet the joint-improvement criterion. Table 1 places accuracy, authorized-action recall and error rates side by side.

5 RQ2: LOCATE THE EVIDENCE-TO-ACTION GAP

We trace omitted actions to their predicted labels, compare field eligibility with model selection, vary shared-acquisition views, and separately test closure correction.

5.1 MOST OMITTED ACTIONS BECOME DEFERRALS

Of the same 88 FIRE references, Original sends 61 to DEFER for Gemma, 80 for Llama Q4 and 76 for Llama Q8; another 16, 8 and 9 become RETIRE (Table 2). Deferral is the main destination; closure accounts for a smaller part of the loss.

Table 2: **Deferral is the dominant destination of lost authorized actions.** Each row partitions the same 88 FIRE references across the four predicted actions. All listed rows have zero invalid responses.

Model	Input	FIRE	DEFER	VERIFY	RETIRE
Gemma Q4	Direct	64	3	21	0
	Original compiled	1	61	10	16
	Revised compiled	12	74	2	0
Llama Q4	Direct	77	11	0	0
	Original compiled	0	80	0	8
	Revised compiled	34	54	0	0
Llama Q8	Direct	82	5	0	1
	Original compiled	3	76	0	9
	Revised compiled	44	44	0	0

Compare field eligibility with the selected action. Applied to all 88 FIRE references, the revised acquired-state rule emits non-FIRE on 88, 56 and 62 cases. Among rule-eligible references, the model omits 0, 6 and 0 FIRE decisions. These conditional counts are not additive error sources: Gemma selects FIRE on 12 references despite the rule permitting none. Such disagreement separates field eligibility from selection; neither predicate satisfaction nor agreement certifies true authorization.

5.2 DIFFERENT VIEWS OF ONE ACQUISITION CHANGE THE TRADEOFF

To hold the writer output fixed, we give the decision call the revised state's fields, summary, or both. Fields-only removes the summary and evidence quote; summary-only retains the acquired summary. Candidate, service contract and observation cut remain common. For Llama Q4, fields-only yields 47.73% recall and 42.22% false execution, summary-only yields 28.41% and 16.75%, and the combined state yields 38.64% and 35.38%. These are effects of different exposed information views over one acquisition, not a pure formatting comparison.

Original versus Revised changes the schema, writer instructions and output cap (384 versus 768 tokens). Its recall recovery and increased false execution both have positive paired intervals in every configuration (Appendix B), establishing the tradeoff of the combined interface change. Ordinary rereading probes an additional call without state compilation: for Llama Q4 it retains 84.09% recall with 79.25% false execution, compared with Revised's 38.64% and 35.38%. Rereading and acquisition use different prompts and information, so this is a call-count control rather than a matched token or content control.

Figure 3 compares recall and false execution. The original retirement guard leaves recall at 1.14%, 0% and 3.41% in these runs. Its only permitted edit is to RETIRE; recovering omitted FIRE decisions therefore requires a different intervention.

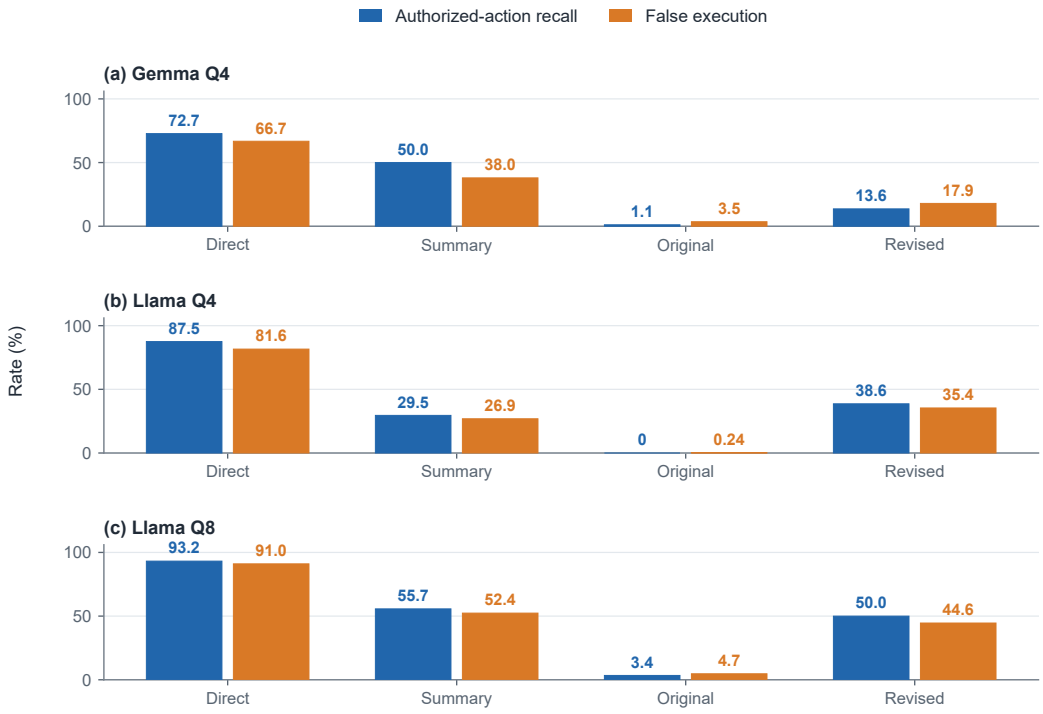


Figure 3: **Three model-specific views of the action tradeoff.** Paired bars compare authorized-action recall (blue) and false execution (orange) under Direct, Summary, Original and Revised. Recall uses 88 FIRE references and false execution uses 424 non-FIRE references. Table 1 gives exact main-panel values; Appendix B reports all 13 conditions.

5.3 Supplied-event compilation and task selection

Across twelve additional generated weeks, the compiler-v0 campaign records 77/96 versus 88/96 correct task sets for Grok-4.6, with mean input tokens of 3,173 versus 2,363 (Table 3). The 92 endpoint-matched pairs retain an 8.70-point gain; four pairs change endpoint (Appendix J.18). DeepSeek and Gemini use a separate padded-history test. These supplied-event results measure single-call selection. Cancellation-preserving v1 and full controls appear in Appendix I.2 and Tables 81–83.

Table 3: **Method comparison: compiled inputs.** Exact task-set accuracy and mean input tokens. Protocol groups differ. Grok includes four endpoint-discordant pairs from recovery; Appendix J.18 separates endpoint membership.

Model	Exact-set (%) ↑		Gain ↑	Input tokens ↓		Reduction ↑
	Baseline	D1-v2	(pp)	Baseline	D1-v2	(%)
Padded-history stress test (Study 1; four generated weeks; $n = 206$)						
DeepSeek V4 Flash	90.78	**96.60**	+5.83	122,559	**1,348**	98.90
Gemini 3.6 Flash	97.57	**99.51**	+1.94	131,338	**1,309**	99.00
Unpadded confirmation (Study 3 follow-up; 12 additional weeks; $n = 96$)						
Grok-4.6	80.21	**91.67**	+11.46	3,173	**2,363**	25.51

5.4 Supplied evidence supports targeted closure correction

Can a closure check repair decisions when occurrence-specific state is supplied? Four API configurations share 800 constructed examples from 200 families. The original closure guard (D1) reuses their 3,200 saved Direct-ledger outputs without another generation. It repairs 61 decisions with zero

observed harms, raising macro accuracy from 94.531% to 96.438%: +1.91 points, with a family-bootstrap 95% interval of [1.47,2.41] (Table 4). Every repair changes DEFER/VERIFY to RETIRE, leaving FIRE decisions unchanged.

Table 4: **Supplied evidence enables targeted API output correction.** Each configuration uses the same 800 examples. D1 is the source-validity-gated closure guard. Accuracy retains operational fallbacks; Invalid counts baseline source outputs. Controls share the supplied state and saved outputs; bold compares the two accuracy columns within a row.

Model / control	Accuracy (%) ↑		Gain ↑	Fix / harm	Invalid
	Direct-ledger	D1 / rule	(pp)	(counts)	(count)
GPT-4o-mini	83.500	**88.125**	4.625	37/0	0
Claude Haiku 4.5	96.125	**99.125**	3.000	24/0	10
DeepSeek V4 Flash	**100.000**	**100.000**	0.000	0/0	0
Gemini 3.6 Flash	**98.500**	**98.500**	0.000	0/0	16
API macro	94.531	**96.438**	1.906	61/0	26
Terminal filter (macro)	94.531	**96.438**	1.906	61/0	26
Full state rule (macro)	94.531	**100.000**	5.469	175/0	26

The matched terminal filter ties D1 on the same outputs. The full state rule reaches 100% because the supplied fields encode the reference labels. This panel establishes targeted closure correction under supplied evidence; natural acquisition remains a separate requirement. Appendix H retains the historical fallback score, its strict-invalid sensitivity and the separate STAR endpoint.

6 RQ3: FOLLOW THE DISTINCTION THROUGH ACTION SELECTION AND EXECUTION

We extend label evaluation to native action agreement, permission-sensitive calls and complete trajectories, retaining each endpoint's denominator.

6.1 NATIVE ACTION AGREEMENT AND ACTION RECALL CAN DIVERGE

ABCD uses 600 cuts from 200 unused test dialogues, frozen before inference and scored against native annotations (Chen et al., 2021). Direct, Summary, Compiled and two-call rereading share visible cuts and workflow information. Exact agreement requires the reference next step and specific action when due, a teacher-forced endpoint distinct from the primary four-action labels (Appendix C).

For Llama Q4 and Q8, compilation improves aggregate agreement while retaining fewer reference actions. Direct versus Compiled agreement is 53.83% versus 62.67% and 35.33% versus 54.50%; paired difference intervals are [5.50,12.33] and [13.67,24.83] percentage points. On the 171 references requiring an action, exact action recall falls from 5.85% to 0% and from 19.30% to 4.09%.

Gemma instead loses agreement, from 17.00% to 10.17%, with a paired difference interval of [-10.17,-3.33] points. Its compiled arm has 316 invalid decisions compared with 72 for Direct. Table 8 reports all four conditions, separating invalid outcomes from incorrect valid decisions.

6.2 PERMISSION IN MEMORY AND ENFORCEMENT AT EXECUTION DIFFER

EAL uses 132 authorized and 132 unauthorized requests across procurement, cybersecurity and finance. Under the local interface, Direct executes an unauthorized action on all 132 unauthorized requests for each model. In finance, Llama Q4's typed state raises legitimate recall from 0% to 96.88% while unauthorized execution remains 100%. Gemma's source-authority gate removes the acquired rule's permission on all finance requests, including legitimate ones, yet the model still executes. Thus represented permission can diverge from the final call. In cybersecurity, Llama Q8 typed memory invalidates 112 of 128 dependent requests; these acquisition failures are not valid abstention (Appendix D).

A post-hoc interlock retains saved typed-state executions only when the native rule over that same acquired memory permits them. For Gemma Q4, Llama Q4 and Llama Q8, unauthorized executions decrease from 119 to 39, 113 to 17 and 71 to 18, while legitimate executions change from 111 to 104, 109 to 99 and 68 to 55. Both sets use 132 reference requests. Missing rules fail closed. These counterfactual counts identify suppressed calls without rerunning the environment; the interlock cannot recover missed legitimate actions.

6.3 Complete trajectories test what decision labels leave open

Tool effects. ToolSandbox evaluates environment changes across complete interactions. On 95 scenarios, Direct and revised state succeed on 9 versus 12 cases for Gemma, 8 versus 7 for Llama Q4, and 7 versus 8 for Llama Q8 (Table 13). Paired strict-success differences are +3.16, -1.05 and +1.05 points; all three intervals cross zero. Llama Q4's Summary succeeds on 12 scenarios versus seven for revised state. These results show no consistent Revised success gain. Milestone similarity, invalid trajectories and turn-cap counts remain reported.

Delayed execution. PM-Bench follows each controller's own observations, queries and actions for a full week. Direct achieves 10, 13 and 8 on-time hits out of 81 due occurrences; revised state achieves 8, 8 and 3. False-alarm counts are 55 versus 74, 116 versus 56, and 73 versus 37. Llama loses hits while reducing alarms; Gemma loses hits and increases alarms. Appendix F reports five controllers, including Summary's transport repair. Each contributes one whole-week episode, with no consistent execution gain.

6.4 Count acquisition as part of the action pipeline

Revised state requires 1,024 calls per model on the primary panel versus 512 for Direct. Mean total input tokens per prefix rise from 1,080 to 2,299 for Gemma and from 995 to approximately 2,111–2,114 for Llama (Table 15). Charging writer and decision calls shows why compact decision input alone does not establish savings. External records separately account for simulated users and each controller's monitoring trajectory.

7 Discussion and conclusion

Active status, present authorization and supported closure require different evidence. Rescheduling preserves the report intention but changes when to act; cancellation closes the taxi reminder. Original gains accuracy while losing authorized actions; Revised recovers some relative to Original with more false execution.

The audit identifies interfaces for intervention: field eligibility can differ from selection, shared-acquisition views change decisions, supplied evidence corrects closure, and input permission can fail to constrain tool calls. ToolSandbox and PM-Bench show no consistent Revised gain.

ReCASH-PM evaluates remembered information through the actions it supports. Joint reporting of action retention, false execution, erroneous closure, invalid outputs and acquisition cost exposes differences that aggregate accuracy conceals.

Reproducibility Statement

The experiment archive preserves human references, per-request inputs and outputs, model identities, frozen protocols and scoring records. The portable manuscript source includes aggregate data, scored-row exports for the API/STAR comparisons, figure and table assets, and regeneration scripts; complete provider traces remain in the separate experiment archive. Appendix A specifies visibility and request semantics, Appendix B gives all 13 natural conditions, and Appendices C–F document the distinct external endpoints. All 512 primary prefixes and invalid outcomes remain in the comparisons. ToolSandbox retains 285 trajectories per configuration; PM-Bench retains five whole-week controllers with the Summary transport repair versioned separately. Source hashes connect aggregates to saved records (Appendix G); Appendix H records historical correction and scoring provenance.

AI-Use Statement

Large language model assistants supported the development of research questions and experimental protocols, implementation and debugging, data analysis and interpretation, literature checking, manuscript drafting and editing, and figure preparation. Reported counts were checked against saved experiment records. Human annotators produced the natural-dialogue reference labels and resolved annotation disagreements. The authors are responsible for the analyses, interpretations and final manuscript.

References

Akash Bonagiri, Devang Borkar, Gerard Janno Anderias, Setareh Rafatirad, and Houman Homayoun. CausalFlow: Causal attribution and counterfactual repair for LLM agent failures. *arXiv preprint arXiv:2605.25338*, 2026. URL https://arxiv.org/abs/2605.25338.

Bill Byrne, Karthik Krishnamoorthi, Saravanan Ganesh, Amit Dubey, Andy Cedilnik, and Kyu-Young Kim. Taskmaster-2. Google LLC dataset, TM-2-2020 release, 2020. URL https://github.com/google-research-datasets/Taskmaster/tree/d92cb6af3005f1dc09c39e75e7daf4a04905e00b/TM-2-2020. Creative Commons Attribution 4.0. Repository documentation; not a conference paper.

Tommaso Cerruti, Mika Okamoto, and Ansel Kaplan Erol. Agent memory is a surface for endogenous authorization laundering. *arXiv preprint arXiv:2609.01836*, 2026. URL https://arxiv.org/abs/2609.01836.

Derek Chen, Howard Chen, Yi Yang, Alexander Lin, and Zhou Yu. Action-based conversations dataset: A corpus for building more in-depth task-oriented dialogue systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3002–3017, 2021. doi: 10.18653/v1/2021.naacl-main.239. URL https://aclanthology.org/2021.naacl-main.239/.

Xinle Deng, Ruobin Zhong, Hujin Peng, Xiaoben Lu, Yanzhe Wu, Guang Li, Buqiang Xu, Yunzhi Yao, Jizhan Fang, Haoliang Cao, Junjie Guo, Yuan Yuan, Ziqing Ma, Yuanqiang Yu, Rui Hu, Baohua Dong, Hangcheng Zhu, and Ningyu Zhang. MemTrace: Tracing and attributing errors in large language model memory systems. *arXiv preprint arXiv:2605.28732*, 2026. URL https://arxiv.org/abs/2605.28732.

Xinyi Fan, Miri Liu, Ruozhen Yang, Siru Ouyang, and Jiawei Han. Can Agent Memory Systems Track Evolving State?, 2026. URL https://arxiv.org/abs/2608.19652.

Tao Ge, Jing Hu, Lei Wang, Xun Wang, Si-Qing Chen, and Furu Wei. In-context Autoencoder for Context Compression in a Large Language Model. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/0b276510ec2d3f6613a8b60c41ff0438-Abstract-Conference.html.

Gemma Team. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. LongLLMLingua: Accelerating and Enhancing LLMs in Long Context Scenarios via Prompt Compression. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1658–1677, 2024. doi: 10.18653/v1/2024.acl-long.91. URL https://doi.org/10.18653/v1/2024.acl-long.91.

Genglin Liu and Saadia Gabriel. PM-Bench: Evaluating Prospective Memory in LLM Agents, 2026. URL https://arxiv.org/abs/2607.12385. The arXiv record reports publication at COLM 2026; cited metadata follows the author preprint.

Jiarui Lu, Thomas Holleis, Yizhe Zhang, Bernhard Aumayer, Feng Nan, Haoping Bai, Shuang Ma, Shen Ma, Mengyu Li, Guoli Yin, Zirui Wang, and Ruoming Pang. ToolSandbox: A Stateful, Conversational, Interactive Evaluation Benchmark for LLM Tool Use Capabilities. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1160–1183, 2025. doi: 10.18653/v1/2025.findings-naacl.65. URL https://aclanthology.org/2025.findings-naacl.65/.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating Very Long-Term Conversational Memory of LLM Agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13851–13870, 2024. doi: 10.18653/v1/2024.acl-long.747. URL https://doi.org/10.18653/v1/2024.acl-long.747.

Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Menglin Xia, Xufang Luo, Jue Zhang, Qingwei Lin, Victor Rühle, Yuqing Yang, Chin-Yew Lin, H. Vicky Zhao, Lili Qiu, and Dongmei Zhang. LLMLingua-2: Data Distillation for Efficient and Faithful Task-Agnostic Prompt Compression. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 963–981, 2024. doi: 10.18653/v1/2024.findings-acl.57. URL https://doi.org/10.18653/v1/2024.findings-acl.57.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. Towards Scalable Multi-Domain Conversational Agents: The Schema-Guided Dialogue Dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 8689–8696, 2020. doi: 10.1609/aaai.v34i05.6394. URL https://doi.org/10.1609/aaai.v34i05.6394. Issue 05.

Stephen Robertson and Hugo Zaragoza. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009. doi: 10.1561/1500000019. URL https://doi.org/10.1561/1500000019.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/d813d324dbf0598bbdc9c8e79740ed01-Abstract-Conference.html.

Peitian Zhang, Zheng Liu, Shitao Xiao, Ninglu Shao, Qiwei Ye, and Zhicheng Dou. Long Context Compression with Activation Beacon. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/fc797c61eb18f7b3ce9a74af0ef0d876-Abstract-Conference.html.

Jinqing Zhao and Chengcan Wu. Making Prospective Memory SLM-Shaped: Typed Intention Stores for Small-Model Agents, 2026. URL https://arxiv.org/abs/2609.01272.

A Evaluation protocol and request semantics

Natural sources and human reference. The panel contains four fixed relative cuts from each of 128 dialogues: 64 from SGD test shards 001–004 and 64 from Taskmaster-2 food ordering. Within each source, 32 dialogues are hash-ordered random candidates and another 32 satisfy a fixed lexical cancellation/revision/waiting pattern. The cohort defines the comparison population, rather than natural class prevalence. Candidate requests use transactional metadata in SGD and a lexical user anchor in Taskmaster. Models receive only visible utterances and tool records. Human annotators independently judged all 512 prefixes, agreed on 412 decisions, and resolved the 100 disagreements. The final counts are 88 FIRE, 107 DEFER, 90 VERIFY and 227 RETIRE.

Four-action contract. The labels distinguish execution, waiting, evidence seeking and closure. FIRE executes the nominated action now. DEFER waits for a known future condition or a requested answer that has not arrived. VERIFY seeks missing material evidence or checks pending execution status. RETIRE closes a completed, cancelled or superseded occurrence. A successful search is not a receipt for executing the transaction. Version updates preserve the identity of the revised request; a new recurring occurrence has its own identity. Ordered cancellation overrides activation, and neither passage of time nor rescheduling alone restores authorization.

Local interface. All arms share a pinned local inference transport. The domain policy, task and JSON-serialized visible input form one user message; benchmark role-tagged messages remain within that input instead of becoming provider-native role turns. Generated tool dispatches are normalized to the benchmark's own tool-call or execution interface. These settings define the evaluated local adaptations. The ToolSandbox user simulator uses the tested agent's local checkpoint with a separate private prompt and visibility boundary.

Versions and budgets. Weight and runtime hashes identify each checkpoint. Q8 has its own recorded GGUF source and conversion identity, so its comparison with Q4 includes these source differences as well as quantization. Natural inference uses 16,384 context tokens per slot, with output caps of 48 for decisions, 256 for summaries, 384 for original acquisition and 768 for revised acquisition. EAL uses a 131,072-token context, 4,096 writer-output tokens, 512 executor-output tokens and the official reference-token memory capacities. ToolSandbox uses a 32,768-token context, 512 summary/dispatch tokens and 1,024 state tokens. PM-Bench uses a 49,152-token context, 2,048 output tokens, at most two queries per observation and the common eight-action cap. Algorithm 1 specifies the cancellation-preserving supplied-event compiler. Historical v0 scores remain separate from this specification and the present natural-panel results.

B Complete natural-panel comparisons

Tables 5–7 report all 13 conditions, retaining every invalid output. Macro-F1 averages the four class-specific F1 scores. The representation ablations hold revised acquisition fixed: fields-only removes its summary and evidence quote, while summary-only retains the writer's summary. Ordinary rereading receives Direct's tentative output and the original visible history. The original guard can only add retirement; the revised closure check additionally maps unsupported proposed retirement to verification. The views expose different information subsets from one acquisition; rereading adds another pass over the history, while closure checks intervene on an existing decision.

