INDEX, FORMULATE, OR RESOLVE? HOW EXTERNAL SUPPORT CHANGES WHAT AGENTS DO

Anonymous authors
Paper under double-blind review

ABSTRACT

When agents make decisions from historical records, retrieving evidence is only the start. They must locate relevant records, formulate and resolve the relation questions needed to use them, then calculate, construct, and submit a usable plan. Across three progressively finer-grained studies, we compare ways of dividing this work by organizing or completing different parts before execution. Study 1 asks whether explicitly identifying which historical records matter and which instructions still apply improves the final artifact, and whether any benefit grows when less interpretation is supplied; the expected gain is not established. Study 2 separates indexing, organizing unresolved questions, and supplying intermediate results. Study 3 further separates reference recovery from judgments about current rules and status. Across all three studies, source evidence remains complete, and agents still have to calculate, construct, and submit their own artifacts. In Study 3, adding correct rule judgments after reference recovery lowers mean provider-reported reasoning in seven of eight configurations; four history-resampling intervals lie entirely below zero. DeepSeek Pro is the only exception: one retained budget-exhausted run flips its overall JUDGE–GROUND mean positive. For Qwen Flash, adding reference results to the unresolved worksheet raises full delivery from 36/60 to 46/60. Further judgments yield its lowest mean reasoning and actor cost, but no additional delivery gain. Relative to sources alone, complete support lowers both reported reasoning and actor cost for Flash, whereas both OpenAI configurations also use less reported reasoning but cost more as cumulative input grows. Within each paired model family, the delivery gap narrows under complete support but remains. External support should therefore be evaluated by what work it completes, which model receives it, and which outcome matters. These are actor-side results under oracle-correct support; support production is not measured.

1 INTRODUCTION

An agent preparing a plan from historical records may find that a later notice withdraws part of an earlier decision while leaving the rest in force. Finding both records does not settle the task: the agent must determine how they relate, which instructions still apply, and what that means for the current plan. Complete evidence does not mean completed work.

External support can take over different parts of this remaining work. An index locates relevant records. A worksheet organizes the questions that still need answers and the evidence for answering them. A filled worksheet goes further by supplying intermediate relation results. The distinction matters: organizing the questions makes the unresolved work explicit, while supplying intermediate relation results resolves some of those relations before execution. The agent still has to carry those results forward into the remaining calculations, construct the artifact, and submit it.

This raises the central question of the paper: when different parts of the work are completed before execution, what changes downstream? The agent may deliver a better plan, use fewer reasoning tokens, or cost less to run, and those changes need not come from the same intervention. Related work has studied neighboring parts of this process: long-term memory benchmarks examine how assistants use information across conversations (Wu et al., 2025), while render-matched evaluations separate presentation effects from memory updates (Jiang et al., 2026). We instead hold source availability

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.

fixed and vary which intermediate results have already been supplied before execution, then follow the consequences through artifact delivery and resource use.

Study 1 begins with the coarsest comparison: whether telling the agent which historical records and instructions apply improves its artifact compared with leaving those relationships unresolved, and whether any benefit is greater when less interpretation has already been supplied. The expected gain is not established. More importantly for the next step, this comparison does not isolate which part of the preparation matters. Study 2 therefore separates record indexing, organizing unresolved questions, and supplying intermediate results. Those intermediate results still combine reference recovery with judgments about current rules and status, so Study 3 separates the two.

That finer decomposition shows that the effects of external support cannot be captured by a single outcome. Adding correct rule judgments after reference recovery often lowers reported reasoning, but the step that improves delivery need not be the one that minimizes reasoning or actor cost, and the pattern varies across models. The within-family comparisons further show that support can improve the cost–delivery trade-off for some receivers, narrowing delivery gaps while preserving large cost advantages. Together, these results shift the question from whether external support helps to which parts of the work are completed before execution, which model receives that support, and which outcome matters.

2 External Support as Work Allocation

2.1 Identifying records and applying rules

Consider a file-retention task. An earlier exception extends retention for both original and backup files; a later notice withdraws only the backup extension. Both records are available, but two pieces of relational work remain. The agent must first identify which exception the task and the notice refer to, and then determine which provisions are still in force. We call these operations reference recovery and rule judgment. Supplying these intermediate relation results changes what has already been resolved before execution, but it does not provide the final plan: the agent must still select the applicable retention periods, calculate expiry dates, construct the required artifact, and submit it.

Figure 1 makes this distinction concrete using Study 3's final decomposition; the earlier studies address the same broader allocation problem with coarser interventions.

A. Common records and task parameters

B51: keep originals for **5 years** and backups for **2 years**. E42: extend these periods to **9 years** and **4 years**.
V71: withdraw **only the backup extension**.
Public lookup records allow both references below to be resolved.
Created: day 100; current day: 2,000; synthetic year: 360 days.

B. Which intermediate answers are supplied?

Question	S	GROUND	JUDGE
Which exception does the task refer to?	Unfilled	E42	E42
Which exception does the notice target?	Unfilled	E42	E42
Does the original-file extension remain in force?	Unfilled	Unfilled	Yes
Does the backup extension remain in force?	Unfilled	Unfilled	No

C. Remaining work: calculate, write, and submit the plan.
Originals: 9 years $\rightarrow 100 + 9 \times 360 = 3{,}340 > 2{,}000 \rightarrow$ protect.
Backups: 2 years $\rightarrow 100 + 2 \times 360 = 820 \leq 2{,}000 \rightarrow$ plan migration.
The artifact must specify originals, backups, migration, and protection.
These calculations and the completed artifact are **not supplied**.
Migration remains a plan; the agent does not move actual files.

Figure 1: What the support supplies, and what the agent still has to do. This translated, condensed SS-14 example separates reference results from rule judgments. S, GROUND, and JUDGE use the same worksheet outside these fields. The worked output explains the required calculations; it is not supplied to the agent or taken from a model execution.

The three studies refine the allocation question in different ways. Study 1 crosses three preprocessing settings (N2, N1, and N0) with four support conditions (C, S, D^+, and D^-), testing both whether correct relational bindings improve the artifact and whether their effect changes as more interpretation is already supplied. Study 2 replaces that design with two reference-representation settings and five support conditions, separating record indexing from the organization of unresolved relation questions and from supplied intermediate relation results. Those supplied results still bundle reference recovery with rule and status judgment. Study 3 then separates reference recovery from rule and status judgment directly.

2.2 Six ways to prepare the same task

Study 3 operationalizes this final split as six preparation conditions (Table 1).

Table 1: All include the complete sources, public rules, task parameters, output requirements, and initial workspace. Each row after U adds preparation to the preceding condition.

Condition	Additional preparation
U — Sources alone	No auxiliary preparation.
I — Record index	Record identities and locations.
QST — Explicit questions	The relation questions relevant to the task.
S — Unresolved worksheet	Candidate referents or values, evidence pools, and potential dependencies.
GROUND — Reference results	Identified referents, designated registration attributes, and their designated comparisons, filled into the worksheet.
JUDGE — Rule judgments	Correct judgments about validity, scope, permission, or current status, added to the grounded worksheet.

These conditions support different kinds of comparison. GROUND–S isolates the effect of supplying reference-recovery results on an otherwise unresolved worksheet. JUDGE–GROUND, our primary comparison, then adds correct rule and status judgments after those references have already been resolved. The supplied intermediate results in both conditions are author-prepared and oracle-correct. GROUND–U and JUDGE–U answer a broader question by comparing the full preparation packages with sources alone, including the earlier indexing and question organization. In every comparison, the source records remain available; only the organization and intermediate relation information supplied before execution change, with corresponding changes in input length.[1]

3 From Bundled Support to Separate Kinds of Work

3.1 Study 1: Does correct binding improve the artifact?

Study 1 tests the coarsest version of the question: whether supplying correct relationships between historical records and current instructions improves the final artifact, and whether any benefit is greater when less interpretation has already been supplied. Four supports provide a prose memo (C), leave relationships unresolved in a scaffold (S_0), fill them correctly (D^+), or introduce one predetermined incorrect binding using a real competing source (D^-). N2 supplies source facts, N1 adds local interpretation, and N0 also adds an ordinary handoff. Twenty tasks, three settings, four supports, and four fixed configurations yield 960 selected executions.

Neither primary comparison supports the prediction. In N2, S_0 scores 97.8125% and D^+ scores 95.3125%: a difference of -2.50 percentage points (pp), with the original joint-primary 97.5% task-resampling interval $[-10.00, 0.00]$. The binding effect in N2 minus its effect in N0 is -0.9375 pp $[-4.375, +0.9375]$.

Of the 960 executions, 918 attain full scores; thirty-two of the forty-two non-full scores are zeros under the original integrity rule, and one task accounts for the entire N2 point difference. The primary comparisons therefore do not establish the expected artifact benefit of correct binding. But the post-run resource diagnostic points to a different pattern. Under N2, incorrect bindings increase cumulative generated output relative to the unresolved scaffold in three of the four configurations,

[1] Intervention construction, candidate sets, and field dependencies: Appendix B.

while the effect of correct bindings is mixed. This diagnostic includes reasoning together with other generated output, so it does not isolate reasoning use. Still, it suggests that relational support can change how an agent executes even when the expected artifact-score gain is absent. Study 2 follows that clue by separating the preparation more finely.

3.2 Study 2: Organizing questions or supplying results

Study 2 decomposes the preparation more finely instead of treating supplied relational support as a bundled intervention. A record index (I) locates records. An unresolved worksheet (S) adds questions, candidate referents or values, evidence routes, and dependencies while leaving the answers unresolved. A correctly filled worksheet (D^+) supplies the intermediate relation results without changing that organization. We call the I$\rightarrow$S addition Formulate and the S$\rightarrow D^+$ addition Resolve. A corrupted worksheet (D^-) changes a prespecified root judgment and its dependent consequences, while a prose memo provides a simpler reference. The worksheet used here differs from Study 1's S_0 scaffold.

Each run crosses twenty histories, two reference representations, five supports, and four configurations: 800 planned coordinates. N2 requires composing alias and registration mappings to identify references; N1 expands identifying attributes in the input. Both retain the source facts. Across three runs, 2,385 of 2,400 planned coordinates produced usable observations. Fifteen technical-missing coordinates were later completed under their original settings for a separately labelled sensitivity. The results below use that completed matrix, while the original missingness remains preserved.

Under N2, the two adjacent additions show different reasoning patterns. Resolve lowers mean reported reasoning in eleven of twelve configuration-by-run comparisons, whereas Formulate increases it in eight. For Grok-4.6, these opposite directions occur in every run: its three-run mean paired difference is $+626.78$ tokens for Formulate and -606.05 tokens for Resolve.

The contrast is recurring, but not universal. DeepSeek's first-run Resolve reduction is 46.67%, while its second-run mean is positive. After averaging repeats, all four N2 configurations use less mean reasoning under D^+ than under S, and more under D^- than under D^+. Artifact and delivery scores do not move in the same way: memos and indexes remain competitive, the fixed corruptions do not consistently lower scores, and N1 shows different resource patterns.

What Study 2 makes clearer is that organizing unresolved work and supplying its intermediate results are not the same intervention. Under N2, Resolve repeatedly coincides with lower reported reasoning, but that pattern does not translate into a single ordering of delivery or artifact quality. One ambiguity remains: D^+ still supplies reference-recovery results and rule or status judgments together. Study 3 separates those two pieces.

4 Study 3: Separating Reference Recovery from Rule Judgment

4.1 Tasks, receivers, and execution

Study 3 carries this final separation into a shared experimental design. It uses twenty synthetic Chinese-language histories derived from Study 2's composition-required setting, with revised worksheets and a typed artifact interface. The tasks cover retention, scheduling, publication, capacity, compatibility, and authorization in bounded local workspaces. Each execution receives one of the six preparation conditions in Table 1 and must calculate, construct, and submit its own artifact.

Study 3a tests DeepSeek V4.1 Flash, Qwen3.7 Max, GPT-5.6 Sol, and Grok-4.6. These configurations fully deliver 1,426 of 1,440 selected executions, so the panel contains little variation in full delivery. After inspecting 3a, we added Study 3b with DeepSeek V4 Pro, Qwen3.7 Flash, GPT-5.6 Luna, and Grok-4.3 under the same histories, support packets, tools, scorers, and resource policy. Because this second panel was added after 3a was inspected, we treat it as an extension rather than an independent confirmation. We retain the 3a/3b labels to preserve that research sequence. Together, the two panels contain 2,880 selected executions: twenty histories, six supports, eight configurations, and three repeats.

Each execution allows 64,000 generated tokens including reasoning, 32 model turns, 80 tool attempts, and 1,200 seconds. Conditions are position-balanced within each receiver. Bounded tools allow

reading, searching, editing, checking, and submitting files; there is no unrestricted shell or Python interpreter. The public checker validates the output interface but does not reveal hidden task correctness. Executions with incorrect content, non-submission, refusal, or budget exhaustion remain in the scientific analysis rather than being filtered out.

4.2 Measuring content, delivery, and resources

The scorer evaluates both the captured artifact and the trusted submission record. The artifact score Q_3 is the fraction of four task-local content groups that pass, subject to the integrity rule. C_3 records a trusted completed submission with a valid public interface. Full delivery requires both:

$$F_3 = 1\{Q_3 = 1 \wedge C_3 = 1\}.$$

These measures distinguish content correctness from successful delivery. A correct but unsubmitted artifact can therefore have $Q_3 = 1$ and $F_3 = 0$, while a valid submission can still contain an incorrect plan.

Reported reasoning is the provider-reported reasoning-token usage summed across an execution. We compare it within each configuration because these counters are not a common unit of cognition across providers. Actor cost is computed from selected-execution usage under the archived tariffs, including cache and input-tier rules. It excludes the cost of producing or maintaining external support and the cost of non-selected attempts. We also retain cumulative input, output, model turns, tool attempts, and wall time.

Within each receiver, each history–condition cell has three executions. We first average those repeats within each history and condition, then compare the resulting twenty paired history means. The saved 100,000 whole-history resamples provide pointwise 95% intervals for sensitivity to history composition, conditional on the observed executions. Relative changes divide each paired mean difference by the corresponding reference-condition mean.[2]

5 What Changes When Work Is Completed in Advance?

5.1 Rule judgments still matter after reference recovery

Once the referenced objects are identified, supplying their rule judgments can still change execution. Adding correct rule judgments after reference recovery lowers mean provider-reported reasoning in seven of eight configurations (Table 2). Four history-resampling intervals lie entirely below zero—Qwen Max, Sol, Luna, and Grok-4.6—while the other four cross zero.

Table 2: Adding rule judgments after reference recovery. ΔR is the JUDGE–GROUND mean difference. Means and original 95% history-resampling intervals are in reported reasoning tokens; relative changes use the GROUND mean. All selected outcomes contribute.

Configuration	GROUND	JUDGE	ΔR	95% history interval	Relative
DeepSeek V4.1 Flash, 3a	4,330.0	3,939.5	-390.5	[-1,227.0, +350.0]	-9.02%
DeepSeek V4 Pro, 3b	3,747.8	4,560.4	+812.7	[-511.9, +2,932.7]	+21.69%
Qwen3.7 Max, 3a	4,357.8	3,807.5	-550.3	[-797.1, -318.7]	-12.63%
Qwen3.7 Flash, 3b	5,789.4	4,873.5	-915.9	[-2,253.2, +207.3]	-15.82%
GPT-5.6 Sol, 3a	337.6	308.7	-28.9	[-45.2, -13.1]	-8.57%
GPT-5.6 Luna, 3b	603.6	513.3	-90.3	[-181.6, -31.5]	-14.96%
Grok-4.6, 3a	4,736.3	4,077.8	-658.5	[-1,006.5, -355.7]	-13.90%
Grok-4.3, 3b	3,642.7	3,322.7	-320.0	[-657.8, +9.7]	-8.78%

The broader pattern is similar when the full support packages are compared with sources alone. All eight configurations have lower mean reasoning under GROUND than U, and seven have lower means under JUDGE than under both GROUND and U (Figure 2). Across the 24 configuration-by-repeat suite means, JUDGE–GROUND and JUDGE–U are each negative in 22. These contrasts answer different questions: JUDGE–GROUND isolates added rule judgments after reference recovery, whereas comparisons with U also include the earlier indexing and organization.

[2]Scoring, accounting, and full paired results: Appendices C–F.

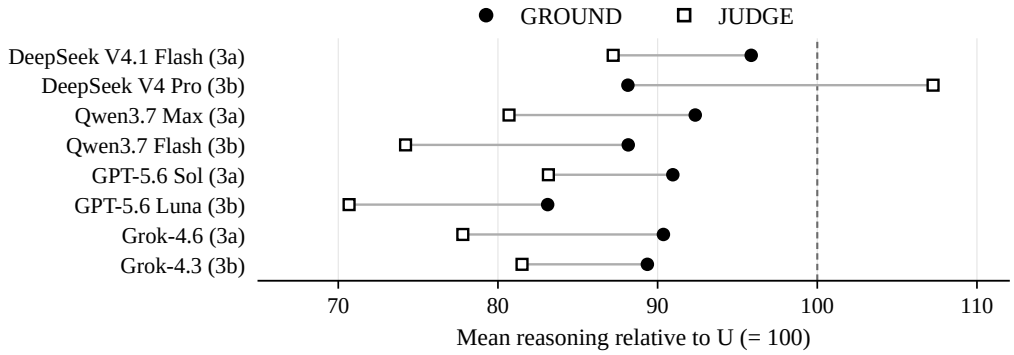

Figure 2: Mean reported reasoning relative to U ($= 100$). Filled circles and open squares denote GROUND and JUDGE. Horizontal segments connect the two package means within each configuration. DeepSeek Pro includes the budget-exhausted run.