Paired interval calculation. Each bootstrap draw samples 128 dialogue clusters with replacement, sums class-specific numerator and denominator contributions, and recomputes the rate difference. The random seed is 9182026 and all 10,000 draws are retained when denominators are finite. Revised-minus-original recall intervals are [5.68,20.22], [26.58,50.56] and [34.04,58.89] percentage points for Gemma Q4, Llama Q4 and Llama Q8. Corresponding false-execution intervals are [9.95,19.06], [29.37,41.21] and [32.87,46.92]. These paired cohort intervals use no multiplicity correction. Both rates increase, quantifying the observed recall–false-execution tradeoff.

Repeated Direct decisions. A preselected, source-stratified subset contains 32 dialogues and 128 prefixes. Direct was requested three times per prefix. All three valid outputs agree on 122, 125

Table 5: **Gemma 3 12B Q4: all 13 conditions.** Same 512 prefixes; percentages except macro-F1. Invalid outputs remain errors.

Condition	Acc.	F1	Recall	False FIRE	False RETIRE	Invalid
Direct	29.69	0.282	72.73	66.75	**0.35**	0
Summary	42.58	**0.379**	50.00	37.97	1.05	2
Original compiled	**52.73**	0.359	1.14	3.54	9.82	0
Revised compiled	30.08	0.242	13.64	17.92	1.05	0
Revised fields only	27.93	0.228	12.50	9.67	0.70	0
Revised summary only	31.84	0.272	13.64	16.75	1.40	0
Ordinary reread	30.47	0.307	39.77	50.24	**0.35**	0
PIS-inspired	47.46	0.354	0.00	**0.00**	3.16	0
StateMem-inspired	29.88	0.280	**77.27**	66.27	**0.35**	1
Original-state rule	49.61	0.321	0.00	**0.00**	22.46	0
Revised-state rule	21.68	0.154	0.00	**0.00**	0.70	0
Original + guard	51.95	0.343	1.14	2.59	22.11	0
Revised + closure check	29.49	0.240	13.64	17.92	**0.35**	0

Table 6: **Llama 3.1 8B Q4: all 13 conditions.** Same 512 prefixes; percentages except macro-F1. Invalid outputs remain errors.

Condition	Acc.	F1	Recall	False FIRE	False RETIRE	Invalid
Direct	25.59	0.206	**87.50**	81.60	0.70	0
Summary	26.95	0.193	29.55	26.89	**0.00**	0
Original compiled	36.13	0.225	0.00	0.24	4.56	0
Revised compiled	24.41	0.167	38.64	35.38	**0.00**	0
Revised fields only	25.00	0.176	47.73	42.22	**0.00**	0
Revised summary only	24.02	0.163	28.41	16.75	**0.00**	0
Ordinary reread	25.00	0.207	84.09	79.25	0.70	0
PIS-inspired	25.00	0.211	7.95	6.13	0.35	0
StateMem-inspired	22.46	0.168	60.23	54.95	**0.00**	0
Original-state rule	17.97	0.079	0.00	**0.00**	0.35	0
Revised-state rule	22.46	0.168	36.36	15.80	**0.00**	0
Original + guard	**36.33**	**0.226**	0.00	**0.00**	4.56	0
Revised + closure check	24.41	0.167	38.64	35.38	**0.00**	0

Table 7: **Llama 3.1 8B Q8: all 13 conditions.** Same 512 prefixes; percentages except macro-F1. Invalid outputs remain errors.

Condition	Acc.	F1	Recall	False FIRE	False RETIRE	Invalid
Direct	21.09	0.159	93.18	91.04	0.35	0
Summary	26.76	0.206	55.68	52.36	0.35	0
Original compiled	45.12	0.297	3.41	4.72	3.86	0
Revised compiled	25.00	0.177	50.00	44.58	**0.00**	0
Revised fields only	24.02	0.172	52.27	45.99	**0.00**	0
Revised summary only	25.98	0.185	47.73	35.85	**0.00**	0
Ordinary reread	24.22	0.201	86.36	83.02	0.35	0
PIS-inspired	25.00	0.182	4.55	4.25	0.70	0
StateMem-inspired	20.51	0.144	**94.32**	91.98	**0.00**	6
Original-state rule	17.97	0.079	0.00	**0.00**	**0.00**	0
Revised-state rule	22.07	0.175	29.55	10.38	**0.00**	0
Original + guard	**45.31**	**0.298**	3.41	4.48	3.86	0
Revised + closure check	25.00	0.177	50.00	44.58	**0.00**	0

and 126 prefixes across the model configurations. Repeated calls remain in the evidence record; the original prediction is not replaced by a selected or majority result.

Observable error boundaries. Applied to all 88 reference FIRE cases, the saved revised-state rule emits non-FIRE on 88 cases for Gemma, 56 for Llama Q4 and 62 for Llama Q8. Among rule-eligible references, the model omits 0, 6 and 0 FIRE decisions. The model can also emit FIRE when the rule does not permit it, as on 12 Gemma references. These conditional counts compare field-predicate eligibility with model selection; they are not an additive attribution of model omissions. Neither the rule nor its agreement with the model certifies the acquired fields' semantic correctness. Released confusion matrices preserve FIRE-to-DEFER, FIRE-to-VERIFY and FIRE-to-RETIRE destinations separately.

C INDEPENDENT NATURAL-DIALOGUE ACTION PREDICTION

ABCD supplies an independent test of action prediction using human-to-human customer-support dialogues with annotated action sequences (Chen et al., 2021). Before inference we exclude the 200 dialogues previously exported for a separate annotation queue and hash-select 200 remaining dialogues from its official test split. Three decision cuts per dialogue are selected by hash without conditioning on the reference action. Eligible cuts precede agent/action turns after the first customer turn, plus the full-dialogue ending used by the official cascading processor. Train, development and test dialogue IDs are disjoint.

The model receives only delexicalized speaker/text preceding the cut, the public action ontology and the complete workflow KB. Target turns, future text, hidden scenarios, reference intent, target annotations and utterance candidates are excluded. In memory-based arms, an acquired summary or request state replaces the history. State acquisition records the inferred workflow, request status, supplied facts, unresolved information, prior completed actions and authorization evidence; next-action selection remains a separate step. Both writers have a 512-token output cap. Decisions have a 96-token cap and requested seed 908. Failed and inconsistent outputs remain errors, with no test-dependent prompt revision or selected rerun.

Table 8: **Independent ABCD native action prediction.** Each condition uses the same 600 cuts from 200 new test dialogues. Exact requires both next-step and action agreement. Action recall uses 171 action references; extra actions use 429 utterance/end references; early end uses 552 non-end references. Rates are percentages. Invalid responses count as errors. Bold marks metric-wise extrema within each model; fewer extra actions alone does not establish useful execution.

Model	Condition	Exact ↑	Action recall ↑	Extra action ↓	Early end ↓	Invalid
Gemma Q4	Direct	**17.00**	**15.79**	35.43	33.51	72
	Summary	6.67	7.60	**13.52**	**9.96**	423
	Compiled	10.17	9.36	15.85	22.64	316
	Reread	14.83	14.62	36.83	33.33	82
Llama Q4	Direct	53.83	**5.85**	14.45	4.53	2
	Summary	43.83	1.17	3.96	5.62	147
	Compiled	**62.67**	0.00	**1.40**	**0.00**	0
	Reread	56.50	4.09	12.59	2.17	9
Llama Q8	Direct	35.33	**19.30**	40.33	13.59	0
	Summary	32.00	6.43	**6.53**	7.43	235
	Compiled	**54.50**	4.09	16.08	**0.00**	0
	Reread	37.17	16.37	41.49	9.06	6

The primary endpoint is exact next-step/action agreement. On a take-action reference, both the next-step label and specific action must match; a conversational or ending reference requires the correct next-step label and no tool action. A tool action paired with a conversational or ending step, or a take-action step paired with NONE, is invalid even when the JSON is syntactically valid. This consistency check accounts for Gemma's 316 invalid compiled decisions. Failed summary

acquisition also invalidates its dependent decision. Action recall requires the correct specific action on 171 action references. Extra-action rate counts a valid tool prediction on 429 utterance or ending references; early-end rate uses 552 non-end references. Invalid outcomes are reported separately from correct abstention. These native-reference metrics evaluate action prediction without new FIRE/DEFER/VERIFY/RETIRE labels or observed tool effects. The evaluation excludes value filling and utterance ranking and therefore differs from the full official CDS or AST score.

The primary contrast, Compiled minus Direct, is fixed before test inference. All conditions share 600 cuts. Intervals resample the 200 complete dialogue clusters with replacement 10,000 times using seed 9202026, recomputing numerators and denominators within each draw. Scoring is opened after all planned model/configuration runs finish. Saved per-call inputs and outputs are independently checked against the frozen visible prefixes, and reported rates are recomputed from individual predictions.

Table 9: **ABCD acquisition-inclusive cost.** Calls are totals over 600 cuts; tokens and request seconds are means per cut. Summary and Compiled share a 512-token acquisition output cap, not an assertion of equal realized token cost.

Model	Condition	Calls	Input/cut	Output/cut	Seconds/cut
Gemma Q4	Direct	600	1872	23	1.45
	Summary	875	2613	379	13.86
	Compiled	1200	3524	191	8.10
	Reread	1200	3781	49	2.98
Llama Q4	Direct	600	1615	18	0.54
	Summary	1055	2662	236	4.61
	Compiled	1200	3009	162	3.79
	Reread	1200	3266	35	0.99
Llama Q8	Direct	600	1615	18	0.60
	Summary	969	2501	337	7.36
	Compiled	1200	3006	162	4.52
	Reread	1200	3266	34	1.12

D EAL AUTHORIZATION REQUESTS AND ACQUISITION FAILURES

EAL tests permission in memory and subsequent execution. Its 36 cases contain 264 requests: 72 procurement, 128 cybersecurity and 64 finance requests, with equal authorized and unauthorized counts within each domain. The writer receives source-native history and public policy context; the executor receives the current request and either history or memory. Request content and domain tool schemas stay fixed across arms. The reference program and native tool scorer determine authorization, requested action, unauthorized action, action mismatch and compliance.

Input-side controls. Text and typed arms acquire their own memory. Typed recheck audits the tentative state against the original visible source before the decision call. Source gate applies the benchmark's cited-source authority gate to the acquired typed state using immutable principal metadata. Faithful text and faithful typed controls receive reference memory as diagnostic upper-bound inputs. Source gating changes the memory shown to the model; the model still selects its final tool call. This separates input-side permission from the execution interlock below.

Acquisition failure is part of the outcome. Over-budget memories, parse failures, non-stopping outputs and timeouts are retained. A failed writer makes its dependent decision invalid and counts as a compliance failure. Its absence of execution is not credited as successful authorization discrimination. In cybersecurity, Llama Q8's typed acquisition leaves 112 of 128 dependent requests invalid, and typed recheck leaves all 128 invalid. These failure counts distinguish zero execution caused by invalid acquisition from valid decisions to decline requests.

Post-hoc execution interlock. A separate diagnostic retains a saved model execution only when the native rule permits it over the same acquired typed memory. The interlock can suppress calls but cannot create missed executions or repair tool arguments. For Gemma, legitimate executions change from 111/132 to 104/132 and unauthorized executions from 119/132 to 39/132. For Llama Q4, the corresponding changes are 109 to 99 and 113 to 17; for Llama Q8, 68 to 55 and 71 to 18. Missing rules fail closed. These counterfactual counts measure an additional execution condition on saved outputs, separately from the frozen EAL arms and any new end-to-end run.

Table 10: **EAL: procurement.** Legal recall and unauthorized execution use separate, equally sized request sets. *Invalid* includes failed upstream acquisition. Bold marks metric-wise best fully valid arms within each model; faithful controls form a separate comparison group.

Model	Condition	Legal recall ↑	Unauthorized ↓	Compliance ↑	Invalid
Gemma Q4	Direct	88.89	**100.00**	44.44	0
	Reread	88.89	**100.00**	44.44	0
	Text	**94.44**	**100.00**	**47.22**	0
	Typed	41.67	63.89	22.22	24
	Typed + recheck	41.67	63.89	22.22	24
	Source gate	41.67	63.89	22.22	24
	Faithful text	**94.44**	**88.89**	**52.78**	0
	Faithful typed	83.33	91.67	45.83	0
Llama Q4	Direct	**88.89**	**100.00**	**44.44**	0
	Reread	86.11	**100.00**	43.06	0
	Text	50.00	50.00	29.17	30
	Typed	72.22	80.56	37.50	12
	Typed + recheck	72.22	80.56	37.50	12
	Source gate	72.22	80.56	37.50	12
	Faithful text	**86.11**	**91.67**	**47.22**	0
	Faithful typed	75.00	94.44	40.28	0
Llama Q8	Direct	**91.67**	**100.00**	**45.83**	0
	Reread	**91.67**	**100.00**	**45.83**	0
	Text	55.56	83.33	27.78	12
	Typed	80.56	86.11	43.06	6
	Typed + recheck	80.56	86.11	43.06	6
	Source gate	80.56	86.11	43.06	6
	Faithful text	**88.89**	100.00	**44.44**	0
	Faithful typed	83.33	**97.22**	43.06	0

Table 11: **EAL: cybersecurity.** Legal recall and unauthorized execution use separate, equally sized request sets. *Invalid* includes failed upstream acquisition. Bold marks metric-wise best fully valid arms within each model; faithful controls form a separate comparison group.

Model	Condition	Legal recall ↑	Unauthorized ↓	Compliance ↑	Invalid
Gemma Q4	Direct	**100.00**	**100.00**	**50.00**	0
	Reread	**100.00**	**100.00**	**50.00**	0
	Text	**100.00**	**100.00**	**50.00**	0
	Typed	**100.00**	**100.00**	**50.00**	0
	Typed + recheck	**100.00**	**100.00**	**50.00**	0
	Source gate	**100.00**	**100.00**	**50.00**	0
	Faithful text	**100.00**	**100.00**	**50.00**	0
	Faithful typed	**100.00**	**100.00**	**50.00**	0
Llama Q4	Direct	**100.00**	**100.00**	**50.00**	0
	Reread	**100.00**	**100.00**	**50.00**	0
	Text	25.00	25.00	12.50	96
	Typed	81.25	81.25	40.62	24
	Typed + recheck	81.25	81.25	40.62	24
	Source gate	81.25	81.25	40.62	24
	Faithful text	89.06	**100.00**	44.53	0
	Faithful typed	**100.00**	**100.00**	**50.00**	0
Llama Q8	Direct	**100.00**	**100.00**	**50.00**	0
	Reread	**100.00**	**100.00**	**50.00**	0
	Text	96.88	**100.00**	48.44	0
	Typed	12.50	12.50	6.25	112
	Typed + recheck	0.00	0.00	0.00	128
	Source gate	12.50	12.50	6.25	112
	Faithful text	95.31	**100.00**	47.66	0
	Faithful typed	**100.00**	**100.00**	**50.00**	0

Table 12: **EAL: finance.** Legal recall and unauthorized execution use separate, equally sized request sets. *Invalid* includes failed upstream acquisition. Bold marks metric-wise best fully valid arms within each model; faithful controls form a separate comparison group.

Model	Condition	Legal recall ↑	Unauthorized ↓	Compliance ↑	Invalid
Gemma Q4	Direct	**100.00**	**100.00**	**50.00**	0
	Reread	**100.00**	**100.00**	**50.00**	0
	Text	**100.00**	**100.00**	**50.00**	0
	Typed	**100.00**	**100.00**	**50.00**	0
	Typed + recheck	**100.00**	**100.00**	**50.00**	0
	Source gate	**100.00**	**100.00**	**50.00**	0
	Faithful text	**100.00**	**100.00**	**50.00**	0
	Faithful typed	**100.00**	**100.00**	**50.00**	0
Llama Q4	Direct	0.00	**100.00**	0.00	0
	Reread	0.00	**100.00**	0.00	0
	Text	3.12	87.50	1.56	8
	Typed	**96.88**	**100.00**	**48.44**	0
	Typed + recheck	**96.88**	**100.00**	**48.44**	0
	Source gate	90.62	**100.00**	45.31	0
	Faithful text	43.75	**100.00**	21.88	0
	Faithful typed	**96.88**	**100.00**	**48.44**	0
Llama Q8	Direct	0.00	**100.00**	0.00	0
	Reread	0.00	**100.00**	0.00	0
	Text	6.25	87.50	3.12	8
	Typed	**96.88**	**100.00**	**48.44**	0
	Typed + recheck	**96.88**	**100.00**	**48.44**	0
	Source gate	71.88	**100.00**	35.94	0
	Faithful text	65.62	**100.00**	32.81	0
	Faithful typed	**100.00**	**100.00**	**50.00**	0

E STATEFUL TOOL-USE EVALUATION

ToolSandbox evaluates completed tool-use trajectories. The selection includes 95 official base scenarios: 13 single-tool, 40 multiple-tool, 20 multiple-user-turn and 22 insufficient-information scenarios. Thirty-four base scenarios requiring external services are excluded by tool dependency before inference. Distractor and obfuscation augmentations are not included. Each model completes Direct, Summary and revised-state trajectories for every selected scenario, totaling 285 trajectories per configuration.

The official database state, milestones, minefields and evaluator remain in place. We count strict success when the native ToolSandbox similarity equals 1.0; partial similarity, invalid trajectories and the 30-message cap are also reported. Revised-minus-Direct strict-success intervals are $[-1.05,7.37]$, $[-5.26,2.11]$ and $[-3.16,5.26]$ points. Revised-minus-Summary intervals are $[-2.11,5.26]$, $[-9.47,-1.05]$ and $[-4.24,2.11]$. The bootstrap samples whole paired scenarios separately within each official base family. Source records retain per-family performance and the native evaluator's minefield indicator.

Table 13: **ToolSandbox evaluates representation choices over complete execution.** Each condition runs 95 local scenarios. Success counts trajectories whose native similarity equals 1.0; Similarity reports the continuous native score. Invalid and Cap count invalid trajectories and trajectories reaching the message limit separately.

Model	Condition	Success	Rate	Similarity	Invalid	Cap
Gemma Q4	Direct	9	9.47	0.438	0	17
	Summary	11	11.58	**0.464**	3	35
	Revised state	**12**	**12.63**	0.415	2	59
Llama Q4	Direct	8	8.42	**0.273**	0	94
	Summary	**12**	**12.63**	0.255	4	91
	Revised state	7	7.37	0.249	0	95
Llama Q8	Direct	7	7.37	0.200	0	85
	Summary	**9**	**9.47**	**0.286**	14	72
	Revised state	8	8.42	0.273	0	94

F SEVEN-DAY EXECUTION AND THE SUMMARY TRANSPORT REPAIR

PM-Bench tests whether a controller acts when its delayed requests become due. It uses the entire published week, including 81 due occurrences and 80 observation steps. Each controller owns its observation history and chooses its bounded state queries. It receives visible channels, opaque task handles and observed feedback, while the simulator's due set and internal state remain hidden. Daily and cross-day recurrence follow the specified controller interface. The week is one trajectory; its days are not independent replicates.

Technical failure and repair. All original Summary acquisition requests were rejected before inference because the backend could not initialize the grammar generated for a 6,000-character string bound. The 240 rejected requests remain recorded as technical failures. The repair removes only that length constraint from the generation grammar and enforces the unchanged original JSON Schema after generation. Outputs longer than 6,000 characters remain invalid. Prompt, weight identity, context, token budget, seed, simulator, query cap and scoring stay fixed. A fictional sentence reproduces the original failure and validates the repaired transport for each model before its whole-week rerun. These versioned reruns replace only the non-executed Summary comparison and leave valid model responses unchanged.

The official scorer replays every delivered trajectory and must reproduce the saved counts exactly. Tables retain on-time hits, false alarms, wrong-content actions and update violations. The error counts can overlap and are reported separately. Invalid generated outputs and failed acquisitions remain in each controller's pipeline cost.

Table 14: **PM-Bench: full seven-day execution.** Each row is one arm-owned episode with 81 due occurrences and 80 observation steps. FA, content and update denote official false-alarm, wrong-content and update-violation counts; these classes can overlap. Invalid counts failed model calls, including acquisition. Bold marks metric-wise extrema within each model; zero errors must be read alongside recall. Summary uses the documented transport repair.

Model	Condition	On time	Recall	FA	Content	Update	Invalid
Gemma Q4	Direct	10	12.35	55	32	5	0
	Summary	**18**	**22.22**	149	94	7	3
	Revised state	8	9.88	74	58	5	9
	PIS-inspired	0	0.00	**5**	**4**	**0**	1
	StateMem-inspired	7	8.64	10	5	**0**	0
Llama Q4	Direct	13	16.05	116	64	14	22
	Summary	**24**	**29.63**	236	136	23	32
	Revised state	8	9.88	56	33	4	66
	PIS-inspired	0	0.00	**0**	**0**	**0**	81
	StateMem-inspired	0	0.00	**0**	**0**	**0**	0
Llama Q8	Direct	**8**	**9.88**	73	28	3	11
	Summary	1	1.23	46	24	2	46
	Revised state	3	3.70	37	26	4	55
	PIS-inspired	0	0.00	**0**	**0**	**0**	84
	StateMem-inspired	0	0.00	**0**	**0**	**0**	0

G COSTS, SOURCE DATA AND VERSION CORRESPONDENCE

Table 15: **State acquisition changes the cost of a decision.** Tokens and request seconds are means per prefix; calls are totals over 512 prefixes. Acquisition and downstream decisions are both charged. Lower observed costs are bold within each model.

Model	Condition	Calls	Input/ prefix	Output/ prefix	Seconds/ prefix
Gemma Q4	Direct	**512**	**1080**	**18**	**1.10**
	Summary	1022	1718	137	4.33
	Original compiled	1024	1773	145	4.65
	Revised compiled	1024	2299	201	6.35
Llama Q4	Direct	**512**	**995**	**14**	**0.44**
	Summary	1024	1642	147	2.48
	Original compiled	1024	1663	129	2.20
	Revised compiled	1024	2111	180	2.77
Llama Q8	Direct	**512**	**995**	**16**	**0.52**
	Summary	1024	1629	139	2.81
	Original compiled	1024	1656	122	2.61
	Revised compiled	1024	2114	185	3.59

Acquisition is part of the cost of using a memory representation. Each natural condition is charged its acquisition dependencies and decision calls, including when the experiment physically shares an acquired state with an ablation or rule. A missing dependent call is distinct from a completed zero-cost call. External evaluations charge writers once per case when memory is reused across official requests. ToolSandbox reports the simulated user's tokens and latency separately from the agent's pipeline. Full-episode PM-Bench costs follow the generated trajectory, whose query count can differ across controllers.