DeepSeek Pro is the only configuration whose overall JUDGE–GROUND mean increases. One retained SS-03 execution has a large influence: in repeat 2, JUDGE uses 62,189 reasoning tokens and reaches the generation limit without full delivery. Excluding that matched U/GROUND/JUDGE repeat changes Pro's mean JUDGE–GROUND difference from $+21.69\%$ to -0.23% with equal history weights. We nevertheless keep the execution in the formal analysis, and it is not the whole explanation: fourteen of the twenty history-level differences are also positive. DeepSeek Flash shows the opposite kind of instability: its overall JUDGE–U mean is negative even though its paired-history median is positive.[3]

5.2 Delivery and reasoning respond differently

Lower reported reasoning, however, does not necessarily mean better delivery. For Qwen Flash, all 360 selected executions complete a valid submission, so differences in full delivery reflect whether the submitted artifact passes every required content check. Figure 3 shows all six preparation conditions.

The largest adjacent delivery gain occurs from S to GROUND, when reference-recovery results are supplied. Full delivery rises from 36/60 to 46/60, a gain of 16.67 pp $[+1.67, +33.33]$, while mean reasoning changes only slightly, from 5,908.02 to 5,789.43 tokens. The reasoning interval crosses zero.

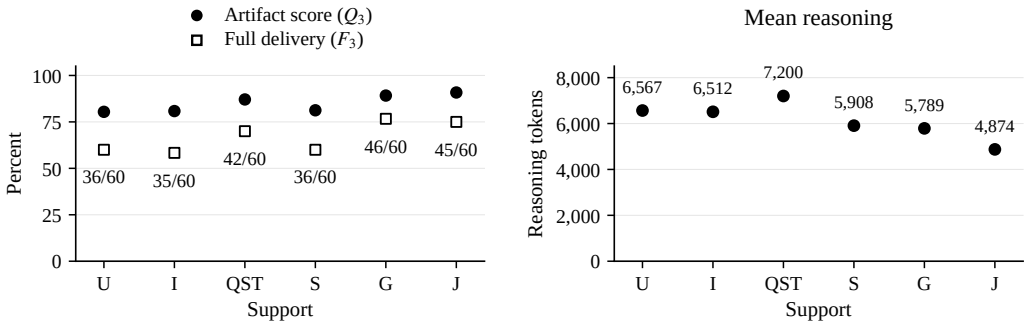

Figure 3: Qwen Flash across all six supports. Left: mean artifact score Q_3 and full delivery F_3, with delivery counts labelled. Right: mean reported reasoning. G and J denote GROUND and JUDGE. Each condition has sixty executions, and $C_3 = 1$ throughout. Each point is a separate condition mean.

The next addition produces a different pattern. From GROUND to JUDGE, mean reasoning falls to 4,873.53 tokens, while mean artifact score rises from 89.17% to 90.83%. Full delivery, however,

[3]Paired distributions and influence diagnostics: Appendix G.

changes from 46/60 to 45/60. The JUDGE–GROUND reasoning difference is negative in all three repeats, although its history-resampling interval crosses zero; the score and full-delivery intervals also cross zero. In the observed means, the delivery gain appears when reference results are supplied, whereas the lowest reasoning use appears only after rule judgments are added.

Across the full package comparison, JUDGE still improves on sources alone: mean artifact score rises by 10.42 percentage points, full delivery by 15 points, and mean reasoning falls by 1,693.40 tokens. But those net gains are not uniform across executions. Of the nine additional full deliveries, fifteen matched executions improve while six worsen.

The saved SS-14 retention plan shows what one such change looks like. Under U, Qwen Flash assigns backups three years and an expiry day of 1,180. Under JUDGE, it uses the rule-consistent two years and day 820. Both conditions retain the nine-year provision for original files.[4]

Luna shows that the sequence need not look the same. Its full deliveries rise from 54/60 under S to 55/60 under GROUND and 57/60 under JUDGE. From GROUND to JUDGE, mean score rises from 95% to 97.92% and mean reasoning falls from 603.60 to 513.28 tokens. The reasoning interval lies below zero, while the score and delivery intervals cross zero. Across all six conditions, however, QST has Luna's highest observed score and full-delivery count (58/60), whereas JUDGE has its lowest mean reasoning. More completed work therefore does not produce a single ordering across delivery, artifact score, and reasoning use.

5.3 Fewer reasoning tokens do not always cost less

Reasoning use and actor cost can move in different directions. Table 3 compares both GROUND and JUDGE with sources alone, retaining delivery counts alongside resource changes for all eight configurations. Under JUDGE, Sol and Luna use less reported reasoning, yet their actor costs rise by 30.33% and 33.68%. Cumulative input grows by 35.37% and 45.46%, with no reported cache hits. Under the archived tariffs, the savings in generated output are not enough to offset those additional input charges. Because multi-turn execution sends context again, the cost of added support depends on cumulative input rather than only on the initial packet size.

Table 3: Preparation versus sources alone, with U → GROUND (G) and U → JUDGE (J) side by side. Full deliveries are out of sixty per condition. Resource changes are $100(\bar{x}_{\text{support}} - \bar{x}_U)/\bar{x}_U$; input is cumulative. Actor cost uses archived tariffs for selected executions only and excludes producing or maintaining support.

Configuration	Full delivery		ΔReasoning (%)		ΔInput (%)		ΔCost (%)	
	U → G	U → J	U → G	U → J	U → G	U → J	U → G	U → J
DeepSeek V4.1 Flash	58 → 60	58 → 60	-4.14%	-12.78%	+12.44%	+9.25%	-6.25%	-11.69%
DeepSeek V4 Pro	60 → 60	60 → 59	-11.86%	+7.25%	+12.10%	+14.35%	-10.98%	-3.70%
Qwen3.7 Max	59 → 60	59 → 58	-7.64%	-19.31%	+33.85%	+23.00%	+10.82%	-2.24%
Qwen3.7 Flash	36 → 46	36 → 45	-11.84%	-25.79%	+24.79%	+4.71%	+12.99%	-20.15%
GPT-5.6 Sol	60 → 60	60 → 60	-9.05%	-16.84%	+40.14%	+35.37%	+34.93%	+30.33%
GPT-5.6 Luna	54 → 55	54 → 57	-16.89%	-29.32%	+44.97%	+45.46%	+34.53%	+33.68%
Grok-4.6	60 → 59	60 → 60	-9.64%	-22.20%	+12.30%	+12.01%	+7.84%	+0.14%
Grok-4.3	57 → 52	57 → 56	-10.64%	-18.49%	+32.87%	+37.81%	+17.66%	+16.49%

Costs also depend on which support condition is used as the baseline. From GROUND to JUDGE, Sol's and Luna's mean costs fall by 3.41% and 0.64%, despite their increases from U to JUDGE. For Flash, GROUND–S raises mean cost by 5.40%, with an interval crossing zero; JUDGE–GROUND lowers it by 29.33%, with the cost interval below zero. Its delivery gain and cost reduction accompany different additions. Overall, JUDGE lowers Flash's cost by 20.15% relative to U, an absolute difference of -0.00454 CNY $[-0.00949, -0.00065]$. Producing and maintaining the support remains outside these actor-cost comparisons.

[4] Saved artifacts and matched delivery transitions: Appendix H.

5.4 How does support change the comparison between models?

So far, we have compared preparation conditions while holding the receiver fixed. We next examine how support changes the gap between configurations within the same model family. After both Study 3 panels were complete, we used the saved condition means for descriptive comparisons within Qwen3.7 Max/Flash and GPT-5.6 Sol/Luna. These comparisons introduce no new executions or inferential tests. The original configurations and controls remain unchanged, and the pairs are named configurations rather than experimentally calibrated capability levels. Figure 4 shows U, GROUND, and JUDGE.

For Qwen, Flash improves from 36/60 full deliveries under U to 45/60 under JUDGE, while Max changes from 59/60 to 58/60. The same-support delivery gap therefore narrows from 38.3 to 21.7 percentage points. Part of that change comes from Max's one-delivery decline; if Max is instead held at its U result, supported Flash remains 23.3 points behind. Under JUDGE, their mean artifact scores are 90.83% and 97.92%, respectively. Support closes part of Flash's delivery gap without eliminating it.

The OpenAI pair starts closer together. Luna moves from 54/60 full deliveries and a 97.08% mean artifact score under U to 57/60 and 97.92% under JUDGE, while Sol remains at 60/60 with a full artifact score. The observed delivery gap falls from 10 to 5 percentage points. Luna's JUDGE–U intervals for both artifact score and full delivery cross zero, so these observed changes remain uncertain.

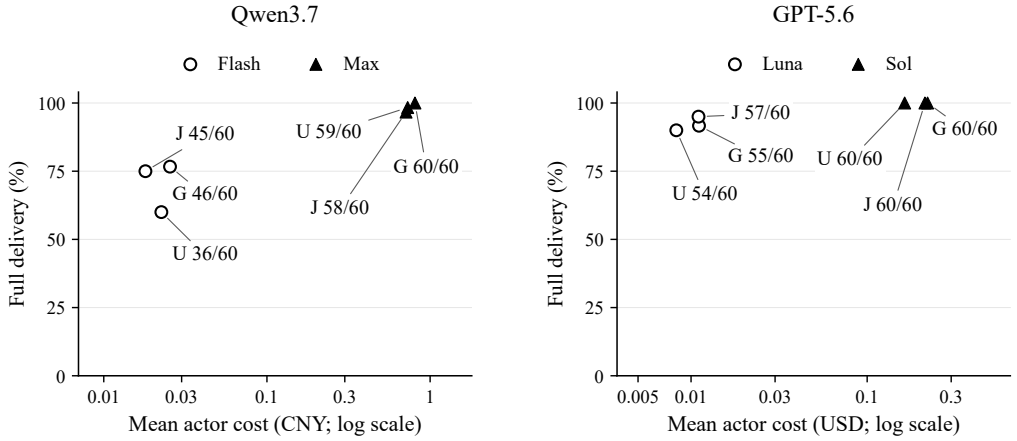

Figure 4: Mean actor cost (logarithmic horizontal axis) and full delivery (vertical axis) under U, GROUND (G), and JUDGE (J). Open circles denote Flash or Luna; filled triangles denote Max or Sol. Each point summarizes sixty selected executions; currencies remain separate. Leader lines attach labels only.

Flash and Luna are already much cheaper than Max and Sol before support is added. Under U, Flash costs 3.09% of Max and Luna 5.05% of Sol. Under JUDGE, the same-support ratios are 2.53% and 5.18%. If Max and Sol are instead held at U, supported Flash costs 2.47% of its reference and supported Luna 6.75%. All ratios compare condition means within one family and currency.

The two families move through this cost–delivery trade-off differently. Flash gains nine full deliveries from U to JUDGE while its own mean actor cost falls by 20.15%; Luna gains three full deliveries while its mean actor cost rises by 33.68%. In both families, a large cost difference remains because most of it was already present before support. These comparisons do not show that support makes the configurations interchangeable. Instead, they show that support can change where a lower-cost receiver sits on an existing cost–delivery trade-off, and that the direction of that change depends on the receiver.

6 Related Work

ReAct, Plan-and-Solve, and Reflexion organize reasoning and action during or across executions (Yao et al., 2023; Wang et al., 2023; Shinn et al., 2023). LongMemEval, StateMem, and Mem2ActBench study how agents use persistent or evolving information, including memory-grounded tool use (Wu et al., 2025; Fan et al., 2026; Shen et al., 2026). Our setting instead keeps the source records available and varies which intermediate relation results are supplied before a fresh execution.

Recent controlled evaluations also separate supplied information from what a model must infer. RENDER varies reader-facing evidence, Presentation, Not Mechanism uses render-matched controls, Explicit State Elicitation tests responses to supplied state labels, and AuthMem-Bench traces memory authority into tool decisions (Si et al., 2026; Jiang et al., 2026; Chen et al., 2026; Zhan et al., 2026). Our S/GROUND/JUDGE comparison holds the worksheet fixed outside the resolved fields and separates supplied reference results from rule judgments; U/I/QST/S additionally vary earlier organization. We measure the resulting changes in artifact delivery, reported reasoning, and actor cost.

7 Discussion and Limitations

These experiments suggest a different way to think about external support. The question is not simply how much information an agent receives, but which parts of the work have already been done before the agent begins. An unresolved question, an identified reference, and a rule judgment may concern the same underlying records, yet they leave very different work for the agent to perform. Across the three studies, moving that boundary changes execution in ways that no single outcome captures: the step that improves delivery need not be the step that reduces reported reasoning, and the step that reduces reasoning need not reduce actor cost. External support is therefore better understood as a redistribution of work than as a uniformly stronger prompt.

That redistribution also has to be considered together with the receiver. Support does not act on an abstract "agent" whose cost and behavior are fixed. The same kind of preparation can move different configurations through the cost–delivery trade-off in different directions. In our within-family comparisons, some of the largest cost differences exist before support is added at all; support can then change delivery without erasing those differences, and in some cases can improve delivery while preserving a large cost advantage. This makes model choice and work allocation separate but interacting decisions. The experiments do not identify a universally best package, nor do they establish that one supported configuration can replace another. They show instead why the useful question is conditional: what work is worth completing in advance, for which receiver, and for which outcome?

The clean separation that makes this question measurable is also an important limitation of the study. The supplied intermediate results are oracle-correct, so the experiments isolate what happens after that work has been completed; they do not measure how difficult, costly, or error-prone it is to produce that support in the first place. Real systems would have to generate, verify, update, and sometimes correct it. The tasks are synthetic Chinese-language histories in bounded workflows, full delivery also requires successful submission, and reported reasoning is provider-reported usage rather than direct access to hidden computation. Three repeats expose some execution variation but cannot characterize rare costly events, and the Study 3 panels ran at different times, so cross-configuration differences may partly reflect provider changes. The within-family comparisons are descriptive rather than tests of equivalence or optimality. These boundaries matter because the present experiments identify consequences of moving work outside the model, not the complete engineering or economic case for doing so.

8 Conclusion

Complete evidence does not complete the task; it only determines what information is available to work with. Across the three studies, moving different parts of that work outside the model changes what remains for the agent, and those changes do not collapse to a single notion of improvement. Correct rule judgments often reduce reported reasoning after references are already resolved, yet delivery and actor cost can move differently, and the same preparation can matter differently for different receivers. The practical question is therefore not whether more external support is better, but how the boundary of work is drawn for a particular receiver and outcome.

Reproducibility Statement

The companion anonymous evidence supplement provides frozen task and support materials, task-local scorers, selection indices, captured artifacts, saved scores, and native usage for the three studies. This PDF contains the figures, numerical tables, and supplementary methods; the editable manuscript source is retained separately by the author and is not part of the submitted evidence attachment. Appendix I describes how to inspect interventions and selected executions and how to verify saved results offline. Manuscript compilation and figure generation are distinct from reproducing an experimental score.

AI Use Statement

The author conceived and led the research, defined the research questions and experimental design, and determined the interpretation and conclusions. Language-model tools were used under the author's direction to discuss ideas, develop synthetic materials and code, retrieve literature, assist with checks and analysis, and draft and revise text. The author takes full responsibility for the paper and its scientific claims. Model-generated suggestions and simulated reviews were not treated as independent scientific validation.

Ethics Statement

The studies use synthetic records and bounded local artifact workflows, not human-participant data or actions in a live organization. The internal historical archive retains original machine paths and execution metadata and is distinct from the curated anonymous evidence supplement. Results from these tasks are not a certification of safe deployment with imperfect memory or external judgments.

References

Alibaba Cloud. qwen3.7-flash. Model Studio documentation, n.d.a. URL https://www.alibabacloud.com/help/en/model-studio/qwen3-7-flash. Accessed September 25, 2026.

Alibaba Cloud. qwen3.7-max. Model Studio documentation, n.d.b. URL https://www.alibabacloud.com/help/en/model-studio/qwen3-7-max. Includes the qwen3.7-max-2026-06-08 snapshot. Accessed September 25, 2026.

Yihang Chen, Pin Qian, Su Wang, Chong Peng, Huan Xu, Shuaiting Li, and Yiqi Sun. Explicit state elicitation is not enough: A controlled audit of memory-policy classification. *arXiv preprint arXiv:2608.17247*, 2026. URL https://arxiv.org/abs/2608.17247.

Xinyi Fan, Miri Liu, Ruozhen Yang, Siru Ouyang, and Jiawei Han. Can agent memory systems track evolving state? *arXiv preprint arXiv:2608.19652*, 2026. URL https://arxiv.org/abs/2608.19652.

Zhaoyang Jiang, Zhizhong Fu, Zicheng Li, Yunsoo Kim, Jiacong Mi, Xuanqi Peng, Fei Teng, and Honghan Wu. Presentation, not mechanism: A render confound in deprecation-aware memory evaluation. *arXiv preprint arXiv:2607.16019*, 2026. URL https://arxiv.org/abs/2607.16019.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition, July 2026. URL https://openai.com/index/gpt-5-6/. Published July 9, 2026. Accessed September 25, 2026.

Yiting Shen, Kun Li, Wei Zhou, and Songlin Hu. Mem2ActBench: A benchmark for evaluating long-term memory utilization in task-oriented autonomous agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8173–8190. Association for Computational Linguistics, July 2026. doi: 10.18653/v1/2026.acl-long.370. URL https://aclanthology.org/2026.acl-long.370/.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *arXiv preprint arXiv:2303.11366*, 2023. URL https://arxiv.org/abs/2303.11366.

Yuan Si, Simeng Han, Daming Li, and Jialu Zhang. RENDER: Controlling reader-facing evidence in LLM memory evaluation. *arXiv preprint arXiv:2608.23568*, 2026. URL https://arxiv.org/abs/2608.23568.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. *arXiv preprint arXiv:2305.04091*, 2023. URL https://arxiv.org/abs/2305.04091.

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://arxiv.org/abs/2410.10813.

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

Qiuyang Zhan, Rui Zhang, Sheng Guo, Lepeng Zhao, and Zhuotao Liu. When memory becomes authority: Benchmarking authority collapse at the memory consolidation boundary. *arXiv preprint arXiv:2608.01679v2*, 2026. URL https://arxiv.org/abs/2608.01679v2.

GUIDE TO THE SUPPLEMENTARY EVIDENCE

The map below links each main-text question to its detailed evidence. All original results, technical-missingness records, and tail diagnostics are retained; the completed Study 2 sensitivity is reported separately from its original runs.

Question or evidence	Location	Page
Actual models, budgets, selection, and intervals	Appendix A	12
Task inputs and reference/judgment construction	Appendix B	15
Artifact checks and full-delivery requirements	Appendix C	18
Tools and archived cost accounting	Appendix D	22
Execution versions and non-selected attempts	Appendices A.3 / D.1	13 / 22
Frozen scoring boundaries and delivery-count decomposition	Appendix C.6	21
Study 1 results and original zero-score evidence	Appendix E.1	24
Study 2: original runs / completed sensitivity	Appendices E.2 / E.3	26 / 27
Study 3: all six condition levels	Appendix F.1	30
Primary repeats / all nine paired contrasts	Appendices F.2 / F.3	31 / 31
Qwen Flash: adjacent additions and intervals	Appendix F.4, Table 29	33–34
Family delivery gaps and cost references	Appendix F.5	34
DeepSeek exception and full-data tail diagnostics	Appendix G	40
SS-05 / SS-14 artifacts and delivery transitions	Appendix H	43
Compact attachment, file lookup, and score readback	Appendix I	44

A PROTOCOL, CONFIGURATIONS, AND MEASUREMENT

A.1 FIXED RECEIVERS AND RESOURCE POLICY

The named receivers identify requested configurations, not immutable server snapshots. Table 4 preserves the actual request labels. The two Study 3 panels share controls within each provider family; this does not imply equal internal computation or a capability ordering. Earlier studies use the four request labels in the 3a column, with their own archived controls.

Table 4: Study 3 requested configurations. The labels in the main text are provider display names.

Provider	Study 3a	Study 3b	Reasoning control
DeepSeek	`deepseek-flash`	`deepseek-v4-pro`	high
Qwen	`qwen3.7-max-2026-06-08`	`qwen3.7-flash`	thinking; preserved
OpenAI	`gpt-5.6-sol`	`gpt-5.6-luna`	high; all-turn context
xAI	`grok-4.6`	`grok-4.3`	high

The within-family summaries pair Qwen3.7 Max with Flash and GPT-5.6 Sol with Luna. Alibaba Cloud describes Max, including the June 8 snapshot used here, as the largest and most capable model in the Qwen3.7 series (Alibaba Cloud, n.d.b). Flash belongs to the same series (Alibaba Cloud, n.d.a) and has lower token prices under the archived tariffs in Table 12. OpenAI describes Sol as its flagship and Luna as its most cost-efficient GPT-5.6 model (OpenAI, 2026). Here *lower-cost configurations* refers to Flash's and Luna's observed actor costs under the archived tariffs. The product descriptions supply positioning, not replacement prices or experimentally calibrated capability levels. The tested OpenAI reference is Sol with the listed `high` control, rather than Sol Pro. Exact request labels, client-side controls, and panel assignments remain unchanged; the pairing neither ranks all eight receivers on a common capability scale nor exhausts the available service modes.

DeepSeek, OpenAI, and xAI use Responses interfaces; Qwen uses its native-compatible chat-completions interface. Streaming is disabled. Qwen preserves thinking and permits parallel tool calls. OpenAI and xAI retain API-supplied encrypted reasoning continuation, permit parallel calls, and disable storage. OpenAI uses low text verbosity, disabled truncation, and explicit thirty-minute caching. No seed is claimed to control provider-internal execution. The separate evidence archive records a projection of the request controls and tariffs; Appendix I distinguishes that archive from the manuscript compilation bundle.

Study 2 and Study 3 attempts allow 32 model turns, 80 tool attempts, 64,000 cumulative generated tokens including reasoning, and 1,200 seconds. Study 2 limits a request to 32,000 generated tokens for DeepSeek and 16,000 for the others. Study 3 permits the remaining cumulative allowance, up to 64,000, in one request. There is no additional local cumulative-input admission gate. Reported input is cumulative native usage, not just the initial prompt length.

A.2 Tools, scoring, and accounting

The agent-facing tools are `list_tree`, `read_text`, `search_text`, `apply_patch`, `run_check`, `request_decision`, and `submit`. There is no unrestricted shell or Python REPL. Reads and searches are bounded; patches operate on allowed files with expected-hash checks. The checker accepts fixed command IDs and validates the public interface, not hidden task correctness. Requesting a decision records an escalation without supplying new facts or an oracle answer. Submission registers a terminal and artifact references; the scorer evaluates the captured artifact independently. A textual imitation of a tool call is not a registered submission.

The trusted terminal schema distinguishes completed, blocked, and technical-failure declarations. Scientific usability concerns whether an observation exists, not whether the task succeeds. In Study 3, four artifact groups contribute equally to Q_3 under the integrity rule, while C_3 also requires completed trusted submission and the public interface. Thus $Q_3 = 1$ does not imply $C_3 = 1$ or $F_3 = 1$. The frozen targets and integrity rules are retained for all results and examples.

Selected-actor cost uses native usage and the archived tariff, including caching and applicable input or time tiers. DeepSeek and Qwen costs are denominated in CNY; OpenAI and xAI in USD. No cross-currency sum is used. These costs exclude support production and maintenance, non-selected attempts, unresolved reservations, and invoice reconciliation. Lower reported reasoning therefore cannot by itself establish lower actor cost or lower total cost.

A.3 Technical selection and original missingness

Table 5: Scientifically usable selected observations. "Original" counts selected first attempts. Replacement and completion describe technical provenance, not improved task performance.

Dataset	Planned	Original	Replacement	Completion	Missing
Study 1	960	954	6	0	0
Study 2, run 1	800	797	3	0	0
Study 2, run 2 (original)	800	791	1	0	8
Study 2, run 3 (original)	800	792	1	0	7
Study 3a	1,440	1,427	8	5	0
Study 3b	1,440	1,417	21	2	0

Study 2's second and third full-design executions were authorized together after the first, before their results were inspected. The original set contains 2,385 usable observations; strict three-run averages exist for 785 of 800 cells. A later, separately labelled sensitivity supplies the fifteen technical-missing coordinates under their original task, support, receiver, and resource settings. Each obtains a usable first completion without a further replacement. These observations are not fifteen new histories. Original missingness and estimates are preserved in Appendix E.

Selection does not discard high reasoning, refusal, incorrect content, non-submission, formatting failure, or budget termination. Study 2 retains fifteen budget terminals (seven DeepSeek, eight Qwen), distinct from its fifteen technical-missing observations. Study 3a retains no budget terminal; Study 3b retains one, the DeepSeek-Pro trajectory discussed in the main text. The archived technical closures record selection before scientific unblinding within the respective frozen campaigns. Study 3b follows the inspection of Study 3a and is not an independently planned confirmation.

Study 3a contains 1,453 attempts and 1,440 selected observations; Study 3b contains 1,465 attempts and 1,440 selected observations. The remaining 38 attempts comprise 36 original sources and two failed replacements. The attachment's `study3/nonselected_attempts_index.csv` links each to its original claim, normalized record, request ledger, available error and terminal events,

authorization excerpts, closure records, and final selected coordinate. These are retained historical records, not additional observations or a new selection.

The release's offline inspection links requests and no-submit terminals to detailed transport-error records for 32 attempts. Two further attempts retain a RemoteDisconnected event and terminal but lack the detailed transport-end fields. Three retain host-stop signals and no-submit terminals. For OpenAI SS-06/QST/repeat 2 in Study 3a, the shared-defect classification and stop snapshot remain, but the original traceback, trusted-terminal file, and root seal were not retained. The index distinguishes these evidence levels. Historical policy excerpts and controller approval timestamps are separate from independently timestamped Owner messages; the latter are not established by file modification times. Original eligibility flags, including initially ineligible sources subsequently authorized for separate completion, remain unchanged.

Study 3 balances condition positions within each receiver: each support occupies each of six positions ten times across sixty history–repeat blocks. Actual execution times are not interchangeable. In 3a a storage interruption separates original and resumed work; four original six-condition blocks cross that boundary, and technical replacements/completions occur later than their source attempts. Position balance does not remove a general provider-time effect.

Study 3b also includes a campaign budget-continuation segment; one original Qwen history–repeat block (SS-10, repeat 3) straddles that boundary, and selected replacements or completions can occur later than their original sources. This continuation advances the campaign's unstarted coordinates; it does not extend a budget-exhausted scientific trajectory. The segment contains 53 original-source coordinates, of which 48 remain selected as originals. The other five are represented by later selected attempts.

A.4 Estimands and interval interpretation

Study 1 weights the four fixed configurations equally within each whole-task vector. Its frozen bootstrap uses 20,000 resamples. The two original primary quantities use Bonferroni-adjusted 97.5% intervals; marginal intervals are 95%. Study 2 and Study 3 report configurations separately, using 100,000 shared whole-history resamples with Python `random.Random`, seeds 20260916 and 20260918 respectively, and Hyndman–Fan type-7 quantiles. Repeated conditions first average the three observed executions within a history, then form paired differences across twenty histories.

The later intervals describe pointwise history-composition sensitivity conditional on the observed executions. They do not integrate all execution noise or provide simultaneous guarantees across comparisons, receivers, or a randomly sampled task population. Missing metrics are not zero-filled, and two available executions do not constitute a three-execution mean. Formal results apply no trimming, winsorization, or success-only selection. Relative changes divide the mean paired difference by the comparison condition mean; paired medians and signs summarize different features of the observed history distribution. The Study 2 contract's 5-pp quality-loss reference is not turned into a new pass/fail test or a population non-inferiority claim.

The within-family gap and cost-ratio summaries in Appendix F.5 are post-hoc descriptive transformations of the saved condition means. They introduce no new cross-configuration interaction intervals, equivalence tests, or non-inferiority thresholds. Any intervals reproduced for the original within-configuration contrasts retain their saved values and interpretation.

B TASK MATERIALS AND THE CONSTRUCTION OF EXTERNAL SUPPORT

B.1 MATERIAL IDENTITY AND WHAT THE MODEL CAN SEE

The archived Study 3 material defines twenty task worlds, all 120 task packets, the public output interfaces, the author reference operations, and the task-local scorers. The same material is used in Study 3a and 3b. Each packet contains the complete record set, public rules, task parameters, and delivery requirements. The initial workspace also contains those records and rules, the public interface, its validator, and a writable artifact template. Author truth files, reference operations, and private scorers are not available to the model. These experimental materials belong to the separate evidence archive; the manuscript compilation bundle has the coverage described in Appendix I.

Within each task, U is constructed by returning the common packet before adding an index or calling the support builder. I appends the record index. QST adds the explicit relation questions. S adds candidate referents or values, evidence pools, and potential dependencies, while leaving the relation fields unresolved. GROUND fills the reference-recovery fields; JUDGE also fills the rule/status judgments on the same worksheet. Neither supplies the completed artifact. Thus source availability is controlled, but intermediate relation results and, in the earlier contrasts, organization are deliberately changed. Equal source access does not mean byte-identical input or equal input-token consumption.

The material directory keeps the actual historical version identifiers. In particular, `W0` inside an author truth file means the base task world; it is not the separate W0 workshop study. The material version ending in `TERMINAL_COMPAT` retains the typed scoring code and extends accepted terminal declarations to include legitimate blocked and technical-failure declarations. A model's declaration of technical failure is not, by itself, an eligible provider failure or permission to replace a poor result.

B.2 THE TWENTY TASKS

Table 6 describes the decision and remaining output work in every task. Counts of reference and judgment slots come from the archived semantic partitions. They are not equal amounts of computation: a slot may contain a record ID, an attribute tuple, or a Boolean, and tasks differ in how many records are involved. The study therefore does not interpret a percentage reduction in reasoning as the price of one elementary cognitive operation.

Table 6: All twenty Study 3 tasks and their remaining artifact work. G/J count supplied reference/judgment slots; CF counts retained author counterfactuals. Complete definitions and unabridged Chinese inputs are retained in the separate evidence archive.

Task	Setting	Remaining work	G	J	CF
SS-01	Media schedule and archive	Select effective segments, channels and retention; calculate durations and archive fields.	17	18	6
SS-02	Image licence and export	Separate internal use from withdrawn public use; calculate sizes and retain attribution.	3	6	5
SS-03	Laboratory containers	Distinguish usable, quarantined and review states; allocate containers and count labels.	5	4	4
SS-04	Cohort publication	Resolve population and aggregate permission; count bins and separate suppression from permission blocks.	9	2	4
SS-05	Dispatch and maintenance	Respect current suspension without losing maintenance duties; compute windows and total duration.	3	6	6
SS-06	Snapshot archive	Check snapshot readiness and signing authority at the relevant time; union ranges and count batches.	3	9	5
SS-07	Calibration revision	Apply the effective coefficient and offset separately; calculate converted sample values.	16	8	4
SS-08	Compatibility fallback	Reject an unapproved preferred choice; intersect build ranges and calculate checksums.	6	2	4
SS-09	Programme authorization	Apply signing authority and prospective suspension; intersect and truncate time intervals.	3	6	5
SS-10	Signature archive	Distinguish rotation, compromise and not-yet-effective credentials; select documents and sum size.	6	12	5
SS-11	Consent and quotations	Separate full withdrawal from location withdrawal; form allowed fields, aliases and excerpts.	4	6	5

Table 6 — continued

Task	Setting	Remaining work	G	J	CF
SS-12	Configuration rollback	Select naming and validation versions separately; clamp defaults and merge required fields.	3	2	4
SS-13	Notification delegation	Check principal, channel, service and scope; form channel sets and retain escalation duties.	5	4	4
SS-14	Partial retention withdrawal	Keep the surviving original exception, withdraw the backup exception; compute expiry and action sets.	2	2	3
SS-15	Capacity after cancellation	Distinguish child-interval cancellation from parent cancellation; calculate occupancy and remaining capacity.	1	4	3
SS-16	Subdelegation	Check parent purpose, term, partial revocation and independent approval; combine quotas and remaining purposes.	3	9	6
SS-17	Dual approval	Check principals, required functions and batch scope; sum ready units and retain distinct defects.	16	12	8
SS-18	Issue versus continuous licence	Separate non-retroactive issue permission from ongoing validity; calculate future permitted duration.	6	12	5
SS-19	Shared quotas	Resolve the relevant pool and valid refunds; apply ordered admission under total and project limits.	5	7	5
SS-20	Alternative authorization	Distinguish requests from actions, emergency paths and absolute prohibitions; aggregate allowed demand.	12	16	6

B.3 Separating references, rules, and residual business calculations

Let L contain the designated reference and registration inputs, P the policy and status inputs, and B the inputs used in the remaining business calculation. The author-side operations follow

$$G = g(L), \qquad J = j(G, P), \qquad A = a(G, J, B),$$

where A is the required artifact. This defines a partition of *supplied results*, not a claim that the model internally computes these functions in this order. The identity of a referenced record can be supplied while its permission, validity, scope, or status remains unresolved. A number or Boolean belongs to the layer determined by its inputs, not by its output type.

Reference candidates are formed using public selectors. In the registry-based tasks, a reference follows an alias and position mapping to identifying attributes; the author operation compares those attributes with the original records among the coarse candidates. Candidate generation retains all candidates matching the public selector rather than using the correct referent to discard alternatives. Evidence pools contain possible relevant registration and rule records, not just the gold path. Dependencies follow the public reference fields and rule templates. The full slot definitions, including candidate sets, direct reads, evidence pools, downstream consumers, and judgment families, are in each archived task's `author/SEMANTIC_PARTITION.json`.

The business calculation then uses the identified records and rule results to build the artifact. It may intersect intervals, sum quantities, select a compatible build, calculate expiry, or construct allowed and blocked sets. Some scored fields closely follow a supplied value; the design does not require every output group to involve a further difficult inference. Conversely, a correct intermediate result need not settle the final action. In SS-05, even a permitted job must fit its time window. The supplied judgment does not perform that duration calculation.

The reference results and rule judgments in Study 3 are oracle-correct. This design measures what happens when those intermediate results are supplied, rather than testing an automatic resolver that generates them. Study 3 has no corrupted-judgment condition; the fixed corruptions in Study 2 do not estimate a deployed resolver's error distribution. The model may use the worksheet in ways that differ from the author-side operations; the partition does not assign measured internal computation to individual fields. Evaluating a deployed system would additionally require the errors and costs of producing, checking, and maintaining its support.

B.4 Exact worked partition for SS-14

The following field names are the original ones, with questions translated. S leaves all four unresolved; GROUND fills the first two; JUDGE also fills the last two. All other worksheet content is held fixed.

Table 7: The four SS-14 supplied fields. Both reference questions retain E21, E33, and E42 as candidates before resolution.