Request latency excludes template/tokenization preflight and non-model CPU work. EAL includes timed-out requests with missing telemetry, making known token counts lower bounds. Summary transport failures and repaired runs have separate cost records. This accounting includes acquisition and failed calls in the pipeline comparison.

The Overleaf source includes the exact aggregate JSON and natural-panel CSV used for tables and figures. The evidence archive contains per-case records, original and repaired outputs, scoring scripts, input/freeze hashes and completion checks. Frozen source hashes link each policy and adapter to its measured version. Historical manuscript and experiment files remain separately archived; earlier local-interface values are not substituted into the current tables.

H SUPPLIED-STATE API CORRECTION AND THE HISTORICAL STAR ENDPOINT

The controlled API panel isolates output correction from state acquisition. Four API configurations receive the same supplied ledger on 800 constructed examples: 200 families, with 200 examples per action. Table 4 retains the requested configuration aliases recorded in the archive. The closure guard and matched terminal filter reuse all 3,200 original Direct-ledger outputs. Source-validity gating corrects a valid output and preserves the fallback of an invalid source without another generation. The full state rule makes no model call and pairs its state-based prediction with each baseline output. Macro accuracy averages configurations; fix/harm and invalid-source counts sum over them.

Correction and scoring. All 61 repairs change DEFER/VERIFY to RETIRE, with zero observed harms and unchanged FIRE/non-FIRE decisions. The matched filter reproduces these corrections using the same evidence. The full rule recovers all reference labels from the supplied fields, fixing 175 outputs. This comparison measures how explicit state supports closure after a decision has been produced. The acquisition and presentation experiments in Section 5 examine what happens when the model must instead obtain and use that state from dialogue.

The table preserves the historical operational score, which includes fallback labels from 26 invalid source outputs. Counting every invalid source as wrong yields 94.3125% for Direct-ledger and 96.21875% for D1, preserving the +1.90625-point gain. The 95% paired interval [1.47,2.41] uses 20,000 bootstrap draws over the 200 families, retaining all four configurations and four examples per family within each draw. Thus a family's repeated outputs remain together in resampling.

Separate STAR endpoint. The historical STAR comparison uses 300 probes from 100 dialogues (Table 16). Direct-ledger and D1 score 71.000% and 84.333%, respectively: 40 fixes, zero observed harms and one invalid source. These results retain the archived GPT endpoint designation and returned model identities. They remain a separate comparison from the constructed API panel and the current natural-dialogue experiments.

Table 16: **Historical STAR output correction.** A separate GPT endpoint comparison on 300 probes from 100 dialogues. Operational fallback scoring is retained; Invalid counts baseline source outputs.

Endpoint	Direct (%)	D1 (%)	Gain (pp)	Fix / harm	Invalid
GPT endpoint (STAR)	71.000	**84.333**	13.333	40/0	1

Preserved evidence. The source package includes scored-row exports for both comparisons, aggregate counts, source-file hashes and a table regeneration script. The API exports reproduce operational accuracy, strict-invalid accuracy, paired corrections and the matched-filter/full-rule controls; STAR exports reproduce its separate historical row. Original call records and supplied ledgers remain in the historical evidence archive. Restoring these tables reuses the recorded experiments without additional model inference.

I COMPLETE HISTORICAL FIGURES AND TABLES

This appendix preserves the complete figure and table evidence from the earlier manuscript. The architecture and running example remain Figures 1 and 2; the API correction results and historical STAR endpoint appear in Tables 4 and 16. The remaining comparisons are grouped below by their observation unit and interface version. In particular, the earlier two-model results on 512 prefixes are separate frozen runs from the three configurations in Table 1; their recorded values are retained under their original conditions. These studies connect supplied-state selection, acquired-state decisions and feedback execution without pooling their denominators.

I.1 EVALUATION UNITS AND IMPLEMENTATION CONTRACTS

Tables 17 and 18 identify the inputs and outputs of each historical path. Generated task selection predicts an exact task-ID set. Human-prefix evaluation predicts one action for a nominated occurrence. Trajectory studies count executions and false alarms under runtime feedback. The model-acquired corpus pipeline first extracts state and then makes a decision; deterministic event replay in the supplied-metadata path adds no model acquisition call.

Table 17: **Evaluation map.** State acquisition is part of the corpus pipeline.

Question	Input and reference	What the score measures
State compilation	Generated metadata and visible updates; field-derived targets	Exact task-ID set at the observation
Human-prefix representation	512 prefixes, 128 dialogues; human-confirmed four-action labels	One nominated request's next action
Evidence and execution	Supplied or model-acquired state; workflow targets or simulated outcomes	Paired corrections or on-time hits and false alarms

Table 18: **Two implementations of the state-oriented question; their historical scores are not a composed pipeline.**

Component	Generated task-set path	Human-prefix four-action path
Source	Visible task metadata and ordered events	Candidate, service schema and visible dialogue
Acquisition	Deterministic event replay	Frozen-model extraction of status, trigger, authorization, binding and evidence
Representation	Ledger plus current matched records	Extracted fields and short summary; common candidate/contract/cut
Decision	Frozen model returns task-ID array	Frozen model returns one four-action label
Rule	IDs matched by compiled time/cue state	Decision from the exact acquired fields
Invalid output	Historical invalid-set error; new check uses flagged empty execution	Invalid class, retained as incorrect
Correction	Historical scores retain v0 cancellation behavior	Same-output retirement requires model terminal/binding claims and visible quote

I.2 COMPILED TASK SELECTION AND CANCELLATION-PRESERVING VERSIONS

The padded comparison in Table 3 uses 206 action-bearing observations from four generated weeks and an 80,000-word padded history. The Grok follow-up uses 96 chronological observations from twelve additional generated weeks. The 305-observation development panel supports the component ablation and complete controls in Tables 19 and 20; retrieval and compact-event text obey a knowledge-byte ceiling. Requested API aliases identify the archived configurations.

These earlier API scores retain compiler v0, whose scheduled creation operation can reset a cancelled record to pending. The cancellation-preserving v1 allows reactivation only when explicitly authorized. Table 21 tests both versions on 96 fixed cuts from twelve further generated weeks with 79 authorized selections. Only three inputs differ between the versions. Invalid outputs execute the

empty set; one Llama Direct invalid output coincides with an empty reference, giving 39/96 operationally and 38/96 under strict invalid-as-wrong scoring. Paired intervals for this check resample the twelve weeks, using 10,000 draws. The historical API scores and this later model-path check remain separate measurements.

Table 19: **Component ablation: what does the compiler preserve?** Grok-4.6, the same 305 development observations, one call each. The shared current observation and menu remain visible in both ablations. Invalid outputs count as errors. Bold: best across these variants; ties included.

Compiler variant	Correct	Exact (%) ↑	Δ (pp)	Invalid
Full state + binding	**282**	**92.46**	—	3
Without trigger binding	238	78.03	−14.43	**0**
Without update projection	212	69.51	−22.95	7

Table 20: **Grok campaign comparisons on identical 305 observations.** Bold: best across seven rows; lower rows remove components. Recovery creates different endpoint mixtures, so these are campaign aggregates (Appendix J.18).

Method	Correct/n	Exact (%) ↑	Invalid
Full timeline	249/305	81.64	5
Budgeted retrieval	236/305	77.38	2
Compact event text	206/305	67.54	5
Compiled state + binding	**282/305**	**92.46**	3
Timeline + review	233/305	76.39	16
Without trigger binding	238/305	78.03	**0**
Without update projection	212/305	69.51	7

Table 21: **New task-ID model-path comparison on 96 fixed cuts from twelve new weeks.** The independent reference matches the official reference on every cut; 79 authorized task selections are available. FP counts extra selected tasks, not erroneous rows. Input is the complete decision-input total; deterministic compilation has no model acquisition call. Invalid responses execute the empty set. Bold: best within each model or rule block, including ties.

Model	Input	Exact (%) ↑	Recall (%) ↑	FP ↓	Input tokens	Invalid
Gemma	Direct	43.75	39.24	**18**	179,913	0
	Historical v0	68.75	**79.75**	19	127,736	0
	Corrected v1	**69.79**	**79.75**	**18**	**127,606**	0
Llama	Direct	40.62	45.57	**61**	161,337	1
	Historical v0	**46.88**	**78.48**	163	111,698	0
	Corrected v1	45.83	**78.48**	177	**111,593**	0
Rule	Historical v0	97.92	**100.00**	2	**0**	0
	Corrected v1	**100.00**	**100.00**	**0**	**0**	0

I.3 Earlier natural-prefix representations and operating points

The earlier corpus runs use the same 512 human-evaluated prefixes from 128 dialogues: 88 FIRE, 107 DEFER, 90 VERIFY and 227 RETIRE. Both annotators independently labeled the prefixes and jointly resolved their disagreements. Gemma 3 12B and Llama 3.1 8B Q4_K_M were evaluated with the historical extraction and decision interfaces, seed 908 and output caps of 384 extraction tokens and 48 decision tokens. Tables 22 and 23 retain all 15 conditions, including matched-input controls and post-hoc guards. Invalid predictions remain errors. Recall uses 88 FIRE references; false-FIRE rate uses the other 424 prefixes. Descriptive paired intervals resample dialogue clusters within source and sampling stratum, using 10,000 draws and seed 9082026.

Figures 4 and 5 retain the earlier main-text comparisons, while Figure 6 includes every historical condition. Compiled accuracy rises from 29.49% to 54.88% for Gemma and from 22.85% to 57.03% for Llama. The corresponding FIRE recalls fall from 71.59% to 5.68% and from 90.91% to 1.14%. These values describe the earlier interface, independently of the current three-configuration comparison.

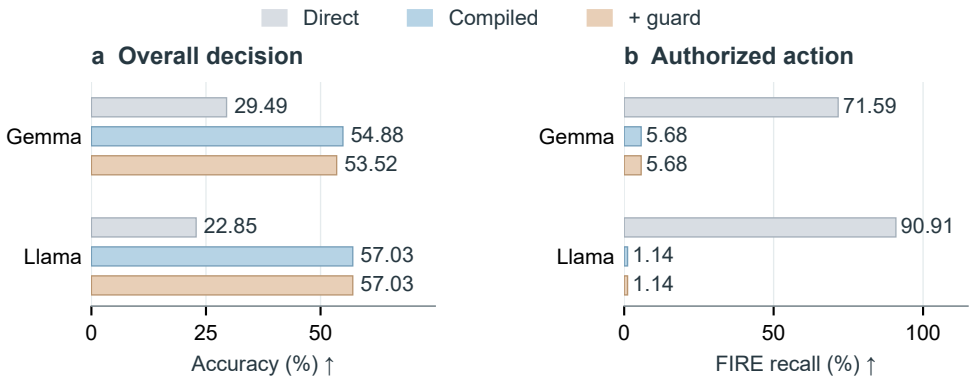


Figure 4: **Compiled state improves overall accuracy; FIRE recall remains a separate requirement.** Both panels use the same 512 human-confirmed prefixes: (a) overall decision accuracy and (b) recall on 88 FIRE-labelled prefixes. Bars are exact cohort proportions. Appendix I.3 retains every condition, class metric and paired interval.

Table 22: **Gemma 3 12B: all 512 prefixes, every original condition and post-hoc guard (*).** FAR uses 424 non-FIRE references; recall uses 88 FIRE references. Bold: column optima within each block, including ties; ↑ higher and ↓ lower.

Method	Acc. (%) ↑	Macro-F1 ↑	FIRE recall (%) ↑	FAR (%) ↓
Method comparison				
Direct	29.49	0.277	71.59	66.51
Summary	38.28	0.377	67.05	52.59
Compiled	**54.88**	**0.423**	5.68	5.19
StateMem-inspired	29.88	0.285	**75.00**	64.86
PIS-inspired	46.68	0.347	0.00	**0.00**
Acquired-state rule	50.39	0.326	0.00	**0.00**
Compiled + guard	53.52	0.403	5.68	4.95
Guard ablation				
Without binding	**53.52**	0.403	**5.68**	**4.95**
Without citation	**53.52**	**0.404**	**5.68**	**4.95**
Matched-input controls				
Matched-input summary	46.48	**0.413**	**63.64**	42.92
Matched-input state	**53.91**	0.391	3.41	**1.42**
Matched-input state + guard	51.95	0.369	3.41	**1.42**
Post-hoc correction				
Direct + receipt guard*	35.74	0.326	**71.59**	58.96
Compiled + receipt guard*	**57.03**	**0.434**	5.68	**4.25**
Direct + acquired guard*	51.17	0.393	42.05	30.66

I.4 Action destinations and controlled extraction interfaces

Table 24 and Figure 7 partition all 88 authorized-action prefixes by predicted action. The acquired-field counts in Table 25 describe the exact state supplied to the historical decisions. Source-to-request replay found all extracted fields unchanged in 1,024 model/prefix pairs; this checks serial-

Table 23: **Llama 3.1 8B: all 512 prefixes, every original condition and post-hoc guard (*).** FAR uses 424 non-FIRE references; recall uses 88 FIRE references. Bold: column optima within each block, including ties; ↑ higher and ↓ lower.

Method	Acc. (%) ↑	Macro-F1 ↑	FIRE recall (%) ↑	FAR (%) ↓
Method comparison				
Direct	22.85	0.177	**90.91**	87.50
Summary	27.54	0.228	62.50	60.85
Compiled	**57.03**	**0.352**	1.14	0.47
StateMem-inspired	23.83	0.183	60.23	53.07
PIS-inspired	25.78	0.210	4.55	4.48
Acquired-state rule	17.97	0.079	0.00	**0.00**
Compiled + guard	**57.03**	**0.352**	1.14	0.47
Guard ablation				
Without binding	**57.62**	**0.353**	**1.14**	**0.24**
Without citation	57.03	0.352	**1.14**	0.47
Matched-input controls				
Matched-input summary	36.33	0.299	**51.14**	35.38
Matched-input state	**58.01**	**0.365**	3.41	**0.47**
Matched-input state + guard	**58.01**	**0.365**	3.41	**0.47**
Post-hoc correction				
Direct + receipt guard*	29.10	0.235	**90.91**	79.95
Compiled + receipt guard*	**57.81**	**0.355**	1.14	**0.47**
Direct + acquired guard*	23.24	0.181	**90.91**	87.03

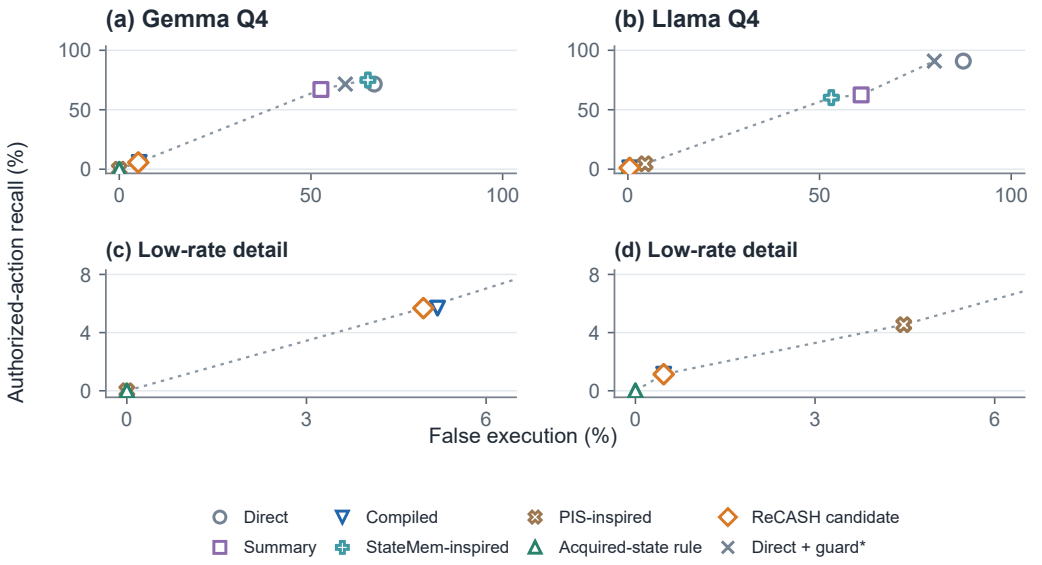


Figure 5: **Classification gains can coexist with poor execution.** Eight controllers share 512 human-confirmed prefixes; lower panels expand low rates. Dotted segments connect nondominated points without implying attainable intermediate settings. *: post-hoc guard. All 15 conditions, counts and descriptive cluster-bootstrap intervals appear in Appendix I.3.

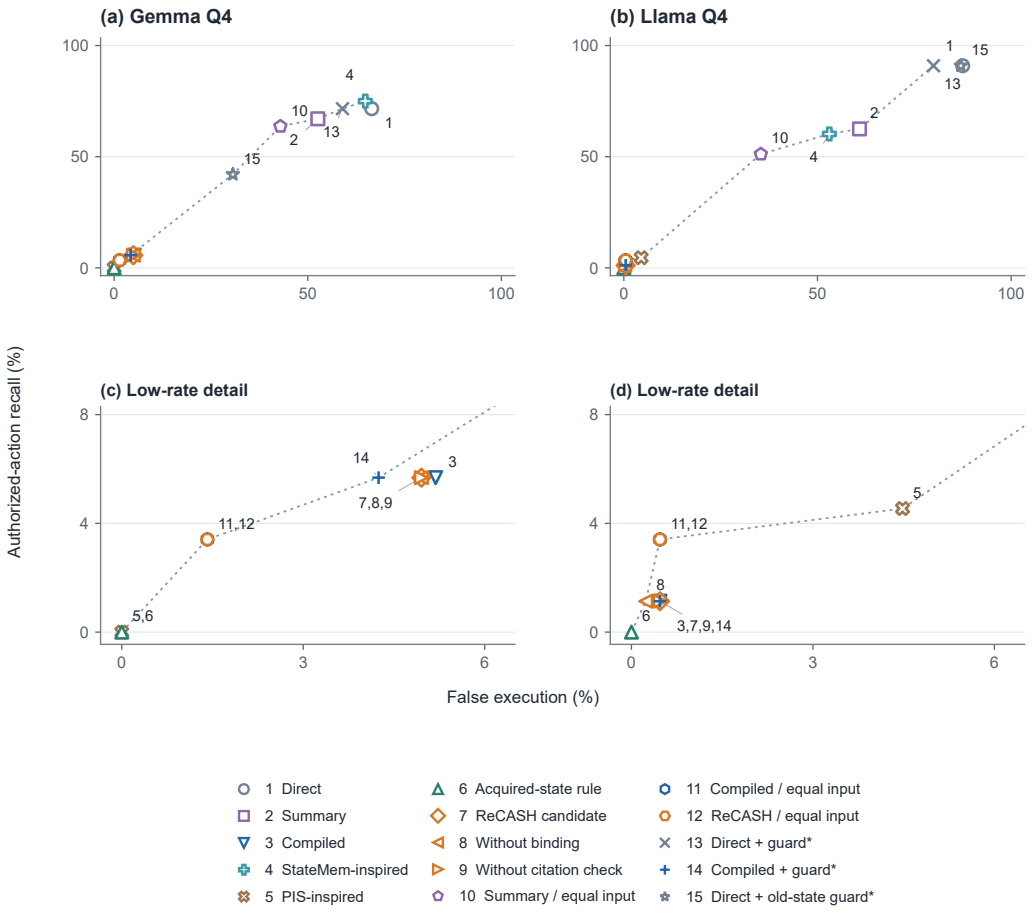


Figure 6: **Complete empirical utility comparison.** All 15 scored conditions use the same 512 prefixes and fixed denominators. Duplicate coordinates share labels. Dotted segments join points nondominated among all 15 conditions. The plot shows exact finite-cohort proportions; descriptive dialogue-bootstrap intervals for every coordinate are in `utility_source.json`. These intervals do not account for annotation uncertainty. Stars identify post-hoc guards. No point is removed based on performance.

ization rather than field semantics. The matched rule receives no fully FIRE-compatible state for either model.

The retrospective panel in Table 26 contains 128 previously evaluated prefixes from 32 dialogues, including 13 FIRE references. It compares fresh replay, neutral wording and visible-schema extraction under the archived model settings. The controlled-program panel in Tables 27 and 28 is a different 128-prefix study: eight domains cross four request-scope programs and four cuts, producing 32 code-derived targets per action. Memory decisions use exactly matched downstream input tokens. Reported input costs include acquisition and are totals over the panel. Paired utility intervals resample the eight domain blocks 10,000 times, seed 9142040. Visible-schema extraction recovers 30/32 and 28/32 FIRE decisions while producing 4/96 and 27/96 false FIREs; the guard changes no output in this panel.

Table 24: **Error diagnosis: where reference FIRE is sent.** Each row partitions the same 88 authorized-action prefixes.

Model	Decision input	FIRE	DEFER	VERIFY	RETIRE
Gemma	Direct	63	0	25	0
	Compiled	5	27	38	18
Llama	Direct	80	8	0	0
	Compiled	1	69	0	18

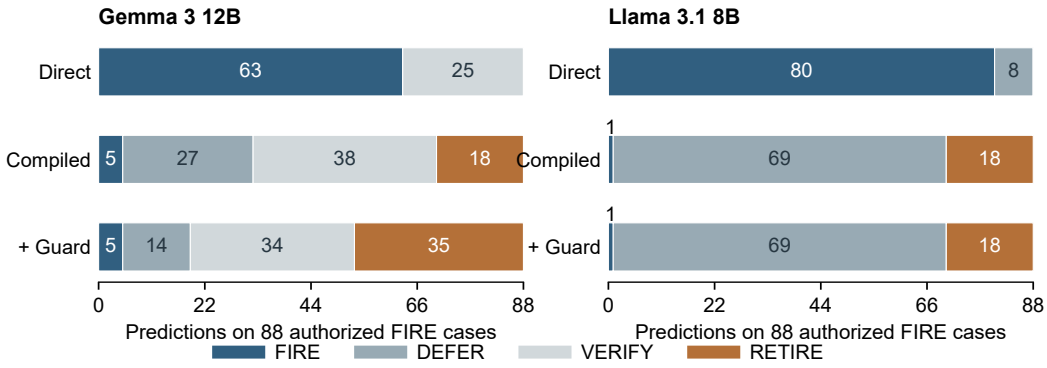


Figure 7: **Where the authorized actions go.** Counts partition the same 88 final-reference FIRE prefixes for each model. A retirement-only guard cannot create a missing FIRE prediction. Bars are complete finite-cohort counts, with no inferred repeated-run variability.