Field	Supplied value	Question answered
`T61.exception_record`	E42	Which exception does the task refer to?
`V71.target_record`	E42	Which exception does the withdrawal target?
`T61.original_exception_retained`	true	Does the original-record provision remain effective?
`T61.backup_exception_retained`	false	Does the backup provision remain effective?

The common source records supply the numerical periods: B51 gives five years for originals and two for backups; E42 extends these to nine and four years. V71 withdraws only the backup extension. Given the filled judgments, the agent still chooses nine years for originals and two for backups, computes expiry from creation day 100 with a 360-day year, and compares it with day 2,000. It must write `config/retention_exception.json`, including the original, backup, migration, and protection groups, and make a trusted submission.

B.5 Author-side counterfactual checks

The material includes 97 author-side counterfactual worlds. These are construction tests, not additional experimental histories or model trials. They change a designated class of input, keep other classes fixed where specified, and compare the reference output against a separately written expected artifact. The files retain the changed world, expected output, observed author-reference output, and the change description. The base world of each task is likewise checked against its expected artifact.

SS-14 illustrates the three input classes. Redirecting the task to E33 changes its reference while leaving the withdrawal directed at E42; both E33 provisions then survive. Changing only V71's withdrawal scope to originals leaves both E42 references unchanged but reverses the two retention judgments. Changing creation day from 100 to 120 leaves references and judgments unchanged and adds twenty days to the two expiry calculations. These checks make the intended dependencies inspectable; they do not establish the model's reasoning mechanism.

Table 8: Saved SS-14 author counterfactuals. The values are reference-operation outputs, not model observations.

Changed input	References (task / notice)	Retained (original / backup)	Expiry (original / backup)
Base world	E42 / E42	true / false	3,340 / 820
Task points to E33	E33 / E42	true / true	2,260 / 1,180
Withdraw original scope	E42 / E42	false / true	1,900 / 1,540
Creation day 120	E42 / E42	true / false	3,360 / 840

B.6 What changed between the earlier tasks and Study 3

Study 3 is not a byte-identical replication of Study 2. Its interface uses explicit typed fields rather than some of the earlier encoded strings, and it separates reference and judgment operations. The archived material catalog records each change. Examples include integer time-window endpoints in SS-05, explicit rule use of service scope in SS-13, a real unapproved preferred build in SS-08, and an additional archive item in SS-10 that requires summing two eligible document sizes. SS-20 explicitly distinguishes request IDs from action names. These revisions help define the new comparison but prevent attributing all cross-study changes to the reference/judgment split alone. The lower per-request generation limits of Study 2 also differ from Study 3.

C SCORING, SUBMISSION, AND KNOWN MEASUREMENT BOUNDARIES

C.1 ARTIFACT CHECKS AND SUBMISSION ARE SEPARATE REQUIREMENTS

The scorer reads the captured artifact; it does not grade a model's explanation or hidden reasoning. Each Study 3 task defines four artifact groups in its public interface. A group passes only when its required fields have the correct types and match the author truth under the specified comparison rules. Table 9 lists the groups for all tasks. They have equal weight, irrespective of how many fields a group contains.

$$Q_{\text{raw}} = \frac{1}{4}\sum_{k=1}^{4} \mathbf{1}\{\text{group } k \text{ passes}\}, \quad Q_3 = Q_{\text{raw}}\mathbf{1}\{\text{integrity passes}\},$$

$$C_3 = \mathbf{1}\{\text{trusted terminal is completed and public interface is valid}\},$$

$$F_3 = \mathbf{1}\{Q_3 = 1 \wedge C_3 = 1\}.$$

An artifact score of one does not establish that a completed submission was registered. Conversely, a valid submission can contain an incorrect plan. These are related measures of requirements, not independent quality replications. For Qwen Flash, $C_3 = 1$ in every selected execution, so its observed full-delivery changes are changes in whether submitted content passes all four groups.

Table 9: Artifact paths and the four equally weighted content groups. Field names are the original schema keys. The full type, range, required-field and comparison rules are in each task's public interface.

Task	Artifact	Groups and fields

SS-01	`config/archive_plan.json`	`segments`: `frames_per_segment`, `segments_per_session`
		`delivery`: `export_channels`, `bytes_per_segment`
		`retention`: `retained_sessions`, `cleanup_batch_sessions`
		`audit`: `digest`, `provenance_sidecar`, `capture_location`

SS-02	`config/image_export.json`	`thumbnail`: `sizes`, `internal_bytes`
		`files`: `members`
		`credits`: `entries`
		`metadata`: `internal_assets`, `keep_days`

SS-03	`config/sample_labels.json`	`sorting`: `released`, `staged`, `zones`, `container_review`
		`labels`: `container_count`, `isolation_count`
		`isolation`: `tube_count`, `lab_records`
		`tracking`: `fields`, `receipt_count`

SS-04	`config/cohort_summary.json`	`cohort`: `members`, `member_count`
		`binning`: `raw_counts`
		`publication`: `public_bins`, `threshold_suppressed_bins`, `permission_blocked_bins`
		`privacy`: `excluded_fields`, `min_bin`

SS-05	`config/work_routes.json`	`routes`: `crew_by_job`
		`windows`: `interval_by_job`
		`maintenance`: `retain_jobs`, `total_minutes`
		`blocked`: `jobs`

SS-06	`config/index_rebuild.json`	`inputs`: `snapshots`
		`batches`: `size`, `count`, `row_count`
		`backfill`: `ranges`
		`validation`: `digest`, `preserve_base`

SS-07	`config/calibration_merge.json`	`conversion`: `values`
		`windows`: `boundaries`
		`units`: `name`, `new_revision_samples`
		`quality`: `flags`, `sample_count`

Table 9 — continued

Task	Artifact	Groups and fields
SS-08	config/package_lock.json	packages: versions dependency: bounds, in_range channels: allowed checks: checksum_total, download
SS-09	config/display_program.json	playback: ranges, play_minutes subtitles: enabled, burn_in language: codes notices: host, retain_accessibility_notice
SS-10	config/signature_archive.json	archive: documents, archive_kib verification: states review: documents retention: days, evidence_count
SS-11	config/interview_release.json	fields: by_interview anonymity: aliases excerpts: text audit: retain_records, real_names
SS-12	config/schema_rollback.json	fields: names constraints: minimum, maximum defaults: count, country preserved: fields, checksum_required
SS-13	config/alert_channels.json	channels: routes escalation: levels, required dedup: minutes, window_seconds obligations: retained_channels, fallback
SS-14	config/retention_exception.json	original: years, expiry_day backup: years, expiry_day migration: to_cold, plan_only protected: no_delete
SS-15	config/lab_capacity.json	availability: free capacity: by_room preserved: bookings conflicts: ranges, maintenance_count
SS-16	config/subdelegation.json	roles: active scope: uses_by_role, daily_units escalation: uses, count audit: retain_records, create_accounts
SS-17	config/release_quorum.json	release: ready_batches, ready_units missing: requirements preserved: signatures evidence: records, perform_release
SS-18	config/continuous_authority.json	allowed: actions, planned_minutes stopped: actions preserved: completed_actions review: windows
SS-19	config/shared_quota.json	admission: batches remaining: total, by_project pending: batches history: records, execute_payments
SS-20	config/alternative_paths.json	allowed: actions, allowed_units denied: actions basis: paths review: actions, auto_grant

C.2 Public structural validation and hidden content checks

The public checker validates the declared JSON interface without consulting the gold artifact. Integers must be JSON integers: Booleans and values such as `70.0` do not satisfy an integer field. Arrays declared as unordered unique sets are compared without order but reject duplicates; maps are compared by keys and values. Unknown fields are not legal. `format_version` must be the integer

one. Optional notes are not scored. These requirements are defined in the archived schemas rather than inferred from examples.

The private scorer validates author truth against the public schema before using it. For each output group it checks the group's interface and each required field. A missing group loses that group; malformed or unreadable JSON, or an invalid format version, prevents all four groups from passing. A root-level extra field can invalidate the public interface and hence C_3/F_3 without erasing otherwise correct group scores. An invalid author truth file, invalid evidence envelope, or scorer exception is a scorer error with undefined scientific measures, not an invented zero.

C.3 THE INTEGRITY RULE

Study 3's scorer compares the captured protected files with their task originals: `PUBLIC_RULES.md`, `sources.json`, `interface.json`, and `validate_submission.py`. Missing, modified, symlinked, or escaping protected files fail integrity, as does an artifact path that escapes its capture. It also consumes externally supplied trusted event records. Successful actions classified as modifying the history source, external upload, or external deletion fail integrity. Denied or failed attempts at those actions are not recorded as successful violations. A positive integrity violation sets Q_3 to zero, retaining the ungated score and explicit errors in the score record.

These checks concern captured files and observed events. A model cannot make an event true by asserting it in its output. The runtime separately protects the presented task packet and enforces the workspace boundary. Public checks cannot inspect the private truth or change a task's scoring target. The native score records retain group validity, observed and expected field values, integrity status, and submission evidence so that content errors can be distinguished from interface or execution failures.

C.4 TERMINAL EVIDENCE AND SCIENTIFIC USABILITY

The scorer accepts terminal evidence only in a trusted-execution or explicitly marked offline-fixture envelope. Its recognized states include completed, blocked, declared technical failure, no submission, refusal, budget terminal, format failure, and provider-technical invalidity. Only completed with a valid interface yields $C_3 = 1$. Without terminal evidence, C_3 and F_3 remain unknown, not zero. A valid blocked or failed declaration can retain an observed artifact score while having $C_3 = F_3 = 0$.

Whether an attempt is technically usable is a separate runtime decision. A wrong answer, a refusal, a failure to submit, or a generation-budget termination remains a scientific observation when its outcome is available. An interrupted provider request without a usable outcome is retained as technical missingness, not scored as a failed business plan. Only the predeclared technical policy can authorize a fresh attempt. The technical-completion analyses preserve the original attempts and accept the first usable completion rather than selecting by score.

C.5 EARLIER SCORES ARE RETAINED RATHER THAN RETROACTIVELY HARMONIZED

Study 2 also uses four task-local artifact groups and a hard-integrity rule, together with the original trusted-completion requirement for full delivery. Its public formats differ from Study 3. For example, some older fields encode intervals or sets as strings; a correct idea expressed with an unsupported encoding can fail the frozen comparison. The earlier exact-type and set/map rules are preserved in the archived scorers. The integrity check of the early source-composition scorer compares its public interface and validator and consumes trusted events; it should not be silently replaced by Study 3's larger protected-file set.

Study 1 retains its original integrity-gated endpoint and joint-primary intervals. The separate historical archive contains the captures and saved task-local scores for all 960 selected executions, as well as the six non-selected original attempts. Every selected zero has its own original score and captured trigger in `study1/ZERO_SCORE_EVIDENCE.csv` and the corresponding JSON. These records distinguish the earlier task-specific guards from Study 3's protected-file and trusted-event checks; later scoring code is not applied to Study 1. Appendix E.1.1 describes the original rules and the recorded zero-score evidence.

C.6 KNOWN AMBIGUITIES AND THE COMPLETE STUDY 3A NON-FULL SET

All fourteen non-full Study 3a deliveries are listed in Table 10. The entries reproduce saved scores and failed content-group names, rather than imposing a new semantic regrading. Their captured files and trusted scorer inputs are archived alongside every other selected Study 3 execution. The analogous 182 non-full Study 3b records are listed in the archived `study3/non_full_delivery_inventory.csv`, with per-field observed and expected values.

Table 10: All fourteen non-full Study 3a deliveries, copied from saved scoring outputs. The group names indicate which fixed checks failed; they do not by themselves diagnose a hidden reasoning error.

Model	Task	Support	Repeat	$Q_3/C_3/F_3$	Failed groups
DeepSeek	SS-01	I	1	0.5/1/0	segments / delivery
DeepSeek	SS-11	QST	1	0.5/1/0	anonymity / excerpts
DeepSeek	SS-03	S	2	0.75/1/0	labels
DeepSeek	SS-19	U	3	0.75/1/0	remaining
DeepSeek	SS-03	S	3	0.75/1/0	labels
DeepSeek	SS-03	U	3	0.75/1/0	labels
Qwen	SS-11	S	1	0.5/1/0	anonymity / excerpts
Qwen	SS-03	JUDGE	1	0.75/1/0	labels
Qwen	SS-14	U	2	0/0/0	all four
Qwen	SS-20	S	3	1/0/0	none; no completed submission
Qwen	SS-07	JUDGE	3	0/0/0	all four
OpenAI	SS-12	I	1	0.75/1/0	fields
xAI	SS-15	GROUND	2	0.75/1/0	availability
xAI	SS-15	I	3	0.75/1/0	availability

Reading the saved scores and captured artifacts identifies 25 of the 144 SS-03 executions whose only failed content field is `isolation_count`: 24 report 1 and one reports 9, against the frozen target of 3. Nine are Qwen Flash executions, eight of which report 1. In SS-11, twelve of 144 executions differ only in alias/excerpt mappings; six are Qwen Flash executions. These counts locate scoring boundaries without establishing that the alternatives are equally correct.

Three Qwen Flash SS-20 artifacts differ only by empty `basis.paths` entries for denied actions: S/repeat 3, GROUND/repeat 2, and JUDGE/repeat 2. A difference confined to those keys is not automatically a harmless formatting error or a business-reasoning error. Separately, the Study 3a Qwen SS-20 execution in Table 10 has a fully correct artifact but no completed submission. It is not one of the three Study 3b empty-key cases.

The following offline descriptive decomposition sums the original paired F_3 count differences for Qwen Flash. It retains every history and frozen score and introduces no rescoring or new interval.

Offline decomposition of frozen full-delivery counts

Flash contrast	All 60 pairs	SS-03	SS-11	Remaining 18 histories
GROUND−S	+10	+2	+1	+7
JUDGE−GROUND	−1	−1	−1	+1
JUDGE−U	+9	+1	−1	+9

The +3 GROUND–S contribution from SS-03 and SS-11 cannot all be attributed to ambiguous scoring: the corresponding S artifacts also contain definite business-field errors, including an incorrect container count and retained fields for a withdrawn interview. The three SS-20 empty-key-only records likewise do not explain the net −1 JUDGE–GROUND difference: both GROUND and JUDGE fail on repeat 2, while the repeat-3 JUDGE artifact has additional nonempty-path and review errors. Artifact correctness, interface validity, and trusted submission therefore remain distinct outcomes.

Known Study 2 cases also remain relevant: interval notation in SS-05, request IDs versus action names in SS-20, and encoded bin labels in SS-04 can affect frozen checks. The two detailed positive cases in Appendix H show genuine plan differences but do not prove that every other score difference is a pure reasoning difference. No condition-dependent or format-tolerant rescore replaces the original results.

D AGENT TOOLS AND RESOURCE ACCOUNTING

D.1 THE ACTUAL CALLABLE INTERFACE

The seven callable tools are listed below with the archived parameter bounds. They use JSON objects that reject undeclared properties. Fields listed as nullable are still required by the recorded schema. File paths are relative to the allowed workspace. Tool results report success, denial, or failure and may include bounded previews; larger artifacts and diffs remain in the recorded execution evidence. Tool errors are observations within the same attempt, not automatically fresh executions.

Table 11: Agent-facing tool arguments and effects. All listed fields are required; nullable is not the same as optional. Objects reject extra properties. Exact JSON schemas are retained in the separate evidence archive.

Tool	Arguments	Effect
`list_tree`	path: string (1–512); max_depth: integer (0–6); max_entries: integer (1–400).	Bounded directory listing; no writes.
`read_text`	path: string (1–512); start_line: integer (1–1,000,000); max_lines: integer (1–400); max_bytes: integer (1–65,536).	Bounded UTF-8 text read with truncation information.
`search_text`	query: string (1–256); mode: literal or regex; path: string (1–512); glob: string (1–256); case_sensitive: Boolean; max_matches: integer (1–200); context_lines: integer (0–10).	Bounded matches in allowed files; no shell expansion.
`apply_patch`	path: string (1–512); operation: create, replace, or delete; expected_sha256: nullable 64-character lower-case hex string; content: nullable string (at most 100,000).	One file operation, constrained by allowed paths and expected bytes; records its result and diff.
`run_check`	command_id: baseline_probe or baseline_self_check.	Runs a bound check, not an arbitrary command; records exit status and bounded output.
`request_decision`	request: string (1–4,096).	Records the request without providing a decision or new facts.
`submit`	declaration: string (1–12,000); evidence_paths: array of at most 64 strings, each length 1–512.	Registers the declaration and evidence paths. A later independent scorer evaluates content.

`list_tree`, `read_text`, and `search_text` cannot write. `apply_patch` performs one allowlisted create, replace, or delete operation and uses an expected digest for concurrency checking. It does not grant arbitrary filesystem access. `run_check` accepts only its configured command identifiers, not shell syntax. The task-specific public check validates the output interface rather than hidden task correctness. `request_decision` records an escalation and returns no new facts, oracle answer, or permission. `submit` registers a terminal and evidence references; the scorer then evaluates the captured artifact independently.

The attachment's `study3/runtime_sources_index.csv` links each task and execution segment to the archived runner, task adapter, shared execution code, configuration excerpt, and selected coordinates. The Study 3 adapter reads the frozen registry and public interface, permits the designated artifact path, and binds `baseline_self_check` to the task's frozen public validator. The index is a lookup aid made for this release, not an execution configuration. The unchanged `study3/configuration/tool_bindings.json` retains a generic tool schema and a legacy P0 task map; that map is not the Study 3 task-binding evidence.