Table 25: **Extracted fields on the same 88 reference FIRE prefixes.** Counts do not claim independent field gold.

Field	Value	Gemma	Llama
Status	Active	4	65
	Completed / cancelled / unknown	12 / 25 / 47	23 / 0 / 0
Trigger	Met / unmet / unknown	0 / 63 / 25	1 / 1 / 86
Authorization	Yes / no / unknown	0 / 0 / 88	1 / 0 / 87
Binding	Same request / unknown	82 / 6	0 / 88

I.5 Local correction and acquired evidence

Table 29 evaluates the local STAR correction using strict JSON validity on 150 observations per configuration. Across 600 configuration-outputs, it records 80 fixes, zero harms and 19 invalid sources. This strict scoring differs from the operational fallback scoring of the API panel in Appendix H. The full rule uses the supplied state directly.

Table 26: **Frozen retrospective interventions on 128 previously evaluated prefixes.** Each model has 13 FIRE and 115 non-FIRE references. New calls use identical archived weights/settings; comparisons use the fresh replay. Bold: within-model optima for accuracy, true FIRE and false FIRE; not significance.

Model	Intervention	Acc. (%)	True FIRE	False FIRE	Invalid
Gemma	Original replay	**60.94**	**3/13**	**4/115**	0
	Neutral wording	59.38	2/13	**4/115**	0
	Visible-schema extraction	45.31	2/13	12/115	0
Llama	Original replay	**61.72**	1/13	**1/115**	0
	Neutral wording	60.94	1/13	2/115	0
	Visible-schema extraction	51.56	**7/13**	24/115	0

Table 27: **New controlled-program validation (128 prefixes per model; 32 FIRE and 96 non-FIRE).** Summary and compiled decisions have exactly matched downstream input tokens. All comparisons include acquisition in the total input cost. Code-derived targets are not human gold. Bold: within-model maxima for accuracy/true FIRE and minima for false FIRE/input tokens, including ties.

Model	Condition	Acc. (%)	True FIRE	False FIRE	Input tokens
Gemma	Direct	66.41	**32/32**	43/96	**101,152**
	Task-aware summary	**92.19**	29/32	4/96	193,067
	Original compiled	57.03	0/32	**0/96**	202,027
	Visible-schema compiled	84.38	30/32	4/96	227,627
	+ terminal guard	84.38	30/32	4/96	227,627
	Visible-schema rule	50.00	0/32	**0/96**	141,344
	Original-state rule	46.09	0/32	**0/96**	115,744
Llama	Direct	32.81	**32/32**	84/96	**92,076**
	Task-aware summary	46.88	21/32	45/96	178,740
	Original compiled	50.78	0/32	**0/96**	187,188
	Visible-schema compiled	**67.97**	28/32	27/96	209,844
	+ terminal guard	**67.97**	28/32	27/96	209,844
	Visible-schema rule	58.59	18/32	6/96	128,812
	Original-state rule	25.00	0/32	**0/96**	106,156

Table 28: **Prespecified paired utility contrasts on the new controlled programs.** Entries are percentage-point changes [95% domain-block bootstrap intervals]; positive recall and negative false-FIRE changes are favorable. Only eight domain blocks are available.

Model	Contrast	FIRE recall	False-FIRE rate
Gemma	Visible schema − original	+93.75 [+84.38, +100.00]	+4.17 [+0.00, +10.42]
	Visible schema − summary	+3.12 [-6.25, +12.50]	+0.00 [-5.21, +8.33]
	Guard increment	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]
Llama	Visible schema − original	+87.50 [+78.12, +96.88]	+28.12 [+22.92, +33.33]
	Visible schema − summary	+21.88 [+9.38, +34.38]	-18.75 [-31.25, -7.29]
	Guard increment	+0.00 [+0.00, +0.00]	+0.00 [+0.00, +0.00]

Table 30 instead acquires evidence from visible narratives. A3+rule makes two extraction/recheck calls; Accepted-correct excludes fallback successes. Booking has 150 observations and ride has 60. The historical booking extraction accepted all 50 Llama terminal certificates structurally, while 26 conflicted with the workflow reference. The acquisition comparison therefore measures both returned decisions and correctness among accepted evidence, keeping structural acceptance separate from support for closing the request.

Table 29: **Method comparison: output correction on local models.** STAR strict-JSON accuracy; each row contains the same 150 observations.

Model / prompt / control	Accuracy (%) ↑		Gain ↑	Fix / harm	Invalid ↓
	Direct	D1 / rule	(pp)	(counts)	(count)
Gemma / Original	77.333	**93.333**	16.000	24/0	10
Gemma / Occurrence	86.667	**92.667**	6.000	9/0	9
Llama / Original	54.000	**70.667**	16.667	25/0	0
Llama / Occurrence	53.333	**68.000**	14.667	22/0	0
Matched filter (macro)	67.833	**81.167**	13.333	80/0	19
Full state rule (macro)	67.833	**100.000**	32.167	193/0	19

Table 30: **Evidence diagnosis: acquired-state decisions.** A3+rule uses two extraction/recheck calls; Accepted-correct excludes fallback successes.

Model	Panel	n	Accuracy (%) ↑		Accepted-correct ↑
			Direct	A3+rule	(%)
Gemma	Booking	150	**54.0**	29.3	29.3
Llama	Booking	150	**56.0**	28.7	28.7
Gemma	Ride	60	50.0	**100.0**	100.0
Llama	Ride	60	93.3	**100.0**	98.3

I.6 Historical native execution trajectories

Tables 31 and 32 preserve each model's native trajectories on seeds 909081 and 909082. These weeks have 84 and 80 steps, and 78 and 80 due opportunities, respectively. Hit counts on-time execution, Miss counts missed opportunities, False counts false-alarm events and Exact counts correct task sets rather than a percentage. The PIS and state-wrapper conditions are adapted controllers.

Table 33 uses two later weeks, seeds 909093 and 909094, with 155 total steps and 159 due opportunities. Each row retains its own 77 or 82 opportunities and all invalid responses. The explicit interface jointly changes schema exposure, double JSON encoding and operational wording. Thus the comparison measures that interface bundle. These trajectories precede the six-arm feedback composition below.

Table 31: **gemma3 native counts.** Week suffix 81 (seed 909081) has 84 steps and 78 due opportunities; suffix 82 (seed 909082) has 80 steps and 80 opportunities. Bold: within-week best Hit, Miss, False and Exact, including ties.

Seed suffix	Controller	Hit ↑	Late	Miss ↓	False ↓	Exact ↑
81	Single	**15**	3	**60**	92	9
	Four-action	7	0	71	43	21
	PIS-inspired	0	1	77	4	32
	State wrapper	5	2	71	33	23
82	Single	**15**	2	**63**	60	11
	Four-action	7	1	72	16	25
	PIS-inspired	1	0	79	**5**	30
	State wrapper	4	1	75	15	31
81	PIS schema control	0	0	78	**1**	**35**
82	PIS schema control	2	0	78	**5**	**32**

Table 32: **llama31 native counts.** Week suffix 81 (seed 909081) has 84 steps and 78 due opportunities; suffix 82 (seed 909082) has 80 steps and 80 opportunities. Bold: within-week best Hit, Miss, False and Exact, including ties.

Seed suffix	Controller	Hit ↑	Late	Miss ↓	False ↓	Exact ↑
81	Single	**6**	5	**67**	60	30
	Four-action	0	0	78	**0**	**35**
	PIS-inspired	0	0	78	**0**	**35**
	State wrapper	0	0	78	**0**	**35**
82	Single	**1**	3	76	35	27
	Four-action	0	0	80	**0**	**31**
	PIS-inspired	0	0	80	**0**	**31**
	State wrapper	0	0	80	**0**	**31**
81	PIS schema control	1	1	76	18	32
82	PIS schema control	**1**	4	**75**	23	28

Table 33: **All 12 frozen trajectories.** Week suffixes 93 and 94 identify seeds 909093 and 909094. Hit means on-time execution; Late is delayed execution; Miss is unfulfilled opportunity; FA is false-alarm events. Hit plus Late plus Miss equals 77 or 82 in each row. Invalid calls remain included. Bold: best Hit, Miss and FA within each model/week; ties included.

Model	Interface	Week	Hit ↑	Late	Miss ↓	FA ↓	Invalid ↓
Gemma	Legacy four	93	1	0	76	**0**	0
	Explicit single	93	5	1	71	16	0
	Explicit four	93	**7**	0	**70**	16	0
Gemma	Legacy four	94	7	0	75	34	0
	Explicit single	94	7	0	75	**23**	0
	Explicit four	94	**16**	0	**66**	78	0
Llama	Legacy four	93	0	0	77	**0**	0
	Explicit single	93	**13**	3	**61**	107	17
	Explicit four	93	6	0	71	38	0
Llama	Legacy four	94	0	0	82	**0**	0
	Explicit single	94	**14**	1	**67**	65	0
	Explicit four	94	2	0	80	23	0

I.7 Eight-episode composition with execution feedback

Eight first-day episodes provide 92 observations per controller and 78 due opportunities. Each of six arms follows its own action history. A shared scheduler queries the clock and declared

task-associated state channels before every decision. Direct reads the visible transcript; Summary acquires task-aware prose; C0 uses the original extraction interface; C1 exposes the unchanged schema. C1 also supplies the guard and the acquired-state rule. Summary and compiled model decisions receive exactly 8,192 rendered input tokens. Invalid or oversized actions execute the empty set and remain counted.

Table 34 gives the compact comparison, and Table 35 preserves all outcome columns. Official false-alarm, commission and wrong-content diagnostics are distinct from extra task selections (Set FP). Figure 8 shows the interface and operating points. Paired episode-block intervals in Table 36 compare C1 with C0, C1 with Summary, and the guard increment. The prespecified joint criterion requires a nonnegative lower bound for the hit-rate change and a negative upper bound for false selections per step on both models.

Table 34: **Feedback execution.** Eight new episodes (92 observations) per controller, with its own history. Bold compares the four upper LM-input rows; lower rows add the guard or rule. Hit counts on-time executions; FP/step counts extra selections.

	Gemma		Llama	
Controller	**Hit/eligible ↑**	**FP/step ↓**	**Hit/eligible ↑**	**FP/step ↓**
Direct	**50/78**	**0.38**	**27/78**	1.07
Task-aware summary	47/78	0.55	19/78	0.74
Compiled: original schema	42/78	0.80	11/78	**0.47**
Compiled: visible schema	27/78	0.72	18/78	0.58
Visible schema + guard	27/78	0.71	18/78	0.58
Acquired-state rule	3/78	0.02	3/78	0.21

Table 35: **Complete execution outcomes.** FA, commission and wrong content are separate official scorer diagnostics; their sum must not be substituted for set FP. All 96 trajectories and invalid empty actions remain in the counts. Bold: within-model optima for the outcome columns, excluding Late and constant Comm.; ties included.

Model	**Controller**	**Hit**	**Late**	**Miss**	**FA**	**Comm.**	**Wrong**	**Set FP**	**Exact (%)**
Gemma	Direct	**50**	2	**26**	33	0	17	35	**48.91**
	Summary	47	2	29	49	0	35	51	36.96
	Compiled C0	42	2	34	72	0	43	74	31.52
	Compiled C1	27	3	48	63	0	36	66	22.83
	C1 + guard	27	3	48	62	0	36	65	23.91
	State rule	3	2	73	**0**	0	**1**	**2**	36.96
Llama	Direct	**27**	9	**42**	89	0	63	98	19.57
	Summary	19	6	53	62	0	49	68	26.09
	Compiled C0	11	1	66	42	0	**16**	43	30.43
	Compiled C1	18	4	56	49	0	43	53	28.26
	C1 + guard	18	4	56	49	0	43	53	28.26
	State rule	3	1	74	**18**	0	17	**19**	**34.78**

Table 37 substitutes a guard or rule for one current action on the factual C1 trajectory while holding prior actions fixed. Table 38 checks acquired-field compatibility on each arm's own due-task subset. Raw FIRE/OK and Executed/OK use the rule-compatible denominator; a valid returned decision can exceed the runtime action cap. These diagnostics distinguish local decision changes from complete alternative trajectories.

Table 39 sums logical acquisition and decision usage over all eight episodes per arm. Shared cached calls are charged to every consuming arm; returned token totals exclude unresolved usage from two interrupted Gemma requests, bounded by 9,856 additional input tokens and 8,192 reserved output tokens. Each arm also issues 255 state queries, including 92 clock queries. Gemma uses 741 distinct model requests with 916 logical uses; Llama uses 706 distinct requests with 915 logical uses. These totals keep pipeline accounting separate from the per-step execution rates.

Table 36: **Paired episode-block intervals for the three prespecified contrasts.** The primary contrast is visible versus original extraction schema. No execution-retention margin is introduced after seeing results.

Model	Contrast	Δ hit rate (pp), 95% CI	Δ FP/step, 95% CI
Gemma	Compiled C1 vs. Compiled C0	-19.23 [-37.50, 0.00]	-0.09 [-0.18, 0.05]
	Compiled C1 vs. Summary	-25.64 [-41.03, -8.97]	0.16 [-0.12, 0.50]
	C1 + guard vs. Compiled C1	0.00 [0.00, 0.00]	-0.01 [-0.03, 0.00]
Llama	Compiled C1 vs. Compiled C0	8.97 [0.00, 18.42]	0.11 [-0.27, 0.51]
	Compiled C1 vs. Summary	-1.28 [-8.86, 7.69]	-0.16 [-0.37, 0.09]
	C1 + guard vs. Compiled C1	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]

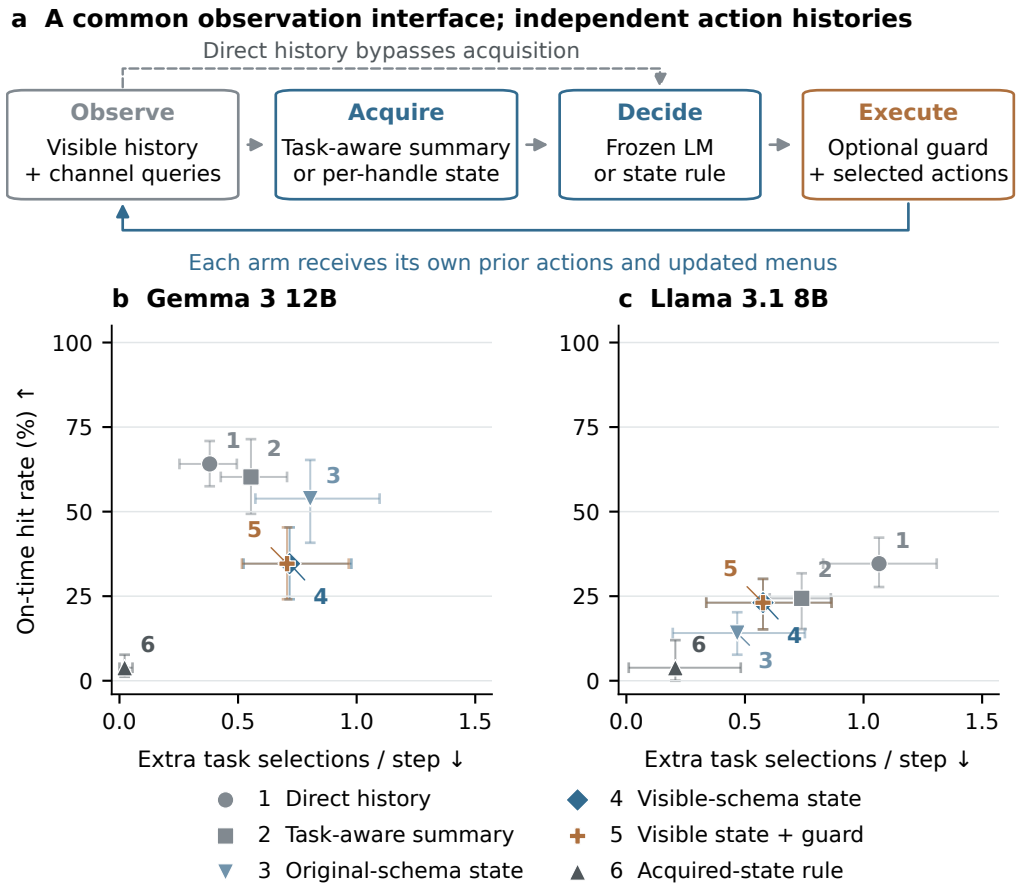


Figure 8: **Execution under a common observation policy.** Each arm has its own action history; the diagram shows the new multi-handle interface. Points use all eight episodes per model, with descriptive 95% episode-bootstrap intervals. Lower extra-task rate and higher on-time hit rate are preferred. The joint paired tests, rather than overlap of marginal intervals, determine the prespecified comparisons in Table 36.

Table 37: **Single-step substitutions on the exact visible-state trajectory.** Earlier actions are fixed. Changes count action sets; fixes and harms count exact-set outcomes. These are diagnostic local effects, not alternate complete trajectories.

Model	Substitute	Steps	Changes	Fixes	Harms	Δ TP	Δ FP
Gemma	Guard	91	1	1	0	0	-1
	Rule	91	38	13	8	-13	-25
Llama	Guard	92	0	0	0	0	0
	Rule	92	30	6	2	-12	-30

Table 38: **Post-hoc field compatibility on official due tasks along each arm's own trajectory.** Rule-compatible means active status, met trigger, authorization yes and same-request binding. Raw FIRE/OK and executed/OK use only compatible states; a valid model decision can still exceed the runtime action cap. This is a diagnostic subset, not a system score.

Model	Input	Due	No state	Rule OK	Raw FIRE/OK	Executed/ OK
Gemma	Compiled C0	78	1	19	18/19	18/19
	Compiled C1	78	1	14	14/14	14/14
Llama	Compiled C0	78	0	13	8/13	4/13
	Compiled C1	78	0	13	7/13	7/13

Table 39: **Logical pipeline cost including acquisition, summed over eight episodes per arm.** Shared cached calls are charged to each consuming arm. Invalid counts model responses; blocked counts empty execution fallbacks, including action-cap violations. These are known returned token totals; interrupted-request bounds accompany the table.

Model	Controller	Acq.	Dec.	Input tokens	Output tokens	Invalid	Blocked
Gemma	Direct	0	92	470,216	6,646	0	0
	Summary	92	91	1,123,588	31,660	1	1
	Compiled C0	92	91	1,125,906	78,669	1	1
	Compiled C1	92	91	1,274,040	73,550	1	1
	C1 + guard	92	91	1,273,986	73,557	1	1
	State rule	92	0	558,562	73,941	0	1
Llama	Direct	0	92	465,986	9,801	0	0
	Summary	92	87	1,067,748	42,718	5	5
	Compiled C0	92	92	1,116,793	84,013	0	3
	Compiled C1	92	92	1,258,736	86,615	0	5
	C1 + guard	92	92	1,258,736	86,615	0	5
	State rule	92	0	511,012	76,023	0	10

I.8 HISTORICAL COST SENSITIVITY AND EVIDENCE CROSSWALK

Table 40 is a hypothetical decision-loss sensitivity analysis on the earlier 512-prefix runs. A wrong decision on a FIRE reference costs λ, another wrong decision costs one and a correct decision costs zero. The table reports total-loss coefficients $a\lambda+b$; division by 512 gives mean loss. This is distinct from measured acquisition tokens. Table 41 preserves the study-to-record crosswalk. Editable table cells, the original vector figures, their available plot coordinates and the historical 512-prefix scoring rows are included under `historical/`; `historical_inventory.json` maps every original figure and table to its current location.

Table 40: **Coefficients of total loss $a\lambda + b$ on 512 prefixes; divide by 512 for mean loss.** The same matrix applies to all methods.

	Gemma		Llama	
Method	a	b	a	b
Direct	25	336	8	387
Summary	29	287	33	338
Compiled	83	148	87	133
StateMem-inspired	22	337	35	355
PIS-inspired	88	185	84	296
Acquired-state rule	88	166	88	332
Compiled + guard	83	155	87	133
Without binding	83	155	87	130
Without citation	83	155	87	133
Matched-input summary	32	242	43	283
Matched-input state	85	151	85	130
Matched-input state + guard	85	161	85	130
Direct + receipt guard*	25	304	8	355
Compiled + receipt guard*	83	137	87	129
Direct + acquired guard*	51	199	8	385

Table 41: **Study numbers used to locate unchanged historical records.**

Study	Archive identifier	Evaluation
1	P203	Padded task-set compilation
2	P183	Controlled API four-action outputs
3	P222	Grok task-selection follow-up
4	P185	Source-validity replay of API outputs
5	P211	Additional STAR local guard evaluation
6	P210	Three-prompt historical STAR challenge
7	P212	Native model-acquired lifecycle state
8	P214	Stronger extraction/recheck controls
9	P205	Earlier supplied-candidate prefix adaptation
10	P224	Reserved-prefix receipt-gate follow-up
11	P199	Separate 750-item human confirmation
12	P198	Sequential lifecycle adapter
13	P213	Executable sandbox

J RESTORED HISTORICAL DATA BY PROTOCOL

This appendix restores detailed tables reported in earlier manuscript versions: condition-specific outcomes, paired intervals, invalid outputs, acquired-evidence diagnostics and full costs. The saved measurements retain their original cohorts, interfaces and scorers. They complement the current results without entering the primary panel's aggregate scores. The source-to-table manifest is included with the editable manuscript; alternate earlier reference versions and transport snapshots remain identifiable in the separate version archive.

J.1 FIVE-PROMPT DEVELOPMENT SENSITIVITY

The five-prompt diagnostic uses 32 previously inspected author-constructed cases (eight per action); it is a development comparison.

Table 42: Study 19 results across five frozen prompts on the 32-case development panel.