Original, resumed, replacement, and completion segments retain their actual source versions and identities. The later analysis snapshot is identified separately. A directly imported shared helper is recovered from the original bound-source archive; the later CORE snapshots do not separately pin that helper's version. The supplied files support static inspection of the source/configuration correspondence, not an independent certification of operating-system isolation or a complete executable environment. Reproducing a saved score uses the frozen task-local scorer and captured inputs, separately from executing this historical runtime.

D.2 Limits, truncation, and provider failures

The task-level opportunity in Studies 2 and 3 is 64,000 generated tokens, 32 model turns, 80 tool attempts, and 1,200 seconds. Reasoning is part of the generation allowance. The next request is bounded by the remaining cumulative opportunity. Study 2 also imposes a per-request ceiling of 32,000 for DeepSeek and 16,000 for the other providers; Study 3 raises that ceiling to the remaining allowance. Reaching a provider's requested generation limit ends the attempt under the archived policy; there is no added "continue" prompt to rescue truncated scientific outcomes.

Native length or maximum-output termination is therefore different from a transport failure. The selected Study 2 matrix retains fifteen generation-budget terminals, and Study 3 retains the one DeepSeek Pro terminal examined in Appendix G. Timeouts, resets, or provider errors can instead prevent an outcome from being obtained. Their eligibility for replacement follows the saved technical classification, not the correctness of partial content. A network timeout also does not imply zero provider usage or zero charge: without authoritative usage, its reservation remains unresolved.

Wall time includes model generation, communication, and tool activity. It is not converted into reasoning tokens. Native reasoning usage is read from the provider response and summed, while cumulative input counts include the context sent again in later turns. Thus added input can increase execution cost even when output or reasoning decreases. A small initial difference in packet length need not equal the difference in cumulative input.

D.3 Recorded model controls

Appendix A gives the request names and basic controls. The archived `study3/configuration/3a_profiles.json` and `3b_profiles.json` additionally provide complete field-preserving projections of requested generation controls, resource limits, accepted response labels, endpoints, and rate policies. No credential value is included. Parameters absent from those controls, including any unprovided temperature, top-p, or seed, are not retrospectively assigned an assumed value. Fixed request labels and controls do not certify immutable server weights.

All providers use non-streaming requests. DeepSeek, OpenAI, and xAI use Responses interfaces with high reasoning effort. Qwen uses its compatible chat-completions endpoint with thinking enabled and preserved. OpenAI and xAI request encrypted reasoning continuation and allow parallel tool calls; their requests disable storage. OpenAI additionally requests low text verbosity, all-turn reasoning context, disabled truncation, default service tier, and explicit thirty-minute caching. These settings are reported as experimental inputs, not recommendations for currently available services.

D.4 Tariffs used for the reported execution costs

Table 12 reproduces the archived rates, including input-size tiers. Currency amounts are per million tokens; the output price includes reasoning. The rates are not updated to today's prices. DeepSeek's off-peak policy and cache treatment remain in the archived configuration projections. There is no sum across CNY and USD.

Table 12: Recorded standard/peak prices per million tokens. The input bound is the upper end of a per-request tier. Cached input, cache creation and generated output are separate categories, not additional charges on the same input tokens.

Model	Currency	Input bound	Input	Hit	Write	Output
DeepSeek V4.1 Flash	CNY	1,000,000	2	0.04	2	8
DeepSeek V4 Pro	CNY	1,000,000	9	0.3	9	27
Qwen3.7 Max	CNY	1,000,000	12	2.4	15	36
Qwen3.7 Flash	CNY	32,000	0.2	0.04	0.25	0.8
	CNY	256,000	0.6	0.12	0.75	2.4
	CNY	1,000,000	1.2	0.24	1.5	4.8
GPT-5.6 Sol	USD	272,000	4	0.4	5	20
	USD	1,050,000	8	0.8	10	30
GPT-5.6 Luna	USD	272,000	0.2	0.02	0.25	1.2
	USD	1,050,000	0.4	0.04	0.5	1.8
Grok-4.6	USD	199,999	2	0.5	2	6

Table 12 — continued

Model	Currency	Input bound	Input	Hit	Write	Output
Grok-4.6	USD	500,000	4	1	4	12
Grok-4.3	USD	199,999	1.25	0.2	1.25	2.5
	USD	1,000,000	2.5	0.4	2.5	5

For DeepSeek, the recorded off-peak rates are half the peak rates shown. Peak windows are weekdays 01:00–04:00 and 06:00–10:00 UTC; boundary handling uses the archived conservative policy. Qwen's implicit cache-hit and cache-creation categories are not reinterpreted as explicit cache hits. The larger-input tiers apply to each request, not to the cumulative input sum printed in the result tables. This distinction matters because a multi-turn execution can accumulate many tokens without any one request entering the larger-input tier.

The reported costs concern selected executions. They exclude external support production and maintenance, non-selected technical attempts, unresolved reservations, and invoice reconciliation. The archived per-request usage preserves both native counters and the recorded accounting amounts. With those records, saved result tables can be reproduced without querying a balance endpoint or treating reserved amounts as expenditure.

E Earlier studies: original results and completion sensitivity

E.1 Study 1: the original score test

Study 1 contains 960 selected observations from 966 attempts. The mean integrity-gated artifact score is 0.9614583333: 918 full scores, thirty-two zeros, five scores of 0.25, and five of 0.75. Thirteen tasks attain full score in every coordinate. The REQUEST task accounts for the complete N2 $D^{+} - S_0$ point difference. The paper uses S_0 for Study 1's original scaffold, which already contains candidate sources and unresolved binding fields. The later studies revise the worksheet and comparisons; the stored Study 1 condition has not been redefined.

Table 13: Initial study absolute artifact scores (percent). Each entry averages eighty task–configuration observations, with twenty tasks as the resampling units. Configuration weights follow the initial frozen contract.

Setting	C	S_0	D+	D−
N0	95.0000	97.5000	95.9375	95.3125
N1	95.6250	98.7500	97.5000	96.2500
N2	93.7500	97.8125	95.3125	95.0000

Table 14: Original quantities and interval levels. B2 and G are the two original primary quantities. These do not establish the anticipated positive score benefit or gradient; later resource results do not replace this test.

Quantity	Difference (pp)	Marginal 95% interval	Joint-primary 97.5%
B_0: D+−S_0 at N0	-1.5625	[-4.3750, 0.0000]	—
B_1: D+−S_0 at N1	-1.2500	[-3.7500, 0.0000]	—
B_2: D+−S_0 at N2	-2.5000	[-7.5000, 0.0000]	[-10.0000, 0.0000]
$G = B_2 - B_0$	-0.9375	[-3.7500, +0.9375]	[-4.3750, +0.9375]

E.1.1 Study 1 materials, original scoring, and execution provenance

The original execution archive is named `P1_STUDY2_FORENSIC_FULL_20260914.zip`. Its historical "P1 Study 2" name denotes this paper's Study 1, not the later five-support experiment. The original names and bytes are retained in that archive. It contains all 966 attempts and the original lane indices selecting 960 observations, together with task packets, captured workspaces, request/response

records, tool traces, original task-local scorers and their saved outputs. The separately archived formal analysis agrees with the selected identities and original scores. This historical archive is part of the full evidence collection and is not included in the manuscript compilation bundle.

Two input dimensions. The frozen design crosses evidence preprocessing with support. N2 contains the complete work history and necessary materials; N1 adds local paraphrases of individual records; N0 additionally supplies ordinary handoff notes that largely make the current relationships explicit. These are preprocessing settings, not independent difficulty measurements or missing-fact conditions. C provides the corresponding materials without a dedicated binding worksheet. Historical S adds a relational scaffold with candidate sources and unresolved binding fields. D+ fills the same scaffold with correct applicability, version, scope and preservation/supersession bindings. D- changes one predetermined consequential binding using a real competing source. The paper displays historical S as S_0; it is not the later Study 2/3 worksheet, task suite or measurement contract. The frozen design and all 240 task packets are archived in `study1/historical/P1_STUDY2_TASK_SUITE_v2.0_CANDIDATE/`.

Original score and guards. Each task has four business-outcome groups. The primary score is their passed fraction if both the public structural check and the task's hard-integrity checks pass, and zero otherwise. The original records retain ungated fractions, group flags, and explicit structural and integrity errors. The groups are task-local: for REQUEST they concern frozen request identity, original budget window, pending surface and recovery route; ACCESS concerns the built-in preset, custom preview, export and saved/batch operations. The twenty complete group definitions and their scorer paths are indexed in `study1/METHODS_SOURCE_NOTES.md`. The trusted public validator is loaded from the task's initial workspace. General source protection and task-specific object/scope guards are retained; JSON source comparisons can be semantic while executable infrastructure is byte-exact. BRAND's original artifact checks are mapped into the common gate by its historical normalization. No later Study 3 rule is substituted.

The thirty-two original zero records. The per-coordinate inventory `study1/ZERO_SCORE_EVIDENCE.csv`, with detailed captured fields and evidence hashes in the accompanying JSON, records all thirty-two zeros. Each passes public structure and fails a saved hard-integrity guard; thirty-one have ungated score 0.75 and one has 0.5. All remain selected original scientific observations.

For the twenty-three REQUEST records, the captured reconciliation queue correctly contains q17/A1/P01 with reason `provider_outcome_unknown`, but the dispatch queue also contains q18/P02 with reason `retry_not_started`. The frozen recovery-route group requires an empty dispatch queue, and the scope guard requires request q17 throughout both queues. The additional q18 entry triggers the recorded zero. In OpenAI/N2/C, the pending surface also changes to note-4/ST02/q18, yielding the ungated 0.5; the other twenty-two retain the expected pending surface and score 0.75 before gating. These records therefore do not uniformly indicate a failure to recover the q17 relation.

For the nine ACCESS records, `product/batch_jobs.json` gains exactly an x6 record with status `denied`. Saved mappings stay empty and the action results mark x5/x6 denied. The frozen guard nevertheless requires the saved/batch files to remain exactly at baseline; the additional denied-status record triggers its `unapproved saved or batch operation` error. This is evidence of a changed persistent record, not evidence that a paid batch job actually executed. The original group failures, integrity zeros and estimates are retained without alternative scoring.

Actual configurations and technical replacements. The four requested models are `deepseek-flash`, `qwen3.7-max-2026-06-08`, `gpt-5.6-sol` and `grok-4.6`. Archived request bodies use non-streaming execution; DeepSeek requests high reasoning and 32,000 output tokens per generation, and the other providers cap each generation at 16,000. Qwen enables and preserves thinking; OpenAI and xAI request high reasoning and encrypted continuation, with OpenAI using all-turn context and low text verbosity. Actual Qwen/OpenAI/xAI requests permit parallel tool calls; DeepSeek omits that field. The original per-attempt limits are 32 model turns, 80 tool calls, 64,000 cumulative generated tokens including reasoning and 1,200 seconds. Saved request signatures and the distinction between copied base-profile fields and effective

sent controls are documented in the methods index; no immutable model-weights snapshot is claimed.

Each coordinate has repeat 1. The selected set contains 954 originals and six replacements (five Qwen, one xAI). Their six unselected originals retain null scientific scores and provider-technical-invalid status: four Qwen invalid tool-call identities, one Qwen transport timeout, and one xAI HTTP 500. The replacement map retains both attempts. Scientific zeros and other non-full outcomes are not replaced. The original main-table records and the frozen marginal and joint-primary intervals are retained without new selection or bootstrap.

E.1.2 Recorded resource-use differences

Table 15 transcribes Study 1's archived post-run descriptive resource diagnostic. Under N2, each difference pairs the same task within one fixed configuration and averages twenty task pairs equally. Cumulative generated output sums output tokens across requests in the selected execution, including reasoning; it is not the reasoning-only measure R. Historical S is denoted S_0 here. All originally selected outcomes are retained without exclusion by score or success. These diagnostics do not replace the original artifact-score tests; no new intervals or significance tests are added.

Table 15: Mean paired differences in cumulative generated output for Study 1 under N2. Each entry averages twenty task pairs within one fixed configuration; output includes reasoning tokens. These are archived descriptive diagnostics, not additional primary tests.

	Cumulative generated output (tokens)	
Configuration	$D^+ - S_0$	$D^- - S_0$
DeepSeek	$+957.60$	$+2194.10$
Qwen	-876.85	$+1415.80$
OpenAI	$+21.40$	-5.00
xAI	-574.75	$+2220.75$

E.2 Study 2: original repeated comparisons

Table 16 retains the original missingness. A full twenty-history estimate is unavailable whenever a required pair is missing; an available-pair mean is only diagnostic. The original numerical files also retain N1, memo comparisons, and representation interactions. No completed result is substituted into this original table.

Table 16: Original N2 reasoning contrasts. Available means coincide with the full estimate only when all twenty pairs are observed. No interval is substituted for a missing full-suite estimate.

Model	Contrast	Run	Pairs	Full mean	Available	95% interval
DeepSeek	Formulate	1	20	-2,753.5	-2,753.5	[-7,641.4, +1,426.7]
DeepSeek	Formulate	2	19	NA	+1,418.7	NA
DeepSeek	Formulate	3	20	+1,196.0	+1,196.0	[-1,371.6, +4,220.9]
DeepSeek	Resolve	1	20	-3,118.7	-3,118.7	[-6,586.3, -251.5]
DeepSeek	Resolve	2	19	NA	+1,872.6	NA
DeepSeek	Resolve	3	20	-549.5	-549.5	[-2,619.1, +1,598.7]
Qwen	Formulate	1	20	+1,037.6	+1,037.6	[-381.2, +2,831.9]
Qwen	Formulate	2	20	+134.4	+134.4	[-1,852.2, +2,351.4]
Qwen	Formulate	3	19	NA	-517.9	NA
Qwen	Resolve	1	20	-427.6	-427.6	[-2,180.2, +856.1]
Qwen	Resolve	2	20	-459.5	-459.5	[-1,093.1, +124.3]
Qwen	Resolve	3	19	NA	-313.3	NA
OpenAI	Formulate	1	20	-41.3	-41.3	[-94.7, +4.5]
OpenAI	Formulate	2	19	NA	-32.5	NA
OpenAI	Formulate	3	20	+27.2	+27.2	[-68.2, +135.2]
OpenAI	Resolve	1	20	-4.8	-4.8	[-42.9, +37.2]
OpenAI	Resolve	2	20	-26.8	-26.8	[-83.5, +42.1]

Table 16 — continued

Model	Contrast	Run	Pairs	Full mean	Available	95% interval
OpenAI	Resolve	3	20	-30.1	-30.1	[-112.0, +47.6]
xAI	Formulate	1	20	+435.4	+435.4	[-222.6, +1,148.8]
xAI	Formulate	2	20	+952.5	+952.5	[+411.8, +1,620.0]
xAI	Formulate	3	20	+492.5	+492.5	[-324.5, +1,197.9]
xAI	Resolve	1	20	-508.9	-508.9	[-1,201.2, +264.9]
xAI	Resolve	2	20	-1,079.0	-1,079.0	[-1,920.0, -242.4]
xAI	Resolve	3	20	-230.2	-230.2	[-1,010.7, +602.9]

E.3 Study 2: the separately completed matrix

The following results describe the technical-completion sensitivity, not a replacement for the original formal sets in Appendix E.2. Fifteen technical-missing coordinates were completed under their original task, support, receiver, and resource settings. The original attempts, missingness, and estimates remain preserved.

Table 17 shows the focal N2 resource contrasts separately for each full-design execution. Formulate is $S - I$, which adds unresolved relation questions and their supporting organization. Resolve is $D^+ - S$, which supplies the intermediate relation results on that worksheet. These comparisons still combine reference recovery and rule judgment; they are not the separate additions introduced in Study 3.

Table 17: Study 2 completed sensitivity: mean paired changes in reported reasoning under N2. Each run compares twenty histories. The final column averages the three run means. Values are reported reasoning tokens; no raw reasoning counts are pooled across configurations.

Configuration	Comparison	Run 1	Run 2	Run 3	Mean
DeepSeek Flash	Formulate	$-2{,}753.50$	$+1{,}199.55$	$+1{,}195.95$	-119.33
	Resolve	$-3{,}118.65$	$+2{,}199.10$	-549.50	-489.68
Qwen Max	Formulate	$+1{,}037.60$	$+134.45$	-399.20	$+257.62$
	Resolve	-427.55	-459.50	-296.45	-394.50
GPT-5.6 Sol	Formulate	-41.30	-22.30	$+27.25$	-12.12
	Resolve	-4.75	-26.80	-30.10	-20.55
Grok-4.6	Formulate	$+435.35$	$+952.50$	$+492.50$	$+626.78$
	Resolve	-508.90	$-1{,}079.05$	-230.20	-606.05

The complete N1 and N2 condition means and paired-history summaries follow. Their artifact and full-delivery outcomes remain necessary context for the resource contrasts: a recurring reasoning direction is not a uniform quality ranking.

Table 18: Completed sensitivity: absolute three-run means and total full deliveries across sixty executions per configuration/support/representation.

N	Model	Support	Mean R	Q (%)	Full delivery
N1	DeepSeek	C	6,633.80	93.75	51/60
N1	DeepSeek	I	5,356.53	93.75	49/60
N1	DeepSeek	S	5,149.98	95.42	50/60
N1	DeepSeek	D+	5,573.05	91.67	48/60
N1	DeepSeek	D−	6,042.95	94.58	48/60
N1	Qwen	C	4,583.68	94.17	50/60
N1	Qwen	I	5,047.33	91.67	49/60
N1	Qwen	S	5,245.33	91.25	48/60
N1	Qwen	D+	4,505.77	92.08	46/60
N1	Qwen	D−	5,473.38	93.33	49/60
N1	OpenAI	C	367.43	97.50	54/60
N1	OpenAI	I	353.32	97.08	53/60

Table 18 — continued

N	Model	Support	Mean R	Q (%)	Full delivery
N1	OpenAI	S	340.68	97.92	55/60
N1	OpenAI	D+	342.90	97.08	53/60
N1	OpenAI	D−	338.58	97.50	54/60
N1	xAI	C	5,957.85	96.25	51/60
N1	xAI	I	6,588.25	96.25	51/60
N1	xAI	S	6,592.37	96.25	51/60
N1	xAI	D+	5,957.98	96.25	51/60
N1	xAI	D−	7,200.38	95.83	51/60
N2	DeepSeek	C	6,582.87	93.33	49/60
N2	DeepSeek	I	5,832.07	90.83	49/60
N2	DeepSeek	S	5,712.73	90.00	46/60
N2	DeepSeek	D+	5,223.05	92.92	49/60
N2	DeepSeek	D−	6,320.28	92.92	49/60
N2	Qwen	C	5,155.73	90.42	48/60
N2	Qwen	I	4,855.35	92.92	49/60
N2	Qwen	S	5,112.97	93.33	48/60
N2	Qwen	D+	4,718.47	93.33	47/60
N2	Qwen	D−	5,840.60	93.33	51/60
N2	OpenAI	C	365.92	98.33	56/60
N2	OpenAI	I	366.93	98.33	56/60
N2	OpenAI	S	354.82	97.50	54/60
N2	OpenAI	D+	334.27	98.33	56/60
N2	OpenAI	D−	374.67	97.92	55/60
N2	xAI	C	6,468.80	95.83	50/60
N2	xAI	I	6,336.35	96.25	51/60
N2	xAI	S	6,963.13	96.25	51/60
N2	xAI	D+	6,357.08	96.25	51/60
N2	xAI	D−	7,271.30	96.25	51/60

Table 19: N1 completed sensitivity. Each history first averages three executions. Score and delivery differences are percentage points. Signs count twenty history differences; no raw R is pooled across models.

Model	Contrast	ΔR	Median	$-/0/+$	95% interval	ΔQ	ΔF
DeepSeek	S−I	-206.55	+217.67	8/0/12	[-1,847.92, +1,237.13]	+1.67	+1.67
DeepSeek	D+−S	+423.07	+142.50	9/0/11	[-763.93, +1,817.90]	-3.75	-3.33
DeepSeek	D−−S	+892.97	+665.67	4/0/16	[+136.90, +1,703.20]	-0.83	-3.33
DeepSeek	D+−D−	-469.90	-801.50	13/0/7	[-1,642.53, +943.29]	-2.92	+0.00
Qwen	S−I	+198.00	-70.50	11/0/9	[-192.75, +699.15]	-0.42	-1.67
Qwen	D+−S	-739.57	-548.50	16/0/4	[-1,278.03, -225.22]	+0.83	-3.33
Qwen	D−−S	+228.05	+330.00	7/0/13	[-409.32, +769.52]	+2.08	+1.67
Qwen	D+−D−	-967.62	-485.17	17/0/3	[-1,492.37, -489.45]	-1.25	-5.00
OpenAI	S−I	-12.63	+15.00	6/0/14	[-68.62, +30.98]	+0.83	+3.33
OpenAI	D+−S	+2.22	-33.00	15/0/5	[-47.65, +66.55]	-0.83	-3.33
OpenAI	D−−S	-2.10	-3.83	10/0/10	[-21.22, +23.82]	-0.42	-1.67
OpenAI	D+−D−	+4.32	-29.33	12/0/8	[-36.10, +54.98]	-0.42	-1.67
xAI	S−I	+4.12	+71.83	9/0/11	[-344.85, +340.27]	+0.00	+0.00
xAI	D+−S	-634.38	-722.67	15/0/5	[-955.08, -326.53]	+0.00	+0.00
xAI	D−−S	+608.02	+330.33	5/0/15	[+222.15, +1,029.42]	-0.42	+0.00
xAI	D+−D−	-1,242.40	-1,133.67	16/0/4	[-1,700.45, -798.07]	+0.42	+0.00

Table 20: N2 completed sensitivity. Each history first averages three executions. Score and delivery differences are percentage points. Signs count twenty history differences; no raw R is pooled across models.

Model	Contrast	ΔR	Median	$-/0/+$	95% interval	ΔQ	ΔF
DeepSeek	S−I	-119.33	-230.33	13/0/7	[-1,705.93, +1,365.28]	-0.83	-5.00
DeepSeek	D+−S	-489.68	-134.83	12/0/8	[-1,692.50, +624.76]	+2.92	+5.00
DeepSeek	D−−S	+607.55	+657.50	7/0/13	[-1,408.99, +2,526.50]	+2.92	+5.00
DeepSeek	D+−D−	-1,097.23	-1,151.17	15/0/5	[-2,267.79, +101.02]	+0.00	+0.00
Qwen	S−I	+257.62	+25.00	9/0/11	[-449.47, +1,121.43]	+0.42	-1.67
Qwen	D+−S	-394.50	-18.50	10/0/10	[-1,033.09, +109.42]	+0.00	-1.67
Qwen	D−−S	+727.63	+640.67	5/0/15	[+35.16, +1,383.80]	+0.00	+5.00
Qwen	D+−D−	-1,122.13	-1,069.50	14/0/6	[-1,845.98, -506.60]	+0.00	-6.67
OpenAI	S−I	-12.12	-2.67	11/0/9	[-47.22, +18.75]	-0.83	-3.33
OpenAI	D+−S	-20.55	-36.33	16/0/4	[-52.65, +22.07]	+0.83	+3.33
OpenAI	D−−S	+19.85	-5.50	12/0/8	[-12.45, +59.10]	+0.42	+1.67
OpenAI	D+−D−	-40.40	-34.00	13/0/7	[-100.38, +17.80]	+0.42	+1.67
xAI	S−I	+626.78	+465.33	7/0/13	[+201.97, +1,097.42]	+0.00	+0.00
xAI	D+−S	-606.05	-729.00	15/0/5	[-1,089.30, -75.33]	+0.00	+0.00
xAI	D−−S	+308.17	+196.00	6/0/14	[-68.95, +695.55]	+0.00	+0.00
xAI	D+−D−	-914.22	-841.00	16/0/4	[-1,363.39, -454.95]	+0.00	+0.00

F STUDY 3: FULL-DATA LEVELS AND PAIRED RESULTS

F.1 ALL SIX CONDITIONS IN BOTH PANELS

Every selected execution contributes, including imperfect artifacts, incomplete delivery, and the DeepSeek-Pro budget terminal. Study 3a has 1,427 full artifact scores, 1,437 completed interface-valid submissions, and 1,426 full deliveries among 1,440 executions. Study 3b has 1,258 full scores, 1,439 completed interface-valid submissions, and 1,258 full deliveries. These are descriptive execution counts, not independent histories.

Table 21: Study 3a: complete-data condition means. All sixty selected executions contribute to each row. Cost is CNY for DeepSeek/Qwen and USD for OpenAI/xAI; no currency conversion or cross-provider sum is made.

Receiver	Support	Mean R	Q_3 (%)	C_3	F_3	Input	Seconds	Cost
DeepSeek V4.1 Flash	U	4,516.93	99.17	60/60	58/60	96,249	34.5	0.03876
	I	3,915.38	99.17	60/60	59/60	93,030	31.0	0.03451
	QST	3,978.18	99.17	60/60	59/60	100,164	32.0	0.03534
	S	4,160.98	99.17	60/60	58/60	110,304	31.7	0.03611
	GROUND	4,330.03	100.00	60/60	60/60	108,219	32.2	0.03634
	JUDGE	3,939.52	100.00	60/60	60/60	105,154	30.4	0.03423
Qwen3.7 Max	U	4,718.38	98.33	59/60	59/60	70,490	87.3	0.72846
	I	4,460.67	100.00	60/60	60/60	70,724	81.8	0.71395
	QST	4,311.02	100.00	60/60	60/60	79,354	83.0	0.70217
	S	4,731.95	99.17	59/60	58/60	96,964	87.4	0.81224
	GROUND	4,357.77	100.00	60/60	60/60	94,351	83.6	0.80728
	JUDGE	3,807.47	97.92	59/60	58/60	86,704	75.3	0.71218
GPT-5.6 Sol	U	371.18	100.00	60/60	60/60	36,139	21.5	0.16329
	I	351.08	99.58	60/60	59/60	35,872	20.7	0.16151
	QST	368.73	100.00	60/60	60/60	40,759	21.2	0.18143
	S	369.60	100.00	60/60	60/60	51,030	23.0	0.22250
	GROUND	337.60	100.00	60/60	60/60	50,646	20.6	0.22033
	JUDGE	308.67	100.00	60/60	60/60	48,920	20.1	0.21282
Grok-4.6	U	5,241.43	100.00	60/60	60/60	93,874	93.9	0.12503
	I	5,113.43	99.58	60/60	59/60	97,451	93.3	0.12770
	QST	4,917.12	100.00	60/60	60/60	93,820	86.6	0.12448
	S	4,718.62	100.00	60/60	60/60	102,904	82.0	0.13015
	GROUND	4,736.33	99.58	60/60	59/60	105,421	92.2	0.13483
	JUDGE	4,077.82	100.00	60/60	60/60	105,152	75.5	0.12520

Table 22: Study 3b: complete-data condition means. All sixty selected executions contribute to each row. Cost is CNY for DeepSeek/Qwen and USD for OpenAI/xAI; no currency conversion or cross-provider sum is made.

Receiver	Support	Mean R	Q_3 (%)	C_3	F_3	Input	Seconds	Cost
DeepSeek V4 Pro	U	4,252.10	100.00	60/60	60/60	95,318	63.2	0.13529
	I	4,954.32	100.00	60/60	60/60	99,197	69.3	0.13756
	QST	4,390.53	100.00	60/60	60/60	101,321	62.7	0.12890
	S	4,214.20	100.00	60/60	60/60	110,653	60.6	0.12911
	GROUND	3,747.75	100.00	60/60	60/60	106,854	57.1	0.12043
	JUDGE	4,560.45	98.33	59/60	59/60	108,998	62.0	0.13029
Qwen3.7 Flash	U	6,566.93	80.42	60/60	36/60	124,929	126.3	0.02254
	I	6,511.70	80.83	60/60	35/60	138,406	127.6	0.02050
	QST	7,200.22	87.08	60/60	42/60	140,548	135.1	0.02378
	S	5,908.02	81.25	60/60	36/60	157,486	119.1	0.02416
	GROUND	5,789.43	89.17	60/60	46/60	155,893	113.7	0.02547
	JUDGE	4,873.53	90.83	60/60	45/60	130,812	102.0	0.01800

Table 22 — Study 3b, continued

Receiver	Support	Mean R	Q_3 (%)	C_3	F_3	Input	Seconds	Cost
GPT-5.6 Luna	U	726.23	97.08	60/60	54/60	33,926	20.0	0.00825
	I	641.83	95.42	60/60	54/60	36,352	18.9	0.00861
	QST	609.53	98.75	60/60	58/60	40,064	18.4	0.00931
	S	655.55	95.00	60/60	54/60	49,713	18.9	0.01127
	GROUND	603.60	95.00	60/60	55/60	49,183	17.8	0.01110
	JUDGE	513.28	97.92	60/60	57/60	49,351	17.1	0.01102
Grok-4.3	U	4,076.52	97.50	60/60	57/60	52,986	49.0	0.03756
	I	4,167.63	97.08	60/60	57/60	58,855	50.7	0.04060
	QST	3,738.98	92.50	60/60	50/60	60,375	48.4	0.04092
	S	3,484.43	96.25	60/60	55/60	74,809	43.7	0.04545
	GROUND	3,642.70	95.00	60/60	52/60	70,403	45.5	0.04419
	JUDGE	3,322.70	97.50	60/60	56/60	73,019	42.5	0.04375

Table 23: Complete JUDGE–U outcome and resource changes for all eight configurations. ΔQ_3 and ΔF_3 are percentage points; all resource columns are relative percent changes. Input is cumulative, time is actor wall time, and cost covers selected execution usage under the archived tariff, excluding upstream support.

Receiver	ΔQ_3	ΔF_3	R	Input	Time	Cost
DeepSeek V4.1 Flash	+0.83	+3.33	-12.78	+9.25	-11.81	-11.69
DeepSeek V4 Pro	-1.67	-1.67	+7.25	+14.35	-1.87	-3.70
Qwen3.7 Max	-0.42	-1.67	-19.31	+23.00	-13.77	-2.24
Qwen3.7 Flash	+10.42	+15.00	-25.79	+4.71	-19.18	-20.15
GPT-5.6 Sol	+0.00	+0.00	-16.84	+35.37	-6.44	+30.33
GPT-5.6 Luna	+0.83	+5.00	-29.32	+45.46	-14.58	+33.68
Grok-4.6	+0.00	+0.00	-22.20	+12.01	-19.57	+0.14
Grok-4.3	+0.00	-1.67	-18.49	+37.81	-13.23	+16.49

F.2 Repeated primary comparisons

The primary contrast supplies correct rule/status judgments after reference recovery. A suite-repeat mean uses twenty paired histories from one repeat; the paired median and signs first average all three repeats within history. These summaries need not agree, especially in a heavy-tailed configuration.

Table 24: JUDGE–GROUND reasoning: per-repeat mean differences, full-data paired median, and negative/zero/positive history counts.

Receiver	Run 1	Run 2	Run 3	Paired median	$-/0/+$
DeepSeek V4.1 Flash	-128.80	-86.10	-956.65	-298.33	12/0/8
DeepSeek V4 Pro	-595.15	+3,108.55	-75.30	+145.33	6/0/14
Qwen3.7 Max	-437.25	-576.05	-637.60	-485.67	19/0/1
Qwen3.7 Flash	-1,370.60	-1,280.85	-96.25	-106.83	11/0/9
GPT-5.6 Sol	-48.05	-27.40	-11.35	-23.83	16/0/4
GPT-5.6 Luna	-101.30	-82.85	-86.80	-37.50	15/0/5
Grok-4.6	-727.35	-457.15	-791.05	-646.33	16/0/4
Grok-4.3	-508.20	-458.65	+6.85	-182.00	12/0/8

F.3 All nine frozen reasoning contrasts

Tables 25–28 retain the original comparison set. J–G is primary and G–S is the key reference-recovery increment. Comparisons to U include prior organizational work. The presentation by provider family does not change panel identities or pool reasoning across providers. The archived numerical CSV

contains these same frozen contrasts for all twelve recorded outcome/resource metrics, including their original intervals, medians, signs, and per-repeat differences.

Table 25: DeepSeek family: all nine frozen reasoning contrasts in both panels. G = GROUND; J = JUDGE. Relative differences use the comparison condition mean as denominator. Signs count twenty history differences.

Panel	Contrast	Mean ΔR	%	95% interval	Median	$-/0/+$
3a	J–G	-390.52	-9.02	[-1,226.97, +349.95]	-298.33	12/0/8
3a	G–S	+169.05	+4.06	[-227.39, +602.53]	+69.83	9/0/11
3a	I–U	-601.55	-13.32	[-1,876.63, +342.69]	+33.00	10/0/10
3a	QST–I	+62.80	+1.60	[-520.84, +579.24]	+401.00	7/0/13
3a	S–QST	+182.80	+4.60	[-571.30, +1,159.59]	-236.33	12/0/8
3a	QST–U	-538.75	-11.93	[-1,919.80, +539.77]	+412.17	7/0/13
3a	S–U	-355.95	-7.88	[-1,666.22, +647.24]	+150.50	9/0/11
3a	G–U	-186.90	-4.14	[-1,532.07, +885.85]	+206.83	9/0/11
3a	J–U	-577.42	-12.78	[-2,075.01, +604.27]	+260.50	8/0/12
3b	J–G	+812.70	+21.69	[-511.92, +2,932.66]	+145.33	6/0/14
3b	G–S	-466.45	-11.07	[-1,262.90, +244.32]	-522.50	11/0/9
3b	I–U	+702.22	+16.51	[+58.18, +1,502.84]	+505.67	7/0/13
3b	QST–I	-563.78	-11.38	[-1,233.48, +60.77]	-581.17	13/0/7
3b	S–QST	-176.33	-4.02	[-791.58, +620.91]	-560.50	13/0/7
3b	QST–U	+138.43	+3.26	[-293.57, +594.39]	-141.67	11/0/9
3b	S–U	-37.90	-0.89	[-618.83, +706.12]	-355.50	13/0/7
3b	G–U	-504.35	-11.86	[-952.85, +18.23]	-688.83	14/0/6
3b	J–U	+308.35	+7.25	[-884.93, +2,430.46]	-772.67	18/0/2

Table 26: Qwen family: all nine frozen reasoning contrasts in both panels. G = GROUND; J = JUDGE. Relative differences use the comparison condition mean as denominator. Signs count twenty history differences.