Model	Prompt	Direct correct ↑	D1 correct ↑	Direct macro-F1 ↑	D1 macro-F1 ↑	Fixes / harms
Gemma3	Prompt 1	**31**	**31**	**0.9686**	**0.9686**	0 / 0
Gemma3	Prompt 2	**32**	**32**	**1.0000**	**1.0000**	0 / 0
Gemma3	Prompt 3	**32**	**32**	**1.0000**	**1.0000**	0 / 0
Gemma3	Prompt 4	**32**	**32**	**1.0000**	**1.0000**	0 / 0
Gemma3	Prompt 5	**32**	**32**	**1.0000**	**1.0000**	0 / 0
Llama3.1	Prompt 1	**23**	**23**	**0.6603**	**0.6603**	0 / 0
Llama3.1	Prompt 2	16	**20**	0.4095	**0.5476**	4 / 0
Llama3.1	Prompt 3	**25**	**25**	**0.7612**	**0.7612**	0 / 0
Llama3.1	Prompt 4	**21**	**21**	**0.6353**	**0.6353**	0 / 0
Llama3.1	Prompt 5	18	**19**	0.5549	**0.5821**	1 / 0

J.2 DISTINCT LIFECYCLE EXTENSIONS

The sequential controller study has 200 due opportunities. Timely execution, repeated execution, wrong execution and physical costs retain their original distinct units.

Table 43: Study 12 Gemma3 sequential outcomes, model calls, and token usage.

Gemma3 policy	Timely / 200	Repeat FIRE	Wrong FIRE	Calls	Tokens
Direct	199	377	384	1,226	623,056
D1 adapter	199	0	0	941	453,079
Source/ status rule	199	0	0	941	453,079
Receipt de-duplication	199	0	7	964	464,023
Additional reviewer	199	319	322	2,380	1,231,566
PIS-style adapter	135	40	52	982	582,454
State adapter	186	216	244	1,193	815,869
Three-action adapter	159	301	336	1,398	736,414
Two-action adapter	154	95	122	1,400	760,649

J.3 Exact Study 19 prompts and local weights

This crosswalk preserves the original protocol-specific meanings of actions; it does not retrospectively relabel a study.

Table 44: Protocol-specific action conventions for visible lifecycle situations.

Visible situation	Original Studies 2 and 24 convention	Studies 20 and 19 fixture convention
Active, known future trigger not met	DEFER	DEFER if a future-triggered operation was explicitly delegated
STAR approval not yet given	DEFER under the explicit confirmation trigger	Missing confirmation is VERIFY
Unknown or untrusted trigger evidence	VERIFY	VERIFY
Observed failed receipt	VERIFY in STAR before retry	VERIFY before retry
Matching successful completed occurrence	RETIRE in STAR	RETIRE if the logged receipt binds to this occurrence
Later distinct request or replacement (interface requirement)	Evaluate its own candidate and version	Evaluate its own anchored occurrence

J.4 Full STAR prompt challenge and dialogue-disjoint check

The original STAR prompt challenge uses 600 probes; the dialogue-disjoint follow-up uses 150 probes from 50 dialogues. Content parsing and strict-JSON scoring are separate reported endpoints. Paired intervals retain their original dialogue resampling and success/failure stratification.

Table 45: Study 6 terminal errors and paired D1 gains under three prompt formulations.

Model / prompt	All false FIRE, Direct / D1	Terminal survival, Direct / D1	Gain (pp)	Pooled dialogue 95% interval	Fixed-quota 95% interval
gemma3 / Original	100 / 0	100 / 0	16.667	[14.333, 19.000]	[16.667, 16.667]
gemma3 / Occurrence restatement	35 / 0	35 / 0	5.833	[4.167, 7.667]	[4.333, 7.333]
gemma3 / State-first restatement	0 / 0	0 / 0	0.000	[0.000, 0.000]	[0.000, 0.000]
llama31 / Original	174 / 74	100 / 0	16.667	[14.333, 19.000]	[16.667, 16.667]
llama31 / Occurrence restatement	130 / 44	86 / 0	14.333	[12.000, 16.667]	[13.167, 15.333]
llama31 / State-first restatement	36 / 25	11 / 0	1.833	[0.833, 3.000]	[0.833, 2.833]

Table 46: Study 6 physical model calls and recorded token usage by prompt.

Model / prompt	Physical calls ↓	Input tokens ↓	Output tokens ↓
gemma3 / Original	600	1,366,271	3,935
gemma3 / Occurrence restatement	600	1,438,871	3,890
gemma3 / State-first restatement	600	1,443,071	3,980
llama31 / Original	600	1,271,935	3,600
llama31 / Occurrence restatement	600	1,343,935	3,614
llama31 / State-first restatement	600	1,346,335	3,701

Table 47: Study 5 accuracy and action errors on additional STAR dialogue IDs.

Model / prompt	Direct ↑	D1 ↑	False FIRE, Direct / D1	Missed FIRE	Terminal survival, Direct / D1	Fixes / harms
gemma3 / Original	82.667	**99.333**	25 / 0	0	25 / 0	25 / 0
gemma3 / Occurrence	92.000	**98.000**	9 / 0	3	9 / 0	9 / 0
llama31 / Original	54.000	**70.667**	45 / 20	0	25 / 0	25 / 0
llama31 / Occurrence	53.333	**68.000**	31 / 11	19	22 / 0	22 / 0

Table 48: Study 5 paired gains and dialogue-resampling intervals.

Model / comparison	Gain (pp)	Pooled dialogue 95% interval	Fixed-quota 95% interval
gemma3/Original D1	16.667	[12.000, 21.333]	[16.667, 16.667]
gemma3/Occurrence D1	6.000	[2.667, 10.000]	[2.667, 9.333]
gemma3/Occurrence vs original Direct	9.333	[4.000, 14.667]	[5.333, 13.333]
llama31/Original D1	16.667	[12.000, 21.333]	[16.667, 16.667]
llama31/Occurrence D1	14.667	[10.000, 19.333]	[12.667, 16.667]
llama31/Occurrence vs original Direct	-0.667	[-8.000, 7.333]	[-8.000, 7.333]

Table 49: Strict-JSON uncertainty corresponding to the main local D1 table. All invalid responses are incorrect and remain unchanged by D1. Gains and intervals are percentage points. This is a post hoc descriptive reconstruction from archived responses, with 20,000 paired whole-dialogue bootstrap draws, 50 dialogue clusters, and seed 22320260908. It is not multiplicity-adjusted confirmatory inference.

Model	Prompt	Gain	Fix/ harm	Pooled dialogue 95% interval
Gemma	Original	16.000	24/0	[11.333, 20.667]
Gemma	Occurrence	6.000	9/0	[2.667, 10.000]
Llama	Original	16.667	25/0	[12.000, 21.333]
Llama	Occurrence	14.667	22/0	[10.000, 19.333]

Table 50: Study 5 physical model calls and recorded token usage by prompt.

Model / prompt	Physical calls ↓	Input tokens ↓	Output tokens ↓
gemma3 / Original	150	320,926	950
gemma3 / Occurrence	150	339,076	961
llama31 / Original	150	298,137	900
llama31 / Occurrence	150	316,137	903

J.5 New native information and interface matrix

Booking has 150 probes from 50 dialogues and four-action support; ride has 60 probes from 30 dialogues and VERIFY/RETIRE support. Orig. and State denote original and state-first prompts. Ext. is model-extracted state; Fence is the fence-tolerant parsing sensitivity. D and D1 are the decision before and after the closure guard; FF, MF and FR count false FIRE, missed FIRE and false RETIRE.

Table 51: Booking results across information views, prompts, and parsers.

Model	View	Prompt	Parser	D ↑	D1 ↑	FF D / D1	MF D / D1	FR D / D1	Inv
Gemma	Oracle	Orig.	Strict	63.33	**76.67**	20/0	15/15	0/0	33
Gemma	Oracle	Orig.	Fence	79.33	**94.67**	23/0	6/6	0/0	3
Gemma	Oracle	Occ.	Strict	70.67	**80.00**	14/0	17/17	0/0	23
Gemma	Oracle	Occ.	Fence	85.33	**94.67**	14/0	8/8	0/0	1
Gemma	Oracle	State	Strict	**77.33**	**77.33**	0/0	19/19	0/0	24
Gemma	Oracle	State	Fence	**93.33**	**93.33**	0/0	10/10	0/0	0
Gemma	raw	Orig.	Strict	51.33	–	18	26	0	45
Gemma	raw	Orig.	Fence	71.33	–	25	18	0	0
Gemma	raw	State	Strict	37.33	–	21	44	0	40
Gemma	raw	State	Fence	54.00	–	25	43	0	0
Gemma	Ext.	Orig.	Strict	**33.33**	**33.33**	9/9	49/49	0/0	46
Gemma	Ext.	Orig.	Fence	**49.33**	**49.33**	11/11	49/49	0/0	2
Gemma	Ext.	State	Strict	**36.67**	**36.67**	22/22	50/50	0/0	34
Gemma	Ext.	State	Fence	**48.67**	**48.67**	24/24	50/50	0/0	0
Llama	Oracle	Orig.	Strict	50.67	**67.33**	54/29	0/0	1/1	0
Llama	Oracle	Orig.	Fence	50.67	**67.33**	54/29	0/0	1/1	0
Llama	Oracle	Occ.	Strict	64.00	**73.33**	33/19	3/3	0/0	0
Llama	Oracle	Occ.	Fence	64.00	**73.33**	33/19	3/3	0/0	0
Llama	Oracle	State	Strict	**74.00**	**74.00**	2/2	25/25	3/3	0
Llama	Oracle	State	Fence	**74.00**	**74.00**	2/2	25/25	3/3	0
Llama	raw	Orig.	Strict	34.00	–	75	0	0	0
Llama	raw	Orig.	Fence	34.00	–	75	0	0	0
Llama	raw	State	Strict	56.67	–	24	37	0	0
Llama	raw	State	Fence	56.67	–	24	37	0	0
Llama	Ext.	Orig.	Strict	38.67	**51.33**	61/33	11/15	13/28	0
Llama	Ext.	Orig.	Fence	38.67	**51.33**	61/33	11/15	13/28	0
Llama	Ext.	State	Strict	**48.67**	44.00	9/7	42/42	20/29	0
Llama	Ext.	State	Fence	**48.67**	44.00	9/7	42/42	20/29	0

Table 52: Ride results across information views, prompts, and parsers.

Model	View	Prompt	Parser	D ↑	D1 ↑	FF D / D1	MF D / D1	FR D / D1	Inv
Gemma	Oracle	Orig.	Strict	**45.00**	**45.00**	15/15	0/0	0/0	6
Gemma	Oracle	Orig.	Fence	**48.33**	**48.33**	18/18	0/0	0/0	0
Gemma	Oracle	Occ.	Strict	**45.00**	**45.00**	25/25	0/0	0/0	5
Gemma	Oracle	Occ.	Fence	**50.00**	**50.00**	26/26	0/0	0/0	1
Gemma	Oracle	State	Strict	**43.33**	**43.33**	23/23	0/0	0/0	7

Model	View	Prompt	Parser	D ↑	D1 ↑	FF D / D1	MF D / D1	FR D / D1	Inv
Gemma	Oracle	State	Fence	**50.00**	**50.00**	23/23	0/0	0/0	0
Gemma	raw	Orig.	Strict	36.67	–	28	0	0	10
Gemma	raw	Orig.	Fence	43.33	–	34	0	0	0
Gemma	raw	State	Strict	46.67	–	22	0	0	7
Gemma	raw	State	Fence	50.00	–	23	0	0	0
Gemma	Ext.	Orig.	Strict	**35.00**	**35.00**	7/7	0/0	0/0	9
Gemma	Ext.	Orig.	Fence	**38.33**	**38.33**	9/9	0/0	0/0	1
Gemma	Ext.	State	Strict	**76.67**	**76.67**	4/4	0/0	0/0	9
Gemma	Ext.	State	Fence	**86.67**	**86.67**	4/4	0/0	0/0	1
Llama	Oracle	Orig.	Strict	23.33	**48.33**	38/23	0/0	0/0	0
Llama	Oracle	Orig.	Fence	23.33	**48.33**	38/23	0/0	0/0	0
Llama	Oracle	Occ.	Strict	31.67	**50.00**	41/30	0/0	0/0	0
Llama	Oracle	Occ.	Fence	31.67	**50.00**	41/30	0/0	0/0	0
Llama	Oracle	State	Strict	**81.67**	**81.67**	0/0	0/0	10/10	0
Llama	Oracle	State	Fence	**81.67**	**81.67**	0/0	0/0	10/10	0
Llama	raw	Orig.	Strict	35.00	–	39	0	0	0
Llama	raw	Orig.	Fence	35.00	–	39	0	0	0
Llama	raw	State	Strict	93.33	–	2	0	2	0
Llama	raw	State	Fence	93.33	–	2	0	2	0
Llama	Ext.	Orig.	Strict	15.00	**36.67**	33/18	0/0	6/8	0
Llama	Ext.	Orig.	Fence	15.00	**36.67**	33/18	0/0	6/8	0
Llama	Ext.	State	Strict	**68.33**	**68.33**	0/0	0/0	18/18	0
Llama	Ext.	State	Fence	**68.33**	**68.33**	0/0	0/0	18/18	0

Table 53: Extractor acceptance, terminal certificates, and rule-only accuracy.

Model	Panel	View parser	n	Valid	Cert	False cert	Rule correct
Llama	Booking	Strict	150	150	50	26	28
Llama	Booking	Fence	150	150	50	26	28
Llama	Ride	Strict	60	60	23	8	52
Llama	Ride	Fence	60	60	23	8	52
Gemma	Booking	Strict	150	0	0	0	25
Gemma	Booking	Fence	150	150	12	0	37
Gemma	Ride	Strict	60	0	0	0	45
Gemma	Ride	Fence	60	60	12	0	57

Table 54: Full-rule and prefilter controls under supplied and extracted state.

Model	Panel	State	Prompt	Arm	Parser	Acc. ↑	FR	Inv
Llama	Booking	Oracle	none	Full	Rule	100.00	0	0
Llama	Booking	Ext.	none	Full	Strict	18.67	26	0
Llama	Booking	Ext.	none	Full	Fence	18.67	26	0
Llama	Ride	Oracle	none	Full	Rule	100.00	0	0
Llama	Ride	Ext.	none	Full	Strict	86.67	8	0
Llama	Ride	Ext.	none	Full	Fence	86.67	8	0
Llama	Booking	Oracle	Orig.	Pre	Strict	67.33	1	0
Llama	Booking	Oracle	Orig.	Pre	Fence	67.33	1	0
Llama	Booking	Oracle	Occ.	Pre	Strict	73.33	0	0
Llama	Booking	Oracle	Occ.	Pre	Fence	73.33	0	0
Llama	Booking	Oracle	State	Pre	Strict	74.00	3	0
Llama	Booking	Oracle	State	Pre	Fence	74.00	3	0
Llama	Ride	Oracle	Orig.	Pre	Strict	48.33	0	0
Llama	Ride	Oracle	Orig.	Pre	Fence	48.33	0	0
Llama	Ride	Oracle	Occ.	Pre	Strict	50.00	0	0
Llama	Ride	Oracle	Occ.	Pre	Fence	50.00	0	0
Llama	Ride	Oracle	State	Pre	Strict	81.67	10	0

Model	Panel	State	Prompt	Arm	Parser	Acc. ↑	FR	Inv
Llama	Ride	Oracle	State	Pre	Fence	81.67	10	0
Llama	Booking	Ext.	State	Pre	Strict	44.00	29	0
Llama	Booking	Ext.	State	Pre	Fence	44.00	29	0
Llama	Booking	Ext.	Orig.	Pre	Strict	51.33	28	0
Llama	Booking	Ext.	Orig.	Pre	Fence	51.33	28	0
Llama	Ride	Ext.	Orig.	Pre	Strict	36.67	8	0
Llama	Ride	Ext.	Orig.	Pre	Fence	36.67	8	0
Llama	Ride	Ext.	State	Pre	Strict	68.33	18	0
Llama	Ride	Ext.	State	Pre	Fence	68.33	18	0
Gemma	Booking	Oracle	none	Full	Rule	100.00	0	0
Gemma	Booking	Ext.	none	Full	Strict	16.67	0	0
Gemma	Booking	Ext.	none	Full	Fence	24.67	0	0
Gemma	Ride	Oracle	none	Full	Rule	100.00	0	0
Gemma	Ride	Ext.	none	Full	Strict	75.00	0	0
Gemma	Ride	Ext.	none	Full	Fence	95.00	0	0
Gemma	Booking	Oracle	Occ.	Pre	Strict	81.33	0	21
Gemma	Booking	Oracle	Occ.	Pre	Fence	94.67	0	1
Gemma	Booking	Oracle	Orig.	Pre	Strict	79.33	0	29
Gemma	Booking	Oracle	Orig.	Pre	Fence	95.33	0	2
Gemma	Booking	Oracle	State	Pre	Strict	78.00	0	23
Gemma	Booking	Oracle	State	Pre	Fence	93.33	0	0
Gemma	Ride	Oracle	Orig.	Pre	Strict	46.67	0	5
Gemma	Ride	Oracle	Orig.	Pre	Fence	48.33	0	0
Gemma	Ride	Oracle	Occ.	Pre	Strict	48.33	0	3
Gemma	Ride	Oracle	Occ.	Pre	Fence	50.00	0	1
Gemma	Ride	Oracle	State	Pre	Strict	46.67	0	5
Gemma	Ride	Oracle	State	Pre	Fence	50.00	0	0
Gemma	Booking	Ext.	Orig.	Pre	Strict	33.33	0	46
Gemma	Booking	Ext.	Orig.	Pre	Fence	49.33	0	2
Gemma	Booking	Ext.	State	Pre	Strict	36.67	0	34
Gemma	Booking	Ext.	State	Pre	Fence	48.67	0	0
Gemma	Ride	Ext.	State	Pre	Strict	76.67	0	9
Gemma	Ride	Ext.	State	Pre	Fence	86.67	0	1
Gemma	Ride	Ext.	Orig.	Pre	Strict	35.00	0	9
Gemma	Ride	Ext.	Orig.	Pre	Fence	38.33	0	1

J.6 Prompt replication and observable staged outcomes

Paired prompt effects are state-first minus original on the same dialogues. The two interval columns retain pooled-dialogue and fixed-quota resampling. Staged extraction/decision counts describe observed joint outcomes.

Table 55: Paired prompt effects and corresponding action-error changes.

Model	Panel	State	Arm	Parser	Delta pp	Dialogue CI	Quota CI	FF-term / MF / FR
Llama	Booking	Oracle	direct	Strict	23.33	[15.33, 31.33]	[15.33, 31.33]	-25/25/2
Llama	Booking	Oracle	D1	Strict	6.67	[-2.67, 16.00]	[-1.33, 14.67]	0/25/2
Llama	Booking	Oracle	direct	Fence	23.33	[15.33, 31.33]	[15.33, 31.33]	-25/25/2
Llama	Booking	Oracle	D1	Fence	6.67	[-2.67, 16.00]	[-1.33, 14.67]	0/25/2
Llama	Booking	raw	direct	Strict	22.67	[16.00, 28.67]	[17.33, 28.00]	-11/37/0
Llama	Booking	raw	direct	Fence	22.67	[16.00, 28.67]	[17.33, 28.00]	-11/37/0
Llama	Booking	Ext.	direct	Strict	10.00	[2.67, 17.33]	[2.67, 17.33]	-23/31/7
Llama	Booking	Ext.	D1	Strict	-7.33	[-13.33, -1.33]	[-13.33, -1.33]	-1/27/1
Llama	Booking	Ext.	direct	Fence	10.00	[2.67, 17.33]	[2.67, 17.33]	-23/31/7
Llama	Booking	Ext.	D1	Fence	-7.33	[-13.33, -1.33]	[-13.33, -1.33]	-1/27/1
Llama	Ride	Oracle	direct	Strict	58.33	[46.67, 70.00]	[48.33, 68.33]	-15/0/10
Llama	Ride	Oracle	D1	Strict	33.33	[23.33, 43.33]	[23.33, 43.33]	0/0/10
Llama	Ride	Oracle	direct	Fence	58.33	[46.67, 70.00]	[48.33, 68.33]	-15/0/10

Model	Panel	State	Arm	Parser	Delta pp	Dialogue CI	Quota CI	FF-term / MF / FR
Llama	Ride	Oracle	D1	Fence	33.33	[23.33, 43.33]	[23.33, 43.33]	0/0/10
Llama	Ride	raw	direct	Strict	58.33	[46.67, 70.00]	[46.67, 68.33]	-11/0/2
Llama	Ride	raw	direct	Fence	58.33	[46.67, 70.00]	[46.67, 68.33]	-11/0/2
Llama	Ride	Ext.	direct	Strict	53.33	[40.00, 66.67]	[43.33, 63.33]	-14/0/12
Llama	Ride	Ext.	D1	Strict	31.67	[23.33, 40.00]	[23.33, 40.00]	0/0/10
Llama	Ride	Ext.	direct	Fence	53.33	[40.00, 66.67]	[43.33, 63.33]	-14/0/12
Llama	Ride	Ext.	D1	Fence	31.67	[23.33, 40.00]	[23.33, 40.00]	0/0/10
Gemma	Booking	Oracle	direct	Strict	14.00	[6.67, 21.33]	[7.33, 20.67]	-20/4/0
Gemma	Booking	Oracle	D1	Strict	0.67	[-6.00, 7.33]	[-5.33, 7.33]	0/4/0
Gemma	Booking	Oracle	direct	Fence	14.00	[8.00, 20.00]	[9.33, 18.67]	-23/4/0
Gemma	Booking	Oracle	D1	Fence	-1.33	[-6.00, 3.33]	[-6.00, 3.33]	0/4/0
Gemma	Booking	raw	direct	Strict	-14.00	[-20.67, -7.33]	[-20.67, -7.33]	3/18/0
Gemma	Booking	raw	direct	Fence	-17.33	[-22.00, -12.67]	[-22.00, -12.67]	0/25/0
Gemma	Booking	Ext.	direct	Strict	3.33	[-2.00, 8.67]	[-2.00, 8.67]	13/1/0
Gemma	Booking	Ext.	D1	Strict	3.33	[-2.00, 8.67]	[-2.00, 8.67]	13/1/0
Gemma	Booking	Ext.	direct	Fence	-0.67	[-3.33, 2.00]	[-3.33, 2.00]	13/1/0
Gemma	Booking	Ext.	D1	Fence	-0.67	[-3.33, 2.00]	[-3.33, 2.00]	13/1/0
Gemma	Ride	Oracle	direct	Strict	-1.67	[-5.00, 0.00]	[-5.00, 0.00]	0/0/0
Gemma	Ride	Oracle	D1	Strict	-1.67	[-5.00, 0.00]	[-5.00, 0.00]	0/0/0
Gemma	Ride	Oracle	direct	Fence	1.67	[0.00, 5.00]	[0.00, 5.00]	0/0/0
Gemma	Ride	Oracle	D1	Fence	1.67	[0.00, 5.00]	[0.00, 5.00]	0/0/0
Gemma	Ride	raw	direct	Strict	10.00	[3.33, 18.33]	[3.33, 16.67]	-3/0/0
Gemma	Ride	raw	direct	Fence	6.67	[1.67, 13.33]	[1.67, 11.67]	-4/0/0
Gemma	Ride	Ext.	direct	Strict	41.67	[33.33, 48.33]	[33.33, 48.33]	-1/0/0
Gemma	Ride	Ext.	D1	Strict	41.67	[33.33, 48.33]	[33.33, 48.33]	-1/0/0
Gemma	Ride	Ext.	direct	Fence	48.33	[41.67, 55.00]	[41.67, 55.00]	-2/0/0
Gemma	Ride	Ext.	D1	Fence	48.33	[41.67, 55.00]	[41.67, 55.00]	-2/0/0

Table 56: Joint outcomes of state extraction and subsequent model decisions.