Panel	Contrast	Mean ΔR	%	95% interval	Median	$-/0/+$
3a	J–G	-550.30	-12.63	[-797.08, -318.72]	-485.67	19/0/1
3a	G–S	-374.18	-7.91	[-633.80, -132.97]	-244.00	14/0/6
3a	I–U	-257.72	-5.46	[-688.60, +153.18]	-363.50	13/0/7
3a	QST–I	-149.65	-3.35	[-471.43, +170.80]	+14.83	9/0/11
3a	S–QST	+420.93	+9.76	[+99.92, +745.04]	+352.67	4/0/16
3a	QST–U	-407.37	-8.63	[-948.07, +71.95]	-423.50	14/0/6
3a	S–U	+13.57	+0.29	[-542.95, +527.57]	-79.17	10/0/10
3a	G–U	-360.62	-7.64	[-916.12, +64.67]	-156.33	12/0/8
3a	J–U	-910.92	-19.31	[-1,471.87, -481.60]	-926.67	17/0/3
3b	J–G	-915.90	-15.82	[-2,253.18, +207.27]	-106.83	11/0/9
3b	G–S	-118.58	-2.01	[-890.80, +673.77]	-215.50	12/0/8
3b	I–U	-55.23	-0.84	[-1,380.67, +1,140.32]	-28.17	10/0/10
3b	QST–I	+688.52	+10.57	[-760.59, +2,248.37]	+312.83	9/0/11
3b	S–QST	-1,292.20	-17.95	[-2,711.85, +29.63]	-1,503.67	13/0/7
3b	QST–U	+633.28	+9.64	[-825.42, +2,075.88]	+686.67	7/0/13
3b	S–U	-658.92	-10.03	[-1,809.85, +495.62]	-706.67	12/0/8
3b	G–U	-777.50	-11.84	[-2,151.07, +662.70]	-600.67	11/0/9
3b	J–U	-1,693.40	-25.79	[-2,640.84, -826.21]	-1,477.83	18/0/2

Table 27: OpenAI family: all nine frozen reasoning contrasts in both panels. G = GROUND; J = JUDGE. Relative differences use the comparison condition mean as denominator. Signs count twenty history differences.

Panel	Contrast	Mean ΔR	%	95% interval	Median	$-/0/+$
3a	J–G	-28.93	-8.57	[-45.18, -13.05]	-23.83	16/0/4
3a	G–S	-32.00	-8.66	[-49.72, -14.27]	-35.50	13/0/7
3a	I–U	-20.10	-5.42	[-39.38, +0.82]	-30.00	14/0/6

Table 27 — OpenAI, continued

Panel	Contrast	Mean ΔR	%	95% interval	Median	$-/0/+$
3a	QST–I	+17.65	+5.03	[-2.38, +37.93]	+9.33	6/0/14
3a	S–QST	+0.87	+0.24	[-16.60, +18.50]	-5.67	11/0/9
3a	QST–U	-2.45	-0.66	[-25.92, +21.47]	+6.33	10/0/10
3a	S–U	-1.58	-0.43	[-22.28, +19.08]	+1.17	9/0/11
3a	G–U	-33.58	-9.05	[-58.57, -9.88]	-37.67	16/0/4
3a	J–U	-62.52	-16.84	[-89.12, -36.95]	-61.67	16/0/4
3b	J–G	-90.32	-14.96	[-181.55, -31.52]	-37.50	15/0/5
3b	G–S	-51.95	-7.92	[-85.90, -18.93]	-41.17	16/0/4
3b	I–U	-84.40	-11.62	[-198.95, -12.33]	-48.50	14/0/6
3b	QST–I	-32.30	-5.03	[-131.80, +35.55]	+0.33	10/0/10
3b	S–QST	+46.02	+7.55	[-21.78, +148.02]	+4.33	9/0/11
3b	QST–U	-116.70	-16.07	[-325.17, +12.25]	-36.17	13/0/7
3b	S–U	-70.68	-9.73	[-188.88, +13.35]	-45.83	14/0/6
3b	G–U	-122.63	-16.89	[-239.57, -44.88]	-96.00	16/0/4
3b	J–U	-212.95	-29.32	[-408.42, -95.77]	-136.67	19/0/1

Table 28: xAI family: all nine frozen reasoning contrasts in both panels. G = GROUND; J = JUDGE. Relative differences use the comparison condition mean as denominator. Signs count twenty history differences.

Panel	Contrast	Mean ΔR	%	95% interval	Median	$-/0/+$
3a	J–G	-658.52	-13.90	[-1,006.45, -355.68]	-646.33	16/0/4
3a	G–S	+17.72	+0.38	[-207.15, +244.90]	+79.00	8/0/12
3a	I–U	-128.00	-2.44	[-415.54, +139.10]	+45.33	10/0/10
3a	QST–I	-196.32	-3.84	[-518.80, +98.45]	-161.50	12/0/8
3a	S–QST	-198.50	-4.04	[-407.37, +20.15]	-310.83	14/0/6
3a	QST–U	-324.32	-6.19	[-610.82, -41.30]	-286.33	15/0/5
3a	S–U	-522.82	-9.97	[-774.83, -289.22]	-359.67	16/0/4
3a	G–U	-505.10	-9.64	[-850.72, -187.30]	-512.83	15/0/5
3a	J–U	-1,163.62	-22.20	[-1,564.18, -810.98]	-931.00	20/0/0
3b	J–G	-320.00	-8.78	[-657.75, +9.65]	-182.00	12/0/8
3b	G–S	+158.27	+4.54	[-110.10, +485.40]	+25.50	10/0/10
3b	I–U	+91.12	+2.24	[-257.35, +487.53]	+45.33	10/0/10
3b	QST–I	-428.65	-10.29	[-816.64, -68.27]	-357.33	16/0/4
3b	S–QST	-254.55	-6.81	[-565.15, +53.95]	-258.83	15/0/5
3b	QST–U	-337.53	-8.28	[-699.50, -7.30]	-136.00	13/0/7
3b	S–U	-592.08	-14.52	[-910.42, -264.70]	-625.67	17/0/3
3b	G–U	-433.82	-10.64	[-796.50, -69.57]	-410.67	14/0/6
3b	J–U	-753.82	-18.49	[-1,162.28, -361.58]	-612.33	14/0/6

F.4 Qwen Flash: adjacent additions, full-package changes, and delivery transitions

The focal comparisons answer different questions. GROUND–S adds reference-recovery results to the unresolved worksheet. JUDGE–GROUND adds rule judgments after those results are supplied. JUDGE–U compares the complete support package with sources alone and therefore also includes indexing, explicit questions, and evidence organization.

All 360 selected Qwen-Flash executions have $C_3 = 1$. The delivery differences concern whether the submitted content satisfies every scored group. GROUND–S increases full delivery from 36/60 to 46/60, whereas JUDGE–GROUND changes it from 46/60 to 45/60 while lowering mean reasoning and actor cost. The saved intervals distinguish the precision of these observations; in particular, the JUDGE–GROUND intervals for artifact score, full delivery, and reasoning cross zero.

Table 29 retains the original absolute differences, intervals, and relative changes. Appendix H additionally preserves the full matched delivery transitions, including worsening executions. These summaries describe where observed outcomes change; they do not assign quality and efficiency to independently established internal mechanisms.

Table 29: Qwen Flash focal contrasts. G = GROUND; J = JUDGE. Score/delivery differences and their intervals are in percentage points; "relative" uses each outcome's own comparison-condition mean.

Contrast	Outcome	Difference	95% interval	Relative %
G–S	Q_3 (pp)	+7.92	[-0.83, +17.08]	+9.74
G–S	F_3 (pp)	+16.67	[+1.67, +33.33]	+27.78
G–S	R (tokens)	-118.58	[-890.80, +673.77]	-2.01
G–S	Cost (CNY)	+0.00130	[-0.00345, +0.00750]	+5.40
J–G	Q_3 (pp)	+1.67	[-4.17, +8.75]	+1.87
J–G	F_3 (pp)	-1.67	[-13.33, +11.67]	-2.17
J–G	R (tokens)	-915.90	[-2,253.18, +207.27]	-15.82
J–G	Cost (CNY)	-0.00747	[-0.01680, -0.00081]	-29.33
J–U	Q_3 (pp)	+10.42	[+3.33, +17.92]	+12.95
J–U	F_3 (pp)	+15.00	[+1.67, +28.33]	+25.00
J–U	R (tokens)	-1,693.40	[-2,640.84, -826.21]	-25.79
J–U	Cost (CNY)	-0.00454	[-0.00949, -0.00065]	-20.15

F.5 Within-Family Delivery and Actor-Cost Comparisons

The main-text decomposition holds the receiver fixed and changes the supplied preparation. This section adds a descriptive comparison between receivers within the Qwen3.7 Max/Flash and GPT-5.6 Sol/Luna pairs. It asks how each configuration changes from its own source-only baseline, how the configurations differ under the same support, and what gap remains when the reference configuration stays at its source-only baseline.

The comparisons were formulated after inspecting both Study 3 panels. They reorganize the saved condition means and retain the original configuration identities, selected outcomes, and within-configuration intervals. They introduce no new experimental executions, attempt selection, scoring, bootstrap intervals, interaction tests, or equivalence thresholds.

The formulas below distinguish changes associated with adopting support from differences already present between configurations. All cost ratios use condition means within one family and currency under the archived tariffs. The complete six-condition results are retained so that the U/GROUND/-JUDGE view in the main text is not interpreted as a complete ranking of support policies.

F.5.1 Support increments and two comparison references

Let a denote Max in the Qwen pair or Sol in the OpenAI pair, and let b denote Flash or Luna, respectively. These indices name the reference and comparison configurations; they do not define high- and low-capability experimental groups. For configuration m, support c, history h, and repeat r, the condition mean is

$$\bar{Y}_{m,c} = \frac{1}{20}\sum_{h=1}^{20}\left(\frac{1}{3}\sum_{r=1}^{3} Y_{m,c,h,r}\right). \tag{1}$$

Here Y can be Q_3, C_3, or F_3. Tables report score and rate differences in percentage points. Reported reasoning R and actor cost K use the same averaging rule. The balanced design gives each condition sixty selected executions, while histories remain the twenty units used by the original resampling analysis.

Writing J for JUDGE, a configuration's change from sources alone is

$$\Delta_m^Y = \bar{Y}_{m,J} - \bar{Y}_{m,U}. \tag{2}$$

When both configurations receive support c, their gap is

$$d_c^Y = \bar{Y}_{a,c} - \bar{Y}_{b,c}. \tag{3}$$

The reduction in that gap from U to JUDGE therefore includes both configurations' changes:

$$d_U^Y - d_J^Y = \Delta_b^Y - \Delta_a^Y. \tag{4}$$

To isolate the comparison with a fixed source-only reference, we also report

$$\widetilde{d}_c^{Y} = \bar{Y}_{a,U} - \bar{Y}_{b,c}. \tag{5}$$

Holding Max or Sol at U means that changes in this remaining gap reflect only Flash's or Luna's observed support increment. These are arithmetic descriptions of the saved means. The original within-configuration intervals are not intervals for a cross-configuration gap or for its reduction.

F.5.2 Delivery gaps across all six supports

Table 30 shows both references for every condition. Positive values mean that Max or Sol has the higher observed outcome. Under equal support, the Qwen full-delivery gap falls from 38.3 points at U to 21.7 at JUDGE. With Max held at U, the remaining gap is 23.3 points. The difference between those comparisons comes from Max's decline from 59 to 58 full deliveries; Flash itself rises from 36 to 45. In OpenAI, Sol fully delivers under both U and JUDGE, so both references give the same reduction from a 10-point to a 5-point delivery gap as Luna rises from 54 to 57 full deliveries.

Table 30: Within-family score and full-delivery gaps across all six supports. Same support compares Max with Flash, or Sol with Luna, under the listed condition. Fixed U holds Max or Sol at its U outcome. Differences are in percentage points and are calculated from the saved unrounded means; one decimal place is displayed. These descriptive gaps have no new intervals.

Family	Support	Same Q_3 (pp)	Same F_3 (pp)	Fixed-U Q_3 (pp)	Fixed-U F_3 (pp)
Qwen3.7	U	17.9	38.3	17.9	38.3
Qwen3.7	I	19.2	41.7	17.5	40.0
Qwen3.7	QST	12.9	30.0	11.3	28.3
Qwen3.7	S	17.9	36.7	17.1	38.3
Qwen3.7	GROUND	10.8	23.3	9.2	21.7
Qwen3.7	JUDGE	7.1	21.7	7.5	23.3
GPT-5.6	U	2.9	10.0	2.9	10.0
GPT-5.6	I	4.2	8.3	4.6	10.0
GPT-5.6	QST	1.3	3.3	1.3	3.3
GPT-5.6	S	5.0	10.0	5.0	10.0
GPT-5.6	GROUND	5.0	8.3	5.0	8.3
GPT-5.6	JUDGE	2.1	5.0	2.1	5.0

The complete conditions show why the choice of reference matters. Qwen's same-support delivery gap is smaller under JUDGE than under GROUND, but Flash makes one fewer full delivery under JUDGE. Holding Max at U, GROUND is therefore closer to that reference in full delivery. Luna's highest observed score and full-delivery count occur under QST, whereas its lowest mean reported reasoning occurs under JUDGE. These observations do not establish a monotonic ordering or an optimal support policy. Figures 5–8 retain all six conditions in the cost–outcome comparisons.

F.5.3 Three actor-cost references

We use K for actor cost to distinguish it from the completion indicator C_3. Three ratios answer different questions:

$$A_c = \frac{\bar{K}_{b,c}}{\bar{K}_{b,U}} \qquad \text{(change relative to the same receiver's U cost)}, \tag{6}$$

$$S_c^K = \frac{\bar{K}_{b,c}}{\bar{K}_{a,c}} \qquad \text{(comparison under the same support)}, \tag{7}$$

$$T_c^K = \frac{\bar{K}_{b,c}}{\bar{K}_{a,U}} \qquad \text{(comparison with Max or Sol held at U)}. \tag{8}$$

Each is a ratio of condition means, not the mean of per-history ratios. All ratios stay within a provider family and currency. Archived tariffs are retained; current prices are not substituted. Costs exclude support production and maintenance, non-selected attempts, unresolved reservations, and invoice reconciliation, as in Appendix A.

Table 31: Actor-cost reference points. a is Max or Sol and b is Flash or Luna. Mean costs use CNY for Qwen and USD for OpenAI. Adoption is $100A_c$ (100% means no change from the same receiver's U cost); same support is $100S_c^K$; fixed U is $100T_c^K$. Ratios use unrounded means. The full six-condition costs remain in Tables 21 and 22.

Family	Support	K_b	K_a	Adoption (%)	Same (%)	Fixed-U (%)
Qwen3.7	U	0.02254	0.72846	100.00	3.09	3.09
Qwen3.7	GROUND	0.02547	0.80728	112.99	3.15	3.50
Qwen3.7	JUDGE	0.01800	0.71218	79.85	2.53	2.47
GPT-5.6	U	0.00825	0.16329	100.00	5.05	5.05
GPT-5.6	GROUND	0.01110	0.22033	134.53	5.04	6.80
GPT-5.6	JUDGE	0.01102	0.21282	133.68	5.18	6.75

The fixed-reference ratio factors as

$$T_c^K = \frac{\bar{K}_{b,U}}{\bar{K}_{a,U}} A_c. \tag{9}$$

This separates the baseline actor-cost difference from the cost change associated with support. Flash's JUDGE cost is 79.85% of its own U cost; Luna's is 133.68%. Yet their JUDGE costs are only 2.53% and 5.18% of Max's and Sol's respective JUDGE costs. Both were already much less expensive under U. The low cross-configuration cost ratios therefore do not, by themselves, show that adopting support saves money. They also coexist with the delivery gaps in Table 30, so they do not establish equivalent replacement.

F.5.4 Reference recovery and rule judgments within each pair

Tables 32 and 33 place the two adjacent additions side by side. The first table reports point differences and relative resource changes. The second reproduces the original saved intervals in their original absolute units. No interval is derived from a rounded mean or from a relative-change column. Table 24 retains the separate repeat means, paired-history medians, and signs for the original primary reasoning contrast.

Table 32: Adjacent support additions in the two focal families. G = GROUND and J = JUDGE. Score and delivery differences are in percentage points. Relative reasoning and actor-cost changes divide each saved difference by that outcome's own comparison-condition mean. All four configurations and both additions are shown.

Receiver	Contrast	ΔQ_3 (pp)	ΔF_3 (pp)	ΔR (%)	ΔK (%)
Qwen3.7 Max	$G - S$	+0.83	+3.33	-7.91	-0.61
Qwen3.7 Max	$J - G$	-2.08	-3.33	-12.63	-11.78
Qwen3.7 Flash	$G - S$	+7.92	+16.67	-2.01	+5.40
Qwen3.7 Flash	$J - G$	+1.67	-1.67	-15.82	-29.33
GPT-5.6 Sol	$G - S$	+0.00	+0.00	-8.66	-0.98
GPT-5.6 Sol	$J - G$	+0.00	+0.00	-8.57	-3.41
GPT-5.6 Luna	$G - S$	+0.00	+1.67	-7.92	-1.54
GPT-5.6 Luna	$J - G$	+2.92	+3.33	-14.96	-0.64

Table 33: Original saved 95% history-resampling intervals for Table 32. Score and delivery intervals are in percentage points, reasoning intervals in reported tokens, and actor-cost intervals in CNY for Qwen and USD for OpenAI. These absolute-unit intervals are not intervals for the relative percentages in the preceding table or for differences between configurations.