Model	Panel	Prompt	Parser	RC / MC	RC / MW	RW / follow	RW / MC	RW / other
Llama	Booking	Orig.	Strict	1	27	11	57	54
Llama	Booking	Orig.	Fence	1	27	11	57	54
Llama	Booking	State	Strict	26	2	37	47	38
Llama	Booking	State	Fence	26	2	37	47	38
Llama	Ride	Orig.	Strict	8	44	6	1	1
Llama	Ride	Orig.	Fence	8	44	6	1	1
Llama	Ride	State	Strict	41	11	8	0	0
Llama	Ride	State	Fence	41	11	8	0	0
Gemma	Booking	Orig.	Strict	13	12	0	37	88
Gemma	Booking	Orig.	Fence	23	2	0	51	74
Gemma	Booking	State	Strict	19	6	1	36	88
Gemma	Booking	State	Fence	25	0	3	48	74
Gemma	Ride	Orig.	Strict	13	32	0	8	7
Gemma	Ride	Orig.	Fence	15	30	0	8	7
Gemma	Ride	State	Strict	38	7	0	8	7
Gemma	Ride	State	Fence	42	3	0	10	5

Table 57: Complete paired D1 effects across information views, prompts, and parsers.

Model	Panel	State	Prompt	Parser	Gain pp	Dialogue CI	Quota CI	Fix / harm / W-to-W
Llama	Booking	Oracle	Orig.	Strict	16.67	[12.00, 21.33]	[16.67, 16.67]	25/0/0
Llama	Booking	Oracle	Orig.	Fence	16.67	[12.00, 21.33]	[16.67, 16.67]	25/0/0

Model	Panel	State	Prompt	Parser	Gain pp	Dialogue CI	Quota CI	Fix / harm / W-to-W
Llama	Booking	Oracle	Occ.	Strict	9.33	[5.33, 13.33]	[6.00, 12.67]	14/0/0
Llama	Booking	Oracle	Occ.	Fence	9.33	[5.33, 13.33]	[6.00, 12.67]	14/0/0
Llama	Booking	Oracle	State	Strict	0.00	[0.00, 0.00]	[0.00, 0.00]	0/0/0
Llama	Booking	Oracle	State	Fence	0.00	[0.00, 0.00]	[0.00, 0.00]	0/0/0
Llama	Booking	Ext.	Orig.	Strict	12.67	[7.33, 18.00]	[9.33, 15.33]	24/5/10
Llama	Booking	Ext.	Orig.	Fence	12.67	[7.33, 18.00]	[9.33, 15.33]	24/5/10
Llama	Booking	Ext.	State	Strict	-4.67	[-8.67, -0.67]	[-8.00, -1.33]	2/9/0
Llama	Booking	Ext.	State	Fence	-4.67	[-8.67, -0.67]	[-8.00, -1.33]	2/9/0
Llama	Ride	Oracle	Orig.	Strict	25.00	[16.67, 33.33]	[25.00, 25.00]	15/0/0
Llama	Ride	Oracle	Orig.	Fence	25.00	[16.67, 33.33]	[25.00, 25.00]	15/0/0
Llama	Ride	Oracle	Occ.	Strict	18.33	[10.00, 26.67]	[13.33, 23.33]	11/0/0
Llama	Ride	Oracle	Occ.	Fence	18.33	[10.00, 26.67]	[13.33, 23.33]	11/0/0
Llama	Ride	Oracle	State	Strict	0.00	[0.00, 0.00]	[0.00, 0.00]	0/0/0
Llama	Ride	Oracle	State	Fence	0.00	[0.00, 0.00]	[0.00, 0.00]	0/0/0
Llama	Ride	Ext.	Orig.	Strict	21.67	[11.67, 31.67]	[16.67, 25.00]	14/1/1
Llama	Ride	Ext.	Orig.	Fence	21.67	[11.67, 31.67]	[16.67, 25.00]	14/1/1
Llama	Ride	Ext.	State	Strict	0.00	[0.00, 0.00]	[0.00, 0.00]	0/0/0
Llama	Ride	Ext.	State	Fence	0.00	[0.00, 0.00]	[0.00, 0.00]	0/0/0
Gemma	Booking	Oracle	Orig.	Strict	13.33	[8.67, 18.00]	[10.67, 16.00]	20/0/0
Gemma	Booking	Oracle	Orig.	Fence	15.33	[10.67, 20.00]	[13.33, 16.67]	23/0/0
Gemma	Booking	Oracle	Occ.	Strict	9.33	[5.33, 13.33]	[6.00, 12.67]	14/0/0
Gemma	Booking	Oracle	Occ.	Fence	9.33	[5.33, 13.33]	[6.00, 12.67]	14/0/0
Gemma	Booking	Oracle	State	Strict	0.00	[0.00, 0.00]	[0.00, 0.00]	0/0/0
Gemma	Booking	Oracle	State	Fence	0.00	[0.00, 0.00]	[0.00, 0.00]	0/0/0
Gemma	Booking	Ext.	Orig.	Strict	0.00	[0.00, 0.00]	[0.00, 0.00]	0/0/0
Gemma	Booking	Ext.	Orig.	Fence	0.00	[0.00, 0.00]	[0.00, 0.00]	0/0/0
Gemma	Booking	Ext.	State	Strict	0.00	[0.00, 0.00]	[0.00, 0.00]	0/0/0
Gemma	Booking	Ext.	State	Fence	0.00	[0.00, 0.00]	[0.00, 0.00]	0/0/0
Gemma	Ride	Oracle	Orig.	Strict	0.00	[0.00, 0.00]	[0.00, 0.00]	0/0/0
Gemma	Ride	Oracle	Orig.	Fence	0.00	[0.00, 0.00]	[0.00, 0.00]	0/0/0
Gemma	Ride	Oracle	Occ.	Strict	0.00	[0.00, 0.00]	[0.00, 0.00]	0/0/0
Gemma	Ride	Oracle	Occ.	Fence	0.00	[0.00, 0.00]	[0.00, 0.00]	0/0/0
Gemma	Ride	Oracle	State	Strict	0.00	[0.00, 0.00]	[0.00, 0.00]	0/0/0
Gemma	Ride	Oracle	State	Fence	0.00	[0.00, 0.00]	[0.00, 0.00]	0/0/0
Gemma	Ride	Ext.	Orig.	Strict	0.00	[0.00, 0.00]	[0.00, 0.00]	0/0/0
Gemma	Ride	Ext.	Orig.	Fence	0.00	[0.00, 0.00]	[0.00, 0.00]	0/0/0
Gemma	Ride	Ext.	State	Strict	0.00	[0.00, 0.00]	[0.00, 0.00]	0/0/0
Gemma	Ride	Ext.	State	Fence	0.00	[0.00, 0.00]	[0.00, 0.00]	0/0/0

J.7 Complete minimal-contrast results

Each minimal-contrast row uses 16 parameter families. FPE/FPL compare future and past FIRE rates at early/late display; LPP/LPF compare late and early display at past/future date. A-E/A-L compare RETIRE for new and old requests. J-E/J-L are the proportions jointly retiring the old occurrence and firing the new one. Joint rows are proportions, not differences.

Table 58: Booking minimal-contrast estimates and confidence intervals.

Model	Prompt	Parser	Contrast	Estimate pp	95% CI	Base / Treated n
Llama	Orig.	Strict	FPE	0.00	[0.00, 0.00]	0/0
Llama	Orig.	Strict	FPL	0.00	[0.00, 0.00]	0/0
Llama	Orig.	Strict	LPP	0.00	[0.00, 0.00]	0/0
Llama	Orig.	Strict	LPF	0.00	[0.00, 0.00]	0/0
Llama	Orig.	Strict	A-E	-25.00	[-50.00, -6.25]	4/0
Llama	Orig.	Strict	A-L	-75.00	[-93.75, -50.00]	12/0
Llama	Orig.	Strict	J-E	0.00	–	–/0

Model	Prompt	Parser	Contrast	Estimate pp	95% CI	Base / Treated n
Llama	Orig.	Strict	J-L	0.00	–	–/0
Llama	Orig.	Fence	FPE	0.00	[0.00, 0.00]	0/0
Llama	Orig.	Fence	FPL	0.00	[0.00, 0.00]	0/0
Llama	Orig.	Fence	LPP	0.00	[0.00, 0.00]	0/0
Llama	Orig.	Fence	LPF	0.00	[0.00, 0.00]	0/0
Llama	Orig.	Fence	A-E	-25.00	[-50.00, -6.25]	4/0
Llama	Orig.	Fence	A-L	-75.00	[-93.75, -50.00]	12/0
Llama	Orig.	Fence	J-E	0.00	–	–/0
Llama	Orig.	Fence	J-L	0.00	–	–/0
Llama	State	Strict	FPE	0.00	[0.00, 0.00]	0/0
Llama	State	Strict	FPL	0.00	[0.00, 0.00]	0/0
Llama	State	Strict	LPP	0.00	[0.00, 0.00]	0/0
Llama	State	Strict	LPF	0.00	[0.00, 0.00]	0/0
Llama	State	Strict	A-E	0.00	[0.00, 0.00]	0/0
Llama	State	Strict	A-L	-25.00	[-50.00, -6.25]	4/0
Llama	State	Strict	J-E	0.00	–	–/0
Llama	State	Strict	J-L	0.00	–	–/0
Llama	State	Fence	FPE	0.00	[0.00, 0.00]	0/0
Llama	State	Fence	FPL	0.00	[0.00, 0.00]	0/0
Llama	State	Fence	LPP	0.00	[0.00, 0.00]	0/0
Llama	State	Fence	LPF	0.00	[0.00, 0.00]	0/0
Llama	State	Fence	A-E	0.00	[0.00, 0.00]	0/0
Llama	State	Fence	A-L	-25.00	[-50.00, -6.25]	4/0
Llama	State	Fence	J-E	0.00	–	–/0
Llama	State	Fence	J-L	0.00	–	–/0
Gemma	Orig.	Strict	FPE	0.00	[0.00, 0.00]	0/0
Gemma	Orig.	Strict	FPL	0.00	[0.00, 0.00]	0/0
Gemma	Orig.	Strict	LPP	0.00	[0.00, 0.00]	0/0
Gemma	Orig.	Strict	LPF	0.00	[0.00, 0.00]	0/0
Gemma	Orig.	Strict	A-E	-100.00	[-100.00, -100.00]	16/0
Gemma	Orig.	Strict	A-L	-100.00	[-100.00, -100.00]	16/0
Gemma	Orig.	Strict	J-E	100.00	–	–/16
Gemma	Orig.	Strict	J-L	100.00	–	–/16
Gemma	Orig.	Fence	FPE	0.00	[0.00, 0.00]	0/0
Gemma	Orig.	Fence	FPL	0.00	[0.00, 0.00]	0/0
Gemma	Orig.	Fence	LPP	0.00	[0.00, 0.00]	0/0
Gemma	Orig.	Fence	LPF	0.00	[0.00, 0.00]	0/0
Gemma	Orig.	Fence	A-E	-100.00	[-100.00, -100.00]	16/0
Gemma	Orig.	Fence	A-L	-100.00	[-100.00, -100.00]	16/0
Gemma	Orig.	Fence	J-E	100.00	–	–/16
Gemma	Orig.	Fence	J-L	100.00	–	–/16
Gemma	State	Strict	FPE	0.00	[0.00, 0.00]	0/0
Gemma	State	Strict	FPL	0.00	[0.00, 0.00]	0/0
Gemma	State	Strict	LPP	0.00	[0.00, 0.00]	0/0
Gemma	State	Strict	LPF	0.00	[0.00, 0.00]	0/0
Gemma	State	Strict	A-E	-100.00	[-100.00, -100.00]	16/0
Gemma	State	Strict	A-L	-100.00	[-100.00, -100.00]	16/0
Gemma	State	Strict	J-E	100.00	–	–/16
Gemma	State	Strict	J-L	100.00	–	–/16
Gemma	State	Fence	FPE	0.00	[0.00, 0.00]	0/0
Gemma	State	Fence	FPL	0.00	[0.00, 0.00]	0/0
Gemma	State	Fence	LPP	0.00	[0.00, 0.00]	0/0
Gemma	State	Fence	LPF	0.00	[0.00, 0.00]	0/0
Gemma	State	Fence	A-E	-100.00	[-100.00, -100.00]	16/0
Gemma	State	Fence	A-L	-100.00	[-100.00, -100.00]	16/0
Gemma	State	Fence	J-E	100.00	–	–/16
Gemma	State	Fence	J-L	100.00	–	–/16

Table 59: Ride minimal-contrast estimates and confidence intervals.

Model	Prompt	Parser	Contrast	Estimate pp	95% CI	Base / Treated n
Llama	Orig.	Strict	FPE	0.00	[0.00, 0.00]	0/0
Llama	Orig.	Strict	FPL	0.00	[0.00, 0.00]	0/0
Llama	Orig.	Strict	LPP	0.00	[0.00, 0.00]	0/0
Llama	Orig.	Strict	LPF	0.00	[0.00, 0.00]	0/0
Llama	Orig.	Strict	A-E	-43.75	[-68.75, -18.75]	7/0
Llama	Orig.	Strict	A-L	-93.75	[-100.00, -81.25]	16/1
Llama	Orig.	Strict	J-E	0.00	–	–/0
Llama	Orig.	Strict	J-L	0.00	–	–/0
Llama	Orig.	Fence	FPE	0.00	[0.00, 0.00]	0/0
Llama	Orig.	Fence	FPL	0.00	[0.00, 0.00]	0/0
Llama	Orig.	Fence	LPP	0.00	[0.00, 0.00]	0/0
Llama	Orig.	Fence	LPF	0.00	[0.00, 0.00]	0/0
Llama	Orig.	Fence	A-E	-43.75	[-68.75, -18.75]	7/0
Llama	Orig.	Fence	A-L	-93.75	[-100.00, -81.25]	16/1
Llama	Orig.	Fence	J-E	0.00	–	–/0
Llama	Orig.	Fence	J-L	0.00	–	–/0
Llama	State	Strict	FPE	0.00	[0.00, 0.00]	0/0
Llama	State	Strict	FPL	0.00	[0.00, 0.00]	0/0
Llama	State	Strict	LPP	0.00	[0.00, 0.00]	0/0
Llama	State	Strict	LPF	0.00	[0.00, 0.00]	0/0
Llama	State	Strict	A-E	-18.75	[-37.50, 0.00]	3/0
Llama	State	Strict	A-L	-100.00	[-100.00, -100.00]	16/0
Llama	State	Strict	J-E	0.00	–	–/0
Llama	State	Strict	J-L	0.00	–	–/0
Llama	State	Fence	FPE	0.00	[0.00, 0.00]	0/0
Llama	State	Fence	FPL	0.00	[0.00, 0.00]	0/0
Llama	State	Fence	LPP	0.00	[0.00, 0.00]	0/0
Llama	State	Fence	LPF	0.00	[0.00, 0.00]	0/0
Llama	State	Fence	A-E	-18.75	[-37.50, 0.00]	3/0
Llama	State	Fence	A-L	-100.00	[-100.00, -100.00]	16/0
Llama	State	Fence	J-E	0.00	–	–/0
Llama	State	Fence	J-L	0.00	–	–/0
Gemma	Orig.	Strict	FPE	0.00	[0.00, 0.00]	0/0
Gemma	Orig.	Strict	FPL	0.00	[0.00, 0.00]	0/0
Gemma	Orig.	Strict	LPP	0.00	[0.00, 0.00]	0/0
Gemma	Orig.	Strict	LPF	0.00	[0.00, 0.00]	0/0
Gemma	Orig.	Strict	A-E	-100.00	[-100.00, -100.00]	16/0
Gemma	Orig.	Strict	A-L	-100.00	[-100.00, -100.00]	16/0
Gemma	Orig.	Strict	J-E	93.75	–	–/15
Gemma	Orig.	Strict	J-L	100.00	–	–/16
Gemma	Orig.	Fence	FPE	0.00	[0.00, 0.00]	0/0
Gemma	Orig.	Fence	FPL	0.00	[0.00, 0.00]	0/0
Gemma	Orig.	Fence	LPP	0.00	[0.00, 0.00]	0/0
Gemma	Orig.	Fence	LPF	0.00	[0.00, 0.00]	0/0
Gemma	Orig.	Fence	A-E	-100.00	[-100.00, -100.00]	16/0
Gemma	Orig.	Fence	A-L	-100.00	[-100.00, -100.00]	16/0
Gemma	Orig.	Fence	J-E	93.75	–	–/15
Gemma	Orig.	Fence	J-L	100.00	–	–/16
Gemma	State	Strict	FPE	0.00	[0.00, 0.00]	0/0
Gemma	State	Strict	FPL	0.00	[0.00, 0.00]	0/0
Gemma	State	Strict	LPP	0.00	[0.00, 0.00]	0/0
Gemma	State	Strict	LPF	0.00	[0.00, 0.00]	0/0
Gemma	State	Strict	A-E	-100.00	[-100.00, -100.00]	16/0
Gemma	State	Strict	A-L	-100.00	[-100.00, -100.00]	16/0
Gemma	State	Strict	J-E	0.00	–	–/0
Gemma	State	Strict	J-L	100.00	–	–/16
Gemma	State	Fence	FPE	0.00	[0.00, 0.00]	0/0
Gemma	State	Fence	FPL	0.00	[0.00, 0.00]	0/0
Gemma	State	Fence	LPP	0.00	[0.00, 0.00]	0/0
Gemma	State	Fence	LPF	0.00	[0.00, 0.00]	0/0

Model	Prompt	Parser	Contrast	Estimate pp	95% CI	Base / Treated n
Gemma	State	Fence	A-E	-100.00	[-100.00, -100.00]	16/0
Gemma	State	Fence	A-L	-100.00	[-100.00, -100.00]	16/0
Gemma	State	Fence	J-E	0.00	–	–/0
Gemma	State	Fence	J-L	100.00	–	–/16

Table 60: New-instance execution, known-waiting, and unresolved-state controls.

Model	Panel	Prompt	Parser	New E	New L	Known	Unresolved
Gemma	Booking	Orig.	Strict	16/16	16/16	16/16	16/16
Gemma	Booking	Orig.	Fence	16/16	16/16	16/16	16/16
Gemma	Booking	State	Strict	16/16	16/16	16/16	16/16
Gemma	Booking	State	Fence	16/16	16/16	16/16	16/16
Llama	Booking	Orig.	Strict	0/16	0/16	16/16	0/16
Llama	Booking	Orig.	Fence	0/16	0/16	16/16	0/16
Llama	Booking	State	Strict	0/16	0/16	16/16	0/16
Llama	Booking	State	Fence	0/16	0/16	16/16	0/16
Gemma	Ride	Orig.	Strict	15/16	16/16	16/16	16/16
Gemma	Ride	Orig.	Fence	15/16	16/16	16/16	16/16
Gemma	Ride	State	Strict	0/16	16/16	16/16	16/16
Gemma	Ride	State	Fence	0/16	16/16	16/16	16/16
Llama	Ride	Orig.	Strict	0/16	0/16	16/16	1/16
Llama	Ride	Orig.	Fence	0/16	0/16	16/16	1/16
Llama	Ride	State	Strict	0/16	0/16	16/16	9/16
Llama	Ride	State	Fence	0/16	0/16	16/16	9/16

J.8 Executable sandbox methods and complete outcomes

The executable sandbox evaluates complete controllers whose actions affect later observations. G/L denote Gemma/Llama; Raw and Ext+R denote raw history and extraction followed by a rule.

Table 61: Sandbox event schedules, required instances, and intended distinctions.

Schedule, in each workflow	Fixed external events	Required instances at the horizon	Intended distinction
Delayed receipt	Authorized request at tick 0; receipt is delivered three ticks after an effect; deadline 2	1	Submission versus an observed outcome; a query can reveal an existing effect before its delayed receipt
Cancellation before confirmation	Unconfirmed request at tick 0, cancellation at tick 1	0	Retention while a known condition is unmet, followed by instance-specific closure
Cancellation and a later new authorization	Unconfirmed old request at tick 0, cancellation at tick 1, distinct authorized request with the same parameters at tick 3; new deadline 5	1	A cancelled old occurrence does not cancel a separately authorized new one
Later request with identical parameters	Authorized request at tick 0, deadline 2; a distinct authorized request with the same parameters at tick 4, deadline 6	2	Identical parameters do not merge operation occurrences
Parameter revision	Unconfirmed version 1 at tick 0; supersession of version 1 and an authorized changed-parameter version 2 at tick 2, deadline 5	1	Closure of the superseded version versus execution of its replacement
Future trigger and explicit reminder	Authorized request at tick 0 with trigger 3 and deadline 4; explicit re-nomination at tick 5	1	A future condition does not imply terminality; a reminder can reopen an incorrectly closed unfinished row

Table 62: Complete sandbox outcomes by model and controller.

G / L	Policy	Req.	Time	Final	Lost H	Closed	Recov.	Wrong	Query	Calls
L	Raw	12	2	2	10	0	0	0	0	88
L	Ext+R	12	0	0	12	0	0	0	96	96
G	Raw	12	8	8	4	2	0	0	0	62
G	Ext+R	12	0	0	12	0	0	0	96	96

J.9 Physical costs, hypothetical placement savings and integrity

Physical requests count actual generations; logical pipeline costs charge dependencies to every consuming arm. Hypothetical placement savings are not measured latency or realized new runs.

Table 63: Physical calls, token usage, runtime, and parser failures by phase.

Model	Phase	Calls	Input tokens ↓	Output tokens ↓	Phase wall s	Strict invalid	Fence invalid
Llama	Initial	1562	2267877	9605	579.76	0	0
Llama	extract	210	344867	11548	134.21	0	0
Llama	Child	420	834814	2616	202.12	0	0
Gemma	Initial	1562	2451495	10731	1385.92	200	5
Gemma	extract	210	357065	11859	291.30	210	0
Gemma	Child	420	817122	3018	419.01	98	4

Table 64: Logical pipeline costs and hypothetical calls avoided by prefilter placement.

Model	Panel	Info	Prompt	Arm	Calls	Input tokens ↓	Output tokens ↓	Skip calls
Llama	Booking	Oracle	Orig.	direct	150	299070	901	0
Llama	Booking	Oracle	Orig.	D1	150	299070	901	0
Llama	Booking	Oracle	Orig.	Pre	125	249856	751	25
Llama	Booking	Oracle	none	Full	0	0	0	150
Llama	Booking	Oracle	State	direct	150	317670	928	0
Llama	Booking	Oracle	State	D1	150	317670	928	0
Llama	Booking	Oracle	State	Pre	125	265356	753	25
Llama	Booking	raw	Orig.	direct	150	251670	900	0
Llama	Booking	raw	State	direct	150	270270	909	0
Llama	Booking	Ext.	Orig.	direct	300	585060	9139	0
Llama	Booking	Ext.	Orig.	D1	300	585060	9139	0
Llama	Booking	Ext.	Orig.	Pre	250	473076	8828	50
Llama	Booking	Ext.	none	Full	150	269520	8226	150
Llama	Booking	Ext.	State	direct	300	603660	9169	0
Llama	Booking	Ext.	State	D1	300	603660	9169	0
Llama	Booking	Ext.	State	Pre	250	485476	8830	50
Llama	Ride	Oracle	Orig.	direct	60	87167	360	0
Llama	Ride	Oracle	Orig.	D1	60	87167	360	0
Llama	Ride	Oracle	Orig.	Pre	45	65623	270	15
Llama	Ride	Oracle	none	Full	0	0	0	60
Llama	Ride	Oracle	State	direct	60	92747	385	0
Llama	Ride	Oracle	State	D1	60	92747	385	0
Llama	Ride	Oracle	State	Pre	45	69808	280	15
Llama	Ride	raw	Orig.	direct	60	68207	364	0
Llama	Ride	raw	State	direct	60	73787	377	0
Llama	Ride	Ext.	Orig.	direct	120	165124	3689	0
Llama	Ride	Ext.	Orig.	D1	120	165124	3689	0
Llama	Ride	Ext.	Orig.	Pre	97	133536	3544	23
Llama	Ride	Ext.	none	Full	60	75347	3322	60
Llama	Ride	Ext.	State	direct	120	170704	3715	0
Llama	Ride	Ext.	State	D1	120	170704	3715	0
Llama	Ride	Ext.	State	Pre	97	136977	3554	23
Gemma	Booking	Oracle	Orig.	direct	150	323219	1061	0
Gemma	Booking	Oracle	Orig.	D1	150	323219	1061	0
Gemma	Booking	Oracle	Orig.	Pre	125	270007	895	25
Gemma	Booking	Oracle	none	Full	0	0	0	150
Gemma	Booking	Oracle	State	direct	150	342419	1045	0
Gemma	Booking	Oracle	State	D1	150	342419	1045	0
Gemma	Booking	Oracle	State	Pre	125	286007	865	25
Gemma	Booking	raw	Orig.	direct	150	262269	1125	0
Gemma	Booking	raw	State	direct	150	281469	1100	0
Gemma	Booking	Ext.	Orig.	direct	300	586578	9724	0
Gemma	Booking	Ext.	Orig.	D1	300	586578	9724	0
Gemma	Booking	Ext.	Orig.	Pre	300	586578	9724	0
Gemma	Booking	Ext.	none	Full	150	279369	8594	150
Gemma	Booking	Ext.	State	direct	300	605778	9664	0

Model	Panel	Info	Prompt	Arm	Calls	Input tokens ↓	Output tokens ↓	Skip calls
Gemma	Booking	Ext.	State	D1	300	605778	9664	0
Gemma	Booking	Ext.	State	Pre	300	605778	9664	0
Gemma	Ride	Oracle	Orig.	direct	60	95142	405	0
Gemma	Ride	Oracle	Orig.	D1	60	95142	405	0
Gemma	Ride	Oracle	Orig.	Pre	45	71613	295	15
Gemma	Ride	Oracle	none	Full	0	0	0	60
Gemma	Ride	Oracle	State	direct	60	101022	410	0
Gemma	Ride	Oracle	State	D1	60	101022	410	0
Gemma	Ride	Oracle	State	Pre	45	76023	295	15
Gemma	Ride	raw	Orig.	direct	60	70856	421	0
Gemma	Ride	raw	State	direct	60	76736	410	0
Gemma	Ride	Ext.	Orig.	direct	120	166508	3673	0
Gemma	Ride	Ext.	Orig.	D1	120	166508	3673	0
Gemma	Ride	Ext.	Orig.	Pre	120	166508	3673	0
Gemma	Ride	Ext.	none	Full	60	77696	3265	60
Gemma	Ride	Ext.	State	direct	120	172388	3675	0
Gemma	Ride	Ext.	State	D1	120	172388	3675	0
Gemma	Ride	Ext.	State	Pre	120	172388	3675	0

J.10 Independent reconstruction, interval sensitivity and concrete errors

False-certificate examples are mechanically selected from saved extraction outputs and compared with evidence visible at the stated cutoff.

Table 65: Mechanically selected false-certificate examples and their visible evidence.

Workflow / reference	False certificates in the stratum	Mechanically selected case	Cited visible evidence and error
Booking / DEFER	1	`P212N-04a3241b21d3b266dfd5`	At cutoff 76, the extractor cites event 73, a `restaurant_book` Check query, and asserts completed/submitted. That query does not establish a successful booking.
Booking / FIRE	4	`P212N-2897031274f1c23c5e96`	At cutoff 50, the extractor cites event 47, an Available Check result, and asserts completed/submitted. The visible workflow requires the still-unsubmitted, confirmed booking to FIRE.
Booking / VERIFY	21	`P212N-1e4a5a19fd5bbb640b92`	At cutoff 52, it cites an Available result at 47 and a Book query at 51, while asserting completion. The selected booking has a failed outcome requiring VERIFY.
Ride modification / VERIFY	8	`P212N-08b8de6cc2ac6b7cff50`	At cutoff 5, it cites only the issued `ride_change` query at 4 and asserts completion. The native reference requires VERIFY.

J.11 Frozen stronger-extraction repair on historical probes

The stronger-extraction comparison reuses previously evaluated booking and ride probes (150 and 60). A3 uses extraction/rechecking; DV adds direct-decision review. These are development repairs, and structural acceptance is separate from semantic correctness.

Table 66: Study 8 extraction and direct-decision accuracy with per-pipeline call counts.

pipeline	calls	Gemma booking	Llama booking	Gemma ride	Llama ride
A0 Original free	1	16.7	20.0	75.0	90.0
A1 Original schema	1	22.7	22.0	96.7	98.3
A2 Evidence schema	1	26.7	22.7	100.0	85.0
A3 Extract-verify	2	29.3	28.7	100.0	100.0
D0 Direct schema	1	54.0	56.0	50.0	93.3
DV Direct-review	2	55.3	33.3	51.7	93.3

Table 67: Study 8 format acceptance, semantic-action accuracy, and certificate diagnostics.

Model	Panel	Arm	N	Acc ↑	Valid correct / all	For-mat fail	Accept fail	False cert/ all cert
gemma3	booking	A0	150	16.7	0.0	150	150	0/0
gemma3	booking	A1	150	22.7	22.7	0	0	0/9
gemma3	booking	A2	150	26.7	26.7	0	0	0/15
gemma3	booking	A3	150	29.3	29.3	0	0	0/19
gemma3	booking	D0	150	54.0	54.0	0	0	0/0
gemma3	booking	DV	150	55.3	55.3	0	0	0/0
gemma3	ride	A0	60	75.0	0.0	60	60	0/0
gemma3	ride	A1	60	96.7	96.7	0	0	0/13
gemma3	ride	A2	60	100.0	100.0	0	0	0/15
gemma3	ride	A3	60	100.0	100.0	0	0	0/15
gemma3	ride	D0	60	50.0	50.0	0	0	0/0
gemma3	ride	DV	60	51.7	51.7	0	0	0/0
llama31	booking	A0	150	20.0	20.0	0	0	21/43
llama31	booking	A1	150	22.0	22.0	0	0	27/52
llama31	booking	A2	150	22.7	22.7	0	0	19/43
llama31	booking	A3	150	28.7	28.7	0	0	3/23
llama31	booking	D0	150	56.0	56.0	0	0	0/0
llama31	booking	DV	150	33.3	33.3	0	0	0/0
llama31	ride	A0	60	90.0	90.0	0	0	6/21
llama31	ride	A1	60	98.3	98.3	0	0	1/16
llama31	ride	A2	60	85.0	85.0	0	0	9/24
llama31	ride	A3	60	100.0	98.3	0	1	0/15
llama31	ride	D0	60	93.3	93.3	0	0	0/0
llama31	ride	DV	60	93.3	93.3	0	0	0/0

Table 68: Study 8 declared and additional descriptive paired comparisons.

Model	Panel	Change	Gain pp	Dialogue interval	Fix/ harm
llama31	booking	A0 to A1	2.0	[-2.0, 6.7]	7/4
llama31	booking	A1 to A2	0.7	[-3.3, 4.7]	5/4
llama31	booking	A2 to A3	6.0	[0.7, 11.3]	13/4
llama31	booking	DV to A3	-4.7	[-12.0, 2.7]	19/26
llama31	booking	D0 to DV	-22.7	[-30.7, -15.3]	7/41
llama31	booking	D0 to A2	-33.3	[-42.7, -24.0]	17/67
llama31	ride	A0 to A1	8.3	[0.0, 16.7]	6/1
llama31	ride	A1 to A2	-13.3	[-23.3, -5.0]	1/9
llama31	ride	A2 to A3	15.0	[6.7, 23.3]	9/0
llama31	ride	DV to A3	6.7	[1.7, 13.3]	4/0
llama31	ride	D0 to DV	0.0	[-10.0, 8.3]	4/4
llama31	ride	D0 to A2	-8.3	[-18.3, 1.7]	3/8
gemma3	booking	A0 to A1	6.0	[2.7, 10.0]	9/0
gemma3	booking	A1 to A2	4.0	[1.3, 7.3]	6/0

Model	Panel	Change	Gain pp	Dialogue interval	Fix/ harm
gemma3	booking	A2 to A3	2.7	[0.7, 5.3]	4/0
gemma3	booking	DV to A3	-26.0	[-31.3, -20.7]	16/55
gemma3	booking	D0 to DV	1.3	[-2.7, 5.3]	6/4
gemma3	booking	D0 to A2	-27.3	[-32.7, -22.0]	15/56
gemma3	ride	A0 to A1	21.7	[13.3, 30.0]	13/0
gemma3	ride	A1 to A2	3.3	[0.0, 8.3]	2/0
gemma3	ride	A2 to A3	0.0	[0.0, 0.0]	0/0
gemma3	ride	DV to A3	48.3	[45.0, 50.0]	29/0
gemma3	ride	D0 to DV	1.7	[0.0, 5.0]	1/0
gemma3	ride	D0 to A2	50.0	[50.0, 50.0]	30/0

Table 69: Study 8 retirement-only correction using each acquired state view.

Model	Panel	State	Calls	Gain pp	Interval	Fix/ harm	Both wrong
llama31	booking	A0	2	-2.7	[-10.0, 4.7]	16/20	1
llama31	booking	A1	2	-2.7	[-10.0, 4.7]	18/22	5
llama31	booking	A2	2	0.7	[-6.7, 8.0]	17/16	3
llama31	booking	A3	3	7.3	[2.7, 12.7]	14/3	0
llama31	ride	A0	2	-10.0	[-18.3, -3.3]	0/6	0
llama31	ride	A1	2	-1.7	[-5.0, 0.0]	0/1	0
llama31	ride	A2	2	-13.3	[-21.7, -6.7]	0/8	0
llama31	ride	A3	3	0.0	[0.0, 0.0]	0/0	0
gemma3	booking	A0	2	0.0	[0.0, 0.0]	0/0	0
gemma3	booking	A1	2	6.0	[2.7, 10.0]	9/0	0
gemma3	booking	A2	2	10.0	[6.0, 14.0]	15/0	0
gemma3	booking	A3	3	12.0	[8.0, 16.7]	18/0	0
gemma3	ride	A0	2	0.0	[0.0, 0.0]	0/0	0
gemma3	ride	A1	2	0.0	[0.0, 0.0]	0/0	0
gemma3	ride	A2	2	0.0	[0.0, 0.0]	0/0	0
gemma3	ride	A3	3	0.0	[0.0, 0.0]	0/0	0

Table 70: Study 8 logical pipeline calls and recorded token usage.

Model	Panel	Arm	Calls	Input tokens ↓	Output tokens ↓
gemma3	booking	A0	150	279369	8578
gemma3	booking	A1	150	279369	8555
gemma3	booking	A2	150	307419	8468
gemma3	booking	A3	300	640894	17008
gemma3	booking	D0	150	281469	1136
gemma3	booking	DV	300	579739	2539
gemma3	ride	A0	60	77696	3263
gemma3	ride	A1	60	77696	3289
gemma3	ride	A2	60	88916	3199
gemma3	ride	A3	120	188076	6403
gemma3	ride	D0	60	76736	415
gemma3	ride	DV	120	160207	940
llama31	booking	A0	150	269520	8074
llama31	booking	A1	150	269520	8203
llama31	booking	A2	150	296820	7280
llama31	booking	A3	300	617672	14316
llama31	booking	D0	150	270270	907
llama31	booking	DV	300	555997	1808
llama31	ride	A0	60	75347	3198
llama31	ride	A1	60	75347	3223
llama31	ride	A2	60	86267	2763
llama31	ride	A3	120	181981	5522

Model	Panel	Arm	Calls	Input tokens ↓	Output tokens ↓
llama31	ride	D0	60	73787	376
llama31	ride	DV	120	153770	751

Table 71: Prespecified strict-interface minimal contrasts (16 families per row). Date: future minus past FIRE rate, early/late display (pp). Position: late minus early FIRE rate, past/future date (pp). Joint: two counts of families correctly retiring the old occurrence and firing the new occurrence, early then late display (each out of 16). Full paired intervals and both parsers are in Appendix J.7.

Panel	Model	Prompt	Date E/ L	Position P/ F	Joint E, L
Booking	Gemma	Original	0.0/0.0	0.0/0.0	16, 16
Booking	Gemma	State-first	0.0/0.0	0.0/0.0	16, 16
Booking	Llama	Original	0.0/0.0	0.0/0.0	0, 0
Booking	Llama	State-first	0.0/0.0	0.0/0.0	0, 0
Ride	Gemma	Original	0.0/0.0	0.0/0.0	15, 16
Ride	Gemma	State-first	0.0/0.0	0.0/0.0	0, 16
Ride	Llama	Original	0.0/0.0	0.0/0.0	0, 0
Ride	Llama	State-first	0.0/0.0	0.0/0.0	0, 0

J.12 Historical external API configuration and scope map

The four requested API configurations share 800 constructed cases. Correct counts retain the original operational fallback convention and source-validity-gated correction.

Table 72: Study 2 requested API configurations and paired source-valid D1 results.

Study 2 requested configuration	Direct correct / 800 ↑	D1 correct / 800 ↑	Paired gain (pp)	Source-invalid rows ↓
gpt-4o-mini	668	**705**	4.625	0
claude-haiku-4-5-20251001	769	**793**	3.000	10
deepseek-v4-flash	**800**	**800**	0.000	0
gemini-3.6-flash	**788**	**788**	0.000	16

J.13 Completed Study 1 cross-model compaction results

The padded-input study contains 206 action-bearing observations from four generated weeks. The 824 scored records are the original two-configuration, two-condition campaign; the separate incomplete eleven-row repeat is not substituted.

Table 73: Study 1 exact-set correctness by generated week and API configuration.

Generated week	Observations	DeepSeek full correct ↑	DeepSeek D1-v2 correct ↑	Gemini full correct ↑	Gemini D1-v2 correct ↑
101	51	**50**	**50**	49	**50**
102	50	42	**49**	**50**	**50**
103	51	44	**50**	50	**51**
104	54	**51**	50	52	**54**
All four weeks	206	187	**199**	201	**205**

Table 74: Study 1 token usage, task-selection counts, and cancellation diagnostics.

Model / arm	Total input tokens ↓	Total completion tokens ↓	TP / FP / FN	Recorded stale	Cancelled-selection rows
DeepSeek / full	25,247,201	435,588	284 / 9 / 24	0	0
DeepSeek / D1-v2	277,716	184,539	301 / 1 / 7	0	1
Gemini / full	27,055,627	180,441	306 / 4 / 2	0	0
Gemini / D1-v2	269,708	78,267	308 / 1 / 0	0	1

J.14 Native-dialogue table and protocol detail

The native booking and ride panels retain their own workflow targets and information interfaces. Supplied-state rules and learned extraction have different inputs and call costs.

Table 75: New native accuracy (%). Booking has 150 probes in 50 dialogues; ride modification has 60 probes in 30 dialogues and only VERIFY/RETIRE support. Cells with a slash show Direct/D1. All entries use strict parsing and primary extracted views. The oracle full rule is 100% in each panel by construction; it has no learned extraction cost.

Panel	Model	Prompt	Oracle D/ D1	Raw D	Extract D/ D1	Extract rule
Booking	Gemma	Original	63.3/76.7	51.3	33.3/33.3	16.7
Booking	Gemma	State-first	77.3/77.3	37.3	36.7/36.7	16.7
Booking	Llama	Original	50.7/67.3	34.0	38.7/51.3	18.7
Booking	Llama	State-first	74.0/74.0	56.7	48.7/44.0	18.7
Ride	Gemma	Original	45.0/45.0	36.7	35.0/35.0	75.0
Ride	Gemma	State-first	43.3/43.3	46.7	76.7/76.7	75.0
Ride	Llama	Original	23.3/48.3	35.0	15.0/36.7	86.7
Ride	Llama	State-first	81.7/81.7	93.3	68.3/68.3	86.7

J.15 Alternative-interface development diagnostic

The alternative-input development comparison uses all 305 steps of four already inspected generated weeks, including 99 no-action observations. Requested endpoint aliases describe archived configurations.

Table 76: Studies 26 and 27 exact-set correctness across alternative input interfaces.

Model	Timeline correct ↑	Ledger correct ↑	Bound correct ↑	Event-filter correct ↑	Direct-review correct ↑
Llama-3.1 local	3/305	9/305	38/305	7/305	38/305
GPT-4.1-nano requested API alias	98/305	99/305	100/305	Not in API protocol	Not in API protocol

Table 77: Study 27 complete Llama pipeline token usage and logical calls.

Llama pipeline	Complete pipeline input tokens	Complete pipeline output tokens	Logical calls ↓
Timeline	534,843	15,572	305
Ledger	413,236	13,563	305
Bound	419,509	11,547	305
Event filter	369,915	11,049	305
Direct review	1,114,233	22,554	610

J.16 Received human judgments and adjudication boundary

This is the separate historical 750-item review, with 727 eligible items and 23 ineligible items. The two reference columns describe that study and are not the 512-prefix primary panel.

Table 78: Post-hoc Study 9 rescore on 727 action-eligible cases from the contributor-attested 750-case human-reviewed return. The 23 INELIGIBLE cases remain separately reported. Old-reference and reviewed-reference columns use the same 727 IDs and unchanged predictions; percentages are accuracy. Correct counts and Macro-F1 use the reviewed reference. Invalid outputs remain incorrect; all displayed arms have zero invalids. Rules are shared zero-model-call predictions, not independent repetitions per model.

Model / strategy	Old ref. (%)	Reviewed correct ↑	Reviewed (%)	Macro-F1 ↑
Gemma / A0 full prefix	45.80	307/727	42.23	0.3661
Gemma / A3 compact+bound	44.70	263/727	36.18	0.3206
Gemma / A6 retirement	44.70	263/727	36.18	0.3206
Gemma / A9 review	44.84	264/727	36.31	0.3214
Llama / A0 full prefix	42.09	247/727	33.98	0.2650
Llama / A3 compact+bound	37.96	232/727	31.91	0.2301
Llama / A6 retirement	37.96	232/727	31.91	0.2301
Llama / A9 review	33.84	212/727	29.16	0.1907
A8 visible-evidence rule	72.63	347/727	47.73	0.4513
A8S extended FSM	67.54	309/727	42.50	0.4194

J.17 Complete condition, error, and cost discussion

Content extraction and strict single-object protocol accuracy are different scoring endpoints on the same 600 historical STAR probes.

Table 79: Gemma plus D1 on the same 600 historical STAR probes (Study 6). Content extraction and strict single-object protocol accuracy are different endpoints; all strict failures remain incorrect.

Prompt	Historical content (%)	Strict protocol (%)
Original	99.000	88.000
Occurrence restatement	98.333	91.167
State-first restatement	98.000	88.833

J.18 Completed frozen Grok follow-up

The completed Grok follow-up contains 2,359 unique planned scientific requests: 2,135 development calls, 192 additional-week calls and 32 padding calls. H/R/C/L/D/N/V and the zero-model-call controls are defined below. Invalid or failed calls remain in the denominator; no score from the incomplete transport snapshot is substituted. These are campaign aggregates across the initial endpoint and recovery, whose membership is specified below.

Table 80: Completed frozen follow-up conditions. R and C use D's per-case UTF-8 knowledge-byte ceiling; this is not equal provider token usage. Review includes its original H call in logical cost.

Arm	Decision input / controller	Comparison purpose
H	Original relevant JSON timeline; empty padding	Unpadded baseline
R	BM25-selected visible history units (Robertson & Zaragoza, 2009) within D's byte ceiling	Retrieval control
C	Task fields and compressed chronological event text within that ceiling	Compression control
L	Original v0 versioned ledger without explicit matches	Binding ablation
D	Original v0 D1-v2 ledger and current matches	Frozen method
N	D with updates removed from ledger reconstruction	Update ablation
V	Full H information plus its first response; one additional review	Extra-call control
R0	Return original v0 matched task IDs directly	Same-ledger rule
R1	Independent event rule retaining prior cancellations	Complete field rule

Table 81: Completed Grok exact-set results. Invalid includes HTTP, transport and strict-format failures and remains wrong. Logical calls count the shared H parent for V; unique physical requests for the entire study are 2,359.

Panel	Arm	Correct / N ↑	Exact (%) ↑	Invalid ↓	Calls / case
A. Method comparisons on 305 development observations					
Old305	H	249/305	81.64	5	1
Old305	R	236/305	77.38	2	1
Old305	C	206/305	67.54	5	1
Old305	D	282/305	**92.46**	3	1
Old305	V	233/305	76.39	16	2
B. Component ablations on the same development observations					
Old305	D	282/305	**92.46**	3	1
Old305	L	238/305	78.03	0	1
Old305	N	212/305	69.51	7	1
C. Additional-week confirmation and padding stress checks					
New96	H	77/96	80.21	0	1
New96	D	88/96	**91.67**	0	1
Padding32	H	24/32	75.00	1	1

Panel	Arm	Correct / N ↑	Exact (%) ↑	Invalid ↓	Calls / case
Padding32	P	22/32	68.75	0	1
Padding32	D	30/32	**93.75**	0	1

Table 82: Prespecified paired comparisons. Bootstrap resamples whole weeks (20,000 draws, seed 20260907222), preserving all within-week observations. Four-week diagnostic intervals are descriptive; the 12-week confirmation remains from the same generator. Multiple diagnostic intervals are unadjusted.

Panel	Contrast	Δ pp	Fix / harm	95% interval	Weeks
Old305	H→D	+10.82	49/16	[7.84, 13.82]	4
Old305	R→D	+15.08	59/13	[11.07, 17.76]	4
Old305	C→D	+24.92	87/11	[16.39, 30.07]	4
Old305	L→D	+14.43	55/11	[11.11, 19.67]	4
Old305	N→D	+22.95	76/6	[18.95, 27.54]	4
Old305	V→D	+16.07	66/17	[13.03, 19.41]	4
New96	H→D	+11.46	15/4	[4.17, 18.75]	12
Padding32	H→P	-6.25	6/8	[-28.12, 9.38]	4
Padding32	H→D	+18.75	7/1	[12.50, 31.25]	4
Padding32	P→D	+25.00	9/1	[12.50, 40.62]	4

The unpadded New96 comparison records 304,590 versus 226,879 input tokens and 151,264 versus 118,657 output tokens for H versus D; all 192 requests have usage. Mean request times are 26.57 versus 18.35 seconds within this campaign. Across the full 2,359-request study, 38 outputs are invalid; usage is missing for 32 failed requests. Known input/output subtotals are 9,699,981/4,108,568 tokens, so full campaign token totals remain unknown. Qwen had no scientific rows after failed preflights and withdrawal.

The archived v0 cancellation behavior remains visible: New96 D selects three cancelled task instances. On the 32-observation padding subset, H, padded H and D are correct on 24, 22 and 30 cases; the padding H-to-P interval includes zero. The task-set gains and token reductions retain these protocol-specific comparisons.

Table 83: Zero-model-call supplied-field controls. R0 preserves the archived v0 cancellation semantics; R1 retains cancellation until an explicit reactivation. Rule scores do not retroactively repair model outputs.

Panel	Rule	Correct / N ↑	Exact (%) ↑
old305	R0	302/305	99.02
old305	R1	305/305	**100.00**
new96	R0	93/96	96.88
new96	R1	96/96	**100.00**

Endpoint membership and recovery. The initial Grok calls use `4sapi.com`; continuation calls use `4sapi.org`, with the same `grok-4.6` model alias. The saved configuration does not establish endpoint equivalence. In New96, H has 93 initial-endpoint and three recovery-endpoint final rows; D has 95 and one. The paired composition is 92 initial/initial, three recovery/initial and one initial/recovery. In Old305, the initial/recovery counts are H: 298/7, R: 297/8, C: 297/8, D: 297/8, L: 299/6 and N: 298/7. All 305 physical V review calls use the recovery endpoint, consuming H parent outputs. Consequently, V versus H changes both the review operation and endpoint. Reported full-campaign gains retain these memberships rather than treating the endpoints as interchangeable.

Same-endpoint sensitivity. We recompute exact-set accuracy on pairs whose selected final responses both use the initial endpoint (Table 84). This is a post-hoc stratification of saved rows, not new inference. The 92 matched New96 pairs yield 76/92 correct for H and 84/92 for D: 12 fixes

and four harms. Input-token totals are 294,961 and 217,348, or means of 3,206.10 and 2,362.48 (26.31% lower). The four discordant pairs contribute H: 1/4 and D: 4/4; their three fixes do not isolate a method effect. Both components sum to the full 77/96 versus 88/96 result. Old305 comparisons use arm-specific endpoint-matched subsets; their denominators differ. V has no initial/initial paired subset.

Table 84: **Grok sensitivity on pairs with the same initial endpoint.** Both responses use `4sapi.com`. Correct counts retain invalid outputs as wrong. Gain is comparison minus H; bold marks the higher accuracy within each paired row. Different Old305 rows have different eligible case sets.

Panel	Comparison	Paired n	H correct	Arm correct	Gain (pp)	Fix / harm
New96	Compiled D	92	76	**84**	+8.70	12 / 4
Old305	Retrieval R	290	**236**	226	−3.45	32 / 42
	Compact text C	290	**238**	202	−12.41	27 / 63
	Compiled D	291	238	**270**	+11.00	46 / 14
	No binding L	292	**237**	227	−3.42	33 / 43
	No updates N	291	**236**	205	−10.65	28 / 59

J.19 Frozen native receipt-gate follow-up

The frozen receipt-gate follow-up has 54 reserved prefixes, of which 19 have determinate author-workflow targets, all DEFER. The remaining 35 are excluded from this accuracy denominator, not reclassified as successes.

Table 85: Frozen native follow-up. Scores use only the 19 determinate author-workflow references, all DEFER, with the same cases for both models. These are not four-class accuracy estimates. The 35 undetermined references are retained outside this accuracy denominator.

Policy	Gemma correct / 19 ↑	Llama correct / 19 ↑	Calls / case
D0	**4**	**9**	1
Old A3+D1	**4**	**9**	3
Proof-gated A3+D1	**4**	**9**	3
Visible proof+D0	**4**	**9**	1
A3 full state rule	0	0	2

J.20 Completed additional corpus and native evaluation

These are the earlier two-model runs on the 512 human-evaluated prefixes. Their four-action supports are 88/107/90/227; source-specific subsets each contain 256 prefixes. They remain distinct from the current three-configuration runs in the main text.

Table 86: Independent human judgments; rows are A, columns B. Total 512; diagonal 412; Cohen kappa 0.717.

A reference / B reference	FIRE	DEFER	VERIFY	RETIRE
FIRE	80	15	13	3
DEFER	4	93	31	0
VERIFY	8	11	21	4
RETIRE	7	4	0	218

Table 87: Gemma on all 512 prefixes under the completed final reference. Invalid outputs remain incorrect; macro-F1 uses the pooled four-class confusion matrix.

Condition	Accuracy (%) ↑	Macro-F1 ↑	FIRE recall (%) ↑	Invalid ↓
A. Representation and controller comparisons				
Direct	29.49	0.277	71.59	0
Summary	38.28	0.377	67.05	1
Compiled	**54.88**	**0.423**	5.68	0
StateMem-inspired	29.88	0.285	**75.00**	1
PIS-inspired	46.68	0.347	0.00	0
Acquired-state rule	50.39	0.326	0.00	0
Compiled+guard	53.52	0.403	5.68	0
B. Guard component ablations (reference: Compiled+guard)				
No binding check	53.52	0.403	5.68	0
No citation check	53.52	0.404	5.68	0
C. Equal decision-input token comparisons				
Equal-input summary	46.48	**0.413**	**63.64**	1
Equal-input state	**53.91**	0.391	3.41	1
Equal-input state+guard	51.95	0.369	3.41	1

Table 88: Llama on all 512 prefixes under the completed final reference. Invalid outputs remain incorrect; macro-F1 uses the pooled four-class confusion matrix.

Condition	Accuracy (%) ↑	Macro-F1 ↑	FIRE recall (%) ↑	Invalid ↓
A. Representation and controller comparisons				
Direct	22.85	0.177	**90.91**	0
Summary	27.54	0.228	62.50	0
Compiled	**57.03**	**0.352**	1.14	0
StateMem-inspired	23.83	0.183	60.23	0
PIS-inspired	25.78	0.210	4.55	0
Acquired-state rule	17.97	0.079	0.00	0
Compiled+guard	**57.03**	**0.352**	1.14	0
B. Guard component ablations (reference: Compiled+guard)				
No binding check	57.62	0.353	1.14	0
No citation check	57.03	0.352	1.14	0
C. Equal decision-input token comparisons				
Equal-input summary	36.33	0.299	**51.14**	0
Equal-input state	**58.01**	**0.365**	3.41	0
Equal-input state+guard	**58.01**	**0.365**	3.41	0

Table 89: Gemma on 256 SGD prefixes under the completed final reference. Invalid outputs remain incorrect; macro-F1 uses the pooled four-class confusion matrix.

Condition	Accuracy (%) ↑	Macro-F1 ↑	FIRE recall (%) ↑	Invalid ↓
A. Representation and controller comparisons				
Direct	26.95	0.236	60.34	0
Summary	41.80	**0.388**	63.79	1
Compiled	**49.22**	0.355	6.90	0
StateMem-inspired	27.73	0.238	**67.24**	1
PIS-inspired	46.88	0.321	0.00	0
Acquired-state rule	42.97	0.238	0.00	0
Compiled+guard	47.66	0.323	6.90	0
B. Guard component ablations (reference: Compiled+guard)				

Condition	Accuracy (%) ↑	Macro-F1 ↑	FIRE recall (%) ↑	Invalid ↓
No binding check	47.66	0.323	6.90	0
No citation check	47.66	0.323	6.90	0
C. Equal decision-input token comparisons				
Equal-input summary	**57.03**	**0.468**	**55.17**	1
Equal-input state	49.61	0.336	3.45	1
Equal-input state+guard	47.27	0.308	3.45	1

Table 90: Llama on 256 SGD prefixes under the completed final reference. Invalid outputs remain incorrect; macro-F1 uses the pooled four-class confusion matrix.

Condition	Accuracy (%) ↑	Macro-F1 ↑	FIRE recall (%) ↑	Invalid ↓
A. Representation and controller comparisons				
Direct	27.73	0.201	**91.38**	0
Summary	29.69	0.238	62.07	0
Compiled	**63.28**	**0.353**	0.00	0
StateMem-inspired	26.95	0.197	60.34	0
PIS-inspired	19.14	0.160	3.45	0
Acquired-state rule	8.59	0.044	0.00	0
Compiled+guard	**63.28**	**0.353**	0.00	0
B. Guard component ablations (reference: Compiled+guard)				
No binding check	64.45	0.358	0.00	0
No citation check	63.28	0.353	0.00	0
C. Equal decision-input token comparisons				
Equal-input summary	35.16	0.302	**58.62**	0
Equal-input state	**65.23**	**0.377**	3.45	0
Equal-input state+guard	**65.23**	**0.377**	3.45	0

Table 91: Gemma on 256 Taskmaster-2 prefixes under the completed final reference. Invalid outputs remain incorrect; macro-F1 uses the pooled four-class confusion matrix.

Condition	Accuracy (%) ↑	Macro-F1 ↑	FIRE recall (%) ↑	Invalid ↓
A. Representation and controller comparisons				
Direct	32.03	0.309	**93.33**	0
Summary	34.77	0.341	73.33	0
Compiled	**60.55**	**0.469**	3.33	0
StateMem-inspired	32.03	0.323	90.00	0
PIS-inspired	46.48	0.362	0.00	0
Acquired-state rule	57.81	0.396	0.00	0
Compiled+guard	59.38	0.456	3.33	0
B. Guard component ablations (reference: Compiled+guard)				
No binding check	59.38	0.456	3.33	0
No citation check	59.38	0.456	3.33	0
C. Equal decision-input token comparisons				
Equal-input summary	35.94	0.333	**80.00**	0
Equal-input state	**58.20**	**0.436**	3.33	0
Equal-input state+guard	56.64	0.418	3.33	0

Table 92: Llama on 256 Taskmaster-2 prefixes under the completed final reference. Invalid outputs remain incorrect; macro-F1 uses the pooled four-class confusion matrix.

Condition	Accuracy (%) ↑	Macro-F1 ↑	FIRE recall (%) ↑	Invalid ↓
A. Representation and controller comparisons				
Direct	17.97	0.156	**90.00**	0
Summary	25.39	0.210	63.33	0
Compiled	**50.78**	**0.348**	3.33	0
StateMem-inspired	20.70	0.177	60.00	0
PIS-inspired	32.42	0.257	6.67	0
Acquired-state rule	27.34	0.107	0.00	0
Compiled+guard	**50.78**	**0.348**	3.33	0
B. Guard component ablations (reference: Compiled+guard)				
No binding check	50.78	0.348	3.33	0
No citation check	50.78	0.348	3.33	0
C. Equal decision-input token comparisons				
Equal-input summary	37.50	0.303	**36.67**	0
Equal-input state	**50.78**	**0.348**	3.33	0
Equal-input state+guard	**50.78**	**0.348**	3.33	0

Table 93: Per-class recall (percent) against the completed 512-prefix reference. Class supports are 88 FIRE, 107 DEFER, 90 VERIFY and 227 RETIRE.

Model	Arm	FIRE	DEFER	VERIFY	RETIRE
Gemma	Direct	71.59	8.41	35.56	20.70
Gemma	Compiled	5.68	36.45	55.56	82.38
Gemma	Compiled+guard	5.68	30.84	52.22	83.26
Llama	Direct	90.91	19.63	0.00	7.05
Llama	Compiled	1.14	97.20	1.11	81.94
Llama	Compiled+guard	1.14	97.20	1.11	81.94

Table 94: Gemma: paired diagnostic contrasts on all 512 prefixes under the completed reference. Every interval uses the same dialogue resamples; equal-input comparisons use the equal-input summary.

Contrast	Fix/harm	Gain (pp)	95% interval
Compiled minus direct	203/73	25.39	[19.92, 30.86]
Compiled+guard minus direct	197/74	24.02	[18.75, 29.30]
Guard minus compiled	2/9	-1.37	[-2.54, -0.20]
Equal-state+guard minus equal-summary	121/93	5.47	[-0.20, 11.13]
Equal-state minus equal-summary	127/89	7.42	[1.76, 13.09]
Binding check added	0/0	0.00	[0.00, 0.00]
Citation check added	0/0	0.00	[0.00, 0.00]

Table 95: Llama: paired diagnostic contrasts on all 512 prefixes under the completed reference. Every interval uses the same dialogue resamples; equal-input comparisons use the equal-input summary.

Contrast	Fix/harm	Gain (pp)	95% interval
Compiled minus direct	254/79	34.18	[28.12, 40.43]
Compiled+guard minus direct	254/79	34.18	[28.12, 40.43]
Guard minus compiled	0/0	0.00	[0.00, 0.00]
Equal-state+guard minus equal-summary	163/52	21.68	[17.19, 26.17]
Equal-state minus equal-summary	163/52	21.68	[17.19, 26.17]
Binding check added	2/5	-0.59	[-1.56, 0.39]
Citation check added	0/0	0.00	[0.00, 0.00]

J.21 Historical padded-input result retained in full

The historical padded-input task-set endpoint uses 206 observations clustered in four generated weeks. These results measure supplied-event selection, not natural-dialogue state acquisition.

Table 96: Completed Study 1 external API experiment: exact task-set selection, distinct from four-action accuracy. Each arm has 206 observations clustered in four generated weeks. Invalid outputs remain incorrect. Input is mean recorded prompt tokens per observation. The field matcher uses no model call. Requested API aliases do not authenticate provider backend revisions.

Configuration	Input	Correct ↑	Exact (%) ↑	Mean input	Invalid ↓
DeepSeek	Full history	187/206	90.78	122,559.23	11
DeepSeek	D1-v2	199/206	**96.60**	**1,348.14**	1
Gemini	Full history	201/206	97.57	131,338.00	0
Gemini	D1-v2	205/206	**99.51**	**1,309.26**	0
Field matcher	Same ledger	205/206	99.51	–	–

J.22 Human adjudication export and conservative receipt repair

These additional analyses use the earlier two-model outputs and the final human reference on the same 512 prefixes. The disagreement subset contains 100 initial disagreements; it has a different class distribution from the full panel. Receipt repairs are post-hoc changes to saved decisions.

Table 97: Conservative receipt repair on the inspected 512-prefix panel under the completed final reference. These are exact paired counts and gains, not confidence bounds.

Model	Baseline	Fix	Harm	Gain (pp)
Gemma	Direct	32	0	6.25
Llama	Direct	32	0	6.25
Gemma	Compiled	11	0	2.15
Llama	Compiled	4	0	0.78

Table 98: Gemma on the same 100 adjudicated disagreements. Accuracy and recalls are percentages; all 12 original conditions and invalid outputs are retained. The constant-VERIFY control makes the skewed reference distribution explicit.

Condition	Acc.	Macro-F1	FIRE	DEFER	VERIFY	RETIRE	Invalid
Direct	28.00	0.215	75.0	7.1	29.0	11.1	0
Summary	17.00	0.122	25.0	21.4	17.4	0.0	1
Compiled	46.00	0.305	0.0	35.7	52.2	55.6	0
StateMem-inspired	29.00	0.221	87.5	7.1	29.0	11.1	0
PIS-inspired	24.00	0.167	0.0	71.4	18.8	11.1	0
Acquired-state rule	54.00	0.273	0.0	0.0	69.6	66.7	0
Compiled+guard	44.00	0.289	0.0	28.6	49.3	66.7	0
No binding check	44.00	0.289	0.0	28.6	49.3	66.7	0
No citation check	44.00	0.288	0.0	28.6	49.3	66.7	0
Equal-input summary	15.00	0.155	37.5	35.7	8.7	11.1	1
Equal-input state	27.00	0.229	0.0	64.3	18.8	55.6	1
Equal-input state+guard	25.00	0.205	0.0	57.1	17.4	55.6	1
Constant VERIFY	69.00	0.204	0.0	0.0	100.0	0.0	0

Table 99: Llama on the same 100 adjudicated disagreements. Accuracy and recalls are percentages; all 12 original conditions and invalid outputs are retained. The constant-VERIFY control makes the skewed reference distribution explicit.

Condition	Acc.	Macro-F1	FIRE	DEFER	VERIFY	RETIRE	Invalid
Direct	8.00	0.077	62.5	21.4	0.0	0.0	0
Summary	12.00	0.102	62.5	50.0	0.0	0.0	0
Compiled	17.00	0.146	0.0	92.9	1.4	33.3	0
StateMem-inspired	12.00	0.098	62.5	50.0	0.0	0.0	0
PIS-inspired	54.00	0.210	0.0	14.3	75.4	0.0	0
Acquired-state rule	69.00	0.204	0.0	0.0	100.0	0.0	0
Compiled+guard	17.00	0.146	0.0	92.9	1.4	33.3	0
No binding check	16.00	0.139	0.0	85.7	1.4	33.3	0
No citation check	17.00	0.146	0.0	92.9	1.4	33.3	0
Equal-input summary	14.00	0.162	25.0	71.4	0.0	22.2	0
Equal-input state	16.00	0.139	0.0	85.7	1.4	33.3	0
Equal-input state+guard	16.00	0.139	0.0	85.7	1.4	33.3	0
Constant VERIFY	69.00	0.204	0.0	0.0	100.0	0.0	0

Table 100: Paired effects on the initial 100 disagreements. Descriptive percentile intervals resample the 71 contributing dialogues within source and original sampling strata (10,000 draws, seed 9082026); no multiplicity correction is applied.

Model	Contrast	Fix/harm	Gain (pp)	95% interval
Gemma	Compiled minus direct	29/11	18.00	[6.32, 29.35]
Gemma	Compiled+guard minus direct	28/12	16.00	[3.16, 28.16]
Gemma	Guard minus compiled	1/3	-2.00	[-5.94, 1.94]
Gemma	Equal-state+guard minus equal-summary	18/8	10.00	[-0.98, 20.43]
Llama	Compiled minus direct	14/5	9.00	[1.94, 16.33]
Llama	Compiled+guard minus direct	14/5	9.00	[1.94, 16.33]
Llama	Guard minus compiled	0/0	0.00	[0.00, 0.00]
Llama	Equal-state+guard minus equal-summary	6/4	2.00	[-4.04, 8.33]

J.23 COMPLETED FROZEN NATIVE INTERFACE COMPARISON

The later native controller comparison uses two generated weeks, with 159 due opportunities and 155 observation steps. Each controller follows its own trajectory; pooled counts and usage retain those units.

Table 101: Pooled finite-week counts, with 159 due opportunities and 155 observation steps per row. Exact is the number of steps with an exactly correct action set, including empty-set steps. Calls include queries and invalid responses.

Model	Interface	Hit ↑	Late	Miss ↓	FA ↓	Exact ↑	Calls
Gemma	Legacy four	8	0	151	**34**	**43**	422
Gemma	Explicit single	12	1	146	39	42	456
Gemma	Explicit four	**23**	0	**136**	94	37	157
Llama	Legacy four	0	0	159	**0**	**51**	155
Llama	Explicit single	**27**	4	**128**	172	26	224
Llama	Explicit four	8	0	151	61	41	157

Table 102: Recorded costs over both weeks. Input and output tokens are server-reported usage, available for every call. Time is summed trajectory duration in seconds and is descriptive; models and arms were run sequentially, not in a randomized latency benchmark.

Model	Interface	Calls	Input tokens ↓	Output tokens ↓	Time (s) ↓
Gemma	Legacy four	422	7,401,182	71,178	3279.9
Gemma	Explicit single	456	6,966,128	**7,899**	2269.9
Gemma	Explicit four	**157**	**1,855,004**	20,469	**831.2**
Llama	Legacy four	**155**	1,958,571	21,865	664.5
Llama	Explicit single	224	2,315,211	**4,493**	**551.6**
Llama	Explicit four	157	**1,790,431**	18,152	555.9