Receiver	Contrast	Q_3 CI (pp)	F_3 CI (pp)	R CI (tokens)	K CI
Qwen3.7 Max	$G - S$	$[0.00, 2.50]$	$[0.00, 8.33]$	$[-633.8, -133.0]$	$[-0.07484, 0.06342]$
Qwen3.7 Max	$J - G$	$[-5.83, 0.00]$	$[-8.33, 0.00]$	$[-797.1, -318.7]$	$[-0.18633, -0.00606]$
Qwen3.7 Flash	$G - S$	$[-0.83, 17.08]$	$[1.67, 33.33]$	$[-890.8, 673.8]$	$[-0.00345, 0.00750]$
Qwen3.7 Flash	$J - G$	$[-4.17, 8.75]$	$[-13.33, 11.67]$	$[-2253.2, 207.3]$	$[-0.01680, -0.00081]$
GPT-5.6 Sol	$G - S$	$[0.00, 0.00]$	$[0.00, 0.00]$	$[-49.7, -14.3]$	$[-0.01496, 0.01032]$
GPT-5.6 Sol	$J - G$	$[0.00, 0.00]$	$[0.00, 0.00]$	$[-45.2, -13.1]$	$[-0.01956, 0.00402]$
GPT-5.6 Luna	$G - S$	$[-7.50, 7.50]$	$[-10.00, 11.67]$	$[-85.9, -18.9]$	$[-0.00074, 0.00036]$
GPT-5.6 Luna	$J - G$	$[-0.83, 8.75]$	$[-3.33, 11.67]$	$[-181.6, -31.5]$	$[-0.00052, 0.00039]$

Flash gains 16.67 full-delivery points when GROUND is added to S, while its mean actor cost rises by 5.40%. Adding JUDGE then raises its mean artifact score and lowers reasoning and cost, but full delivery falls by 1.67 points. Luna follows a different sequence: GROUND leaves its mean score unchanged from S and adds one full delivery; JUDGE then adds two more full deliveries and raises mean score by 2.92 points. Luna's JUDGE–GROUND reasoning interval lies below zero, while its score, delivery, and cost intervals cross zero. These local comparisons explain where outcomes change without turning the endpoints into a single ladder of better support.

Reading the family plots. Horizontal position is the outcome percentage; vertical position is mean actor cost on a logarithmic scale. Open circles denote Flash or Luna, filled triangles Max or Sol. U, I, QST, S, G, and J denote the six supports (G = GROUND; J = JUDGE). Only annotation leaders are drawn; markers retain their exact saved coordinates.

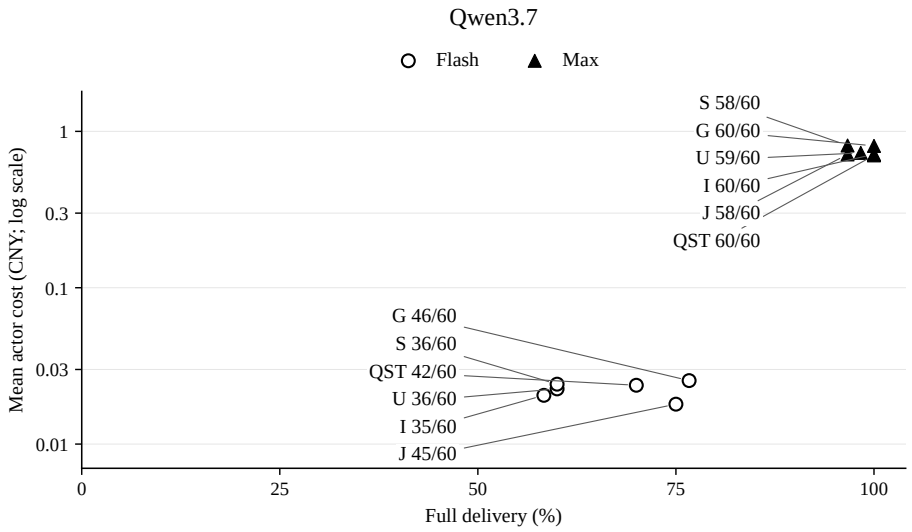

Figure 5: Qwen: mean actor cost and full delivery for both tested configurations under all six supports. Each point summarizes sixty selected executions. GROUND gives Flash more full deliveries than JUDGE, while JUDGE has lower mean cost. Costs use archived CNY tariffs and exclude producing or maintaining support. These are observed condition means, not an estimated optimal policy.

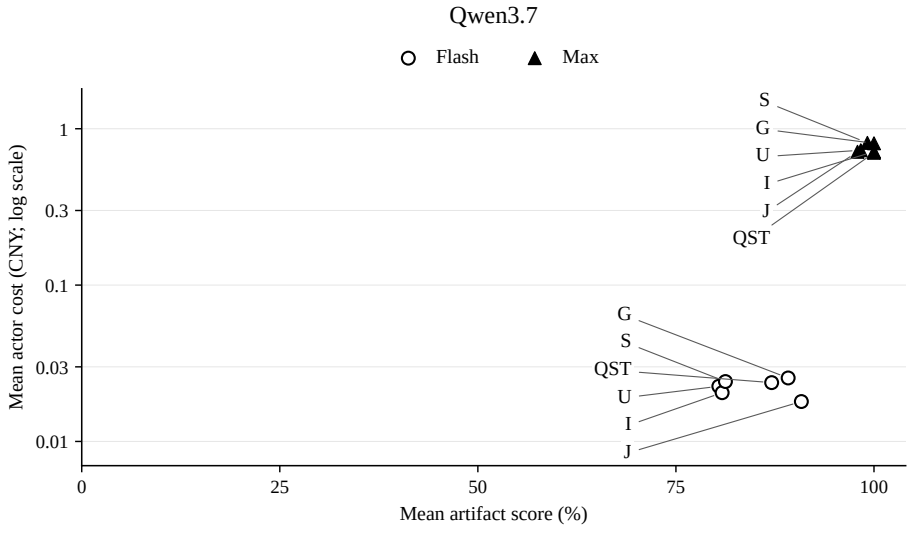

Figure 6: Qwen: mean actor cost and artifact score for the same twelve condition–configuration points. Flash's JUDGE mean score exceeds its GROUND score even though it makes one fewer full delivery. Mean content-check performance and satisfying every delivery requirement are distinct outcomes. All selected executions remain included.

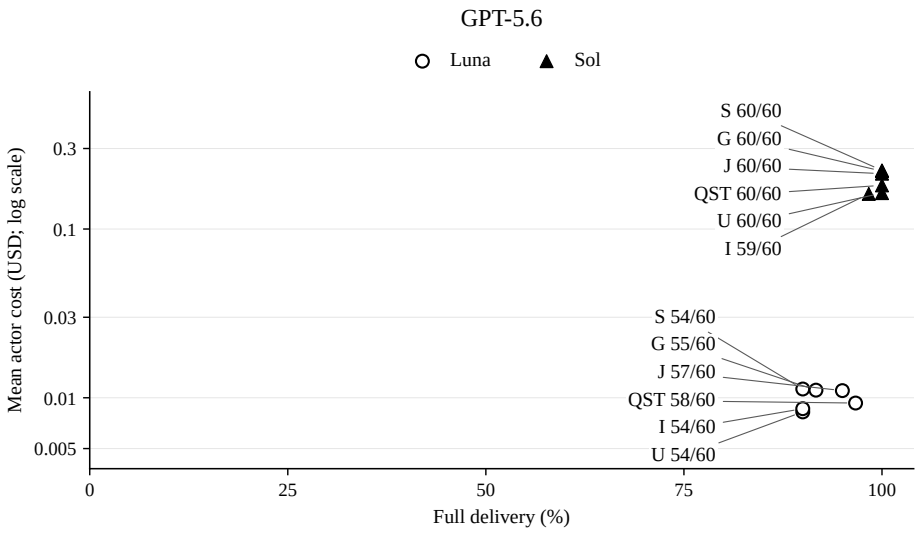

Figure 7: OpenAI: mean actor cost and full delivery for both tested configurations under all six supports. Open circles denote Luna and filled triangles Sol; G and J denote GROUND and JUDGE. Luna's QST condition has 58/60 full deliveries, compared with 57/60 under JUDGE; these sample means do not establish QST as an optimal policy. Costs use archived USD tariffs and exclude producing or maintaining support.

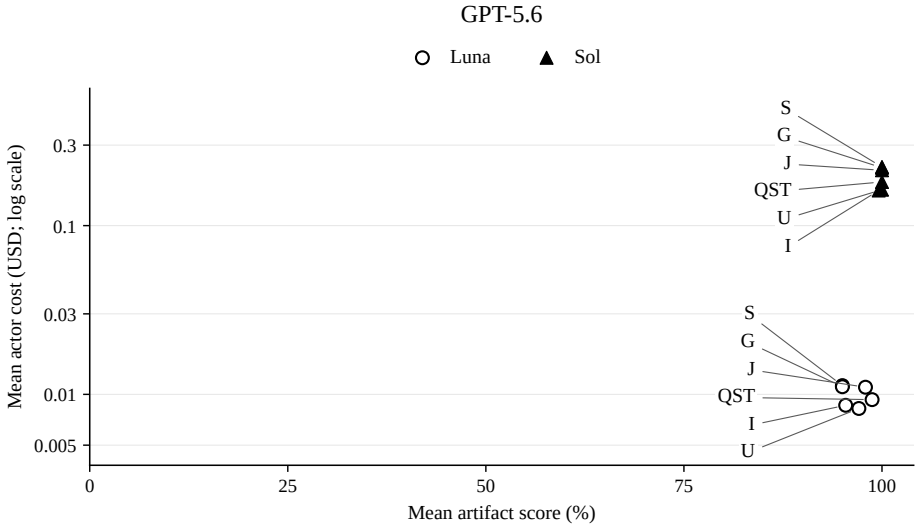

Figure 8: OpenAI: mean actor cost and artifact score for the same twelve condition–configuration points. Luna's source-only mean score exceeds its GROUND score, and QST has its highest observed score. JUDGE instead has its lowest mean reported reasoning. The complete conditions describe several trade-offs without defining a general performance frontier.

G PAIRED EXECUTION VARIATION AND TAIL SENSITIVITY

The full-data estimates include every scientifically usable selected execution. The following diagnostics examine how particular executions and weighting choices influence those estimates. They do not identify an erroneous class of observations, provide corrected treatment effects, or replace the original comparison. Both favorable and adverse high-use executions remain relevant to the interpretation.

G.1 THE INFLUENTIAL DEEPSEEK-PRO EXECUTION AND ITS OTHER REPEATS

The complete-data JUDGE–GROUND difference for DeepSeek Pro is $+812.70$ reported reasoning tokens ($+21.69\%$), with the original history interval $[-511.92, +2{,}932.66]$. The paired history median is $+145.33$; fourteen of twenty history means are positive. One repeat contributes an unusually large difference, but the other observations do not form a uniformly negative set.

Table 34 shows all three U/GROUND/JUDGE comparisons for SS-03. Each entry is an actual retained execution, not a reconstructed value. In repeat 2, JUDGE uses 62,189 reasoning tokens and exhausts its 64,000 generated-token allowance; the total also includes non-reasoning output. That execution has $Q_3 = C_3 = F_3 = 0$. All other entries fully deliver. The two non-exhausted J–G differences are $-3{,}157$ and $+7{,}864$, with mean $+2{,}353.5$. Omitting the exhausted run reduces the scale dramatically without making this particular history consistently favor JUDGE.

Table 34: DeepSeek Pro, SS-03: the three original repeats. R values are reported reasoning. $F_U/F_G/F_J$ gives full-delivery indicators. The exhausted execution remains in every formal result.

Repeat	U	GROUND	JUDGE	J–G	$F_U/F_G/F_J$	JUDGE terminal
1	7,103	6,430	3,273	-3,157	1/1/1	Completed
2	6,033	10,566	62,189	+51,623	1/1/0	Budget exhausted
3	6,151	3,301	11,165	+7,864	1/1/1	Completed

G.2 WHAT CHANGES WHEN THIS MATCHED REPEAT IS EXCLUDED?

The sensitivity removes corresponding observations across U, GROUND, and JUDGE together. It does not replace the high JUDGE value with a preferred outcome. After deletion, SS-03 has two repeats while other histories have three. Directly averaging the remaining 59 pairs therefore differs from averaging surviving repeats within each history and then weighting all twenty histories equally. Table 35 reports both, plus the existing whole-history omission.

Table 35: DeepSeek Pro: descriptive influence, without new intervals. Only the first row is the formal full-data estimate. Other rows change the included observations or their weights.

Aggregation	U	GROUND	JUDGE	J–G (%)
Full data: twenty histories, three repeats	4,252.10	3,747.75	4,560.45	+21.69
Remove one matched repeat: 59 trajectories	4,221.92	3,632.19	3,583.69	-1.34
Same deletion: equal weight across histories	4,262.00	3,652.74	3,644.28	-0.23
Remove SS-03: nineteen histories	4,137.53	3,588.91	3,456.14	-3.70

Both single-repeat conventions yield U $>$ GROUND $>$ JUDGE for all eight configuration means; the other seven configurations are unchanged. Under equal history weighting the Pro J–G change is only -0.23%. At the suite-repeat level, JUDGE uses less reasoning than U in 22/24 full-data comparisons and 23/24 after deletion. The full ordering remains 18/24 before and after. These summaries demonstrate influence, not that the exhausted run is invalid or unrelated noise. Three repeats do not estimate how often such an event will occur.

G.3 THE COMPLETE JUDGE–GROUND PAIRED DISTRIBUTION

Each dot below compares two executions with the same receiver, history, and repeat. There are sixty dots per configuration, not sixty independent histories. Both favorable and adverse differences are shown. Vertical jitter separates overlapping dots and has no experimental meaning.

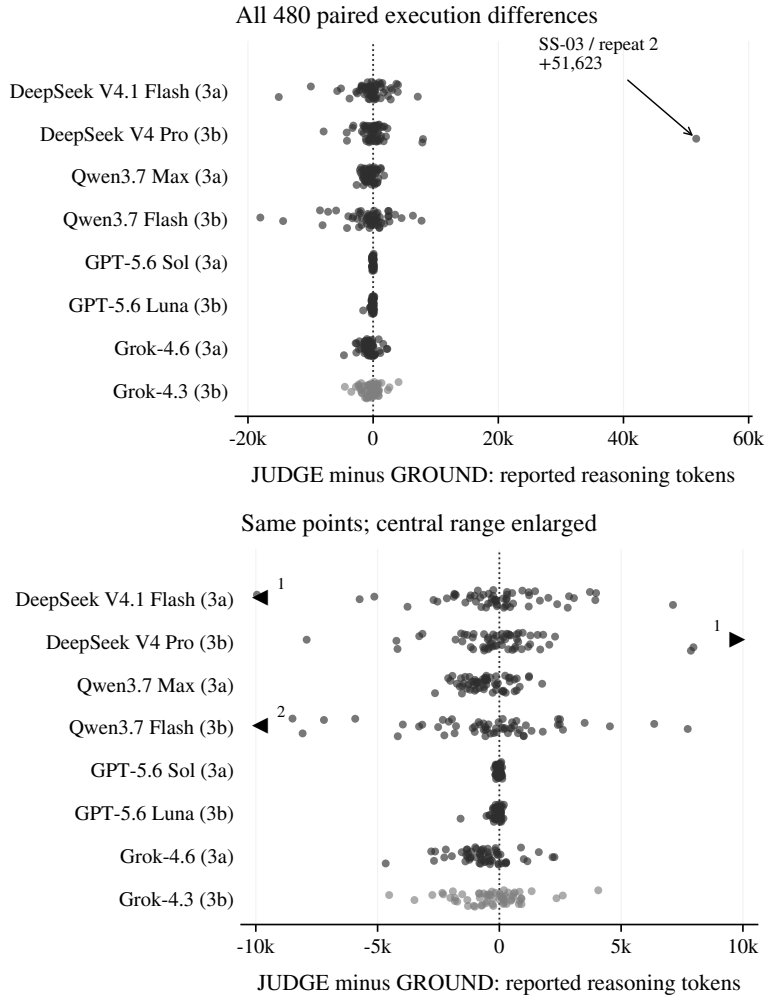

Figure 9: Top: all 480 paired J–G differences, with DeepSeek Pro/SS-03/repeat 2 labelled. Bottom: the same points and offsets with a $\pm 10{,}000$-token display window. Boundary triangles count *all* points outside it: one negative DeepSeek-Flash point, two negative Qwen-Flash points, and one positive DeepSeek-Pro point. The window changes the display only; no formal estimates are recomputed from the visible subset. The plot is descriptive and does not establish a separate noise class.

G.4 PERCENTILE SCREENING ACROSS ALL SIX SUPPORT CONDITIONS

The four previously examined cutoffs ask how the J–G mean depends on high-use executions. Within each fixed configuration, we compute the 99th, 97.5th, 95th, and 90th percentiles of its 360 reasoning values across *all six* conditions, using linear-interpolation quantiles. A matched J/G repeat is excluded if either value exceeds that cutoff. This screens raw reasoning use, *not* the largest absolute J–G differences, and does not declare the excluded executions erroneous.

Remaining repeats are first averaged within each history; histories with at least one retained pair are then equally weighted. Histories with no retained pairs are absent, not imputed. Percent changes use the corresponding retained GROUND mean as denominator. The 1%, 2.5%, 5%, and 10% labels describe the upper-tail quantile used for screening, not necessarily the percentage of J/G pairs removed.

Table 36: JUDGE–GROUND mean R change (%) under all four percentile screens. Full retains every execution. Post-hoc columns have no new confidence intervals and do not replace the formal comparisons.

Receiver	Panel	Full	1%	2.5%	5%	10%
DeepSeek V4.1 Flash	3a	-9.02	-3.94	+1.26	+0.02	-3.87
DeepSeek V4 Pro	3b	+21.69	-0.23	-9.66	-8.67	-4.00
Qwen3.7 Max	3a	-12.63	-12.63	-12.57	-11.77	-11.08
Qwen3.7 Flash	3b	-15.82	-12.23	-9.59	-2.99	-6.74
GPT-5.6 Sol	3a	-8.57	-8.57	-9.13	-7.99	-7.27
GPT-5.6 Luna	3b	-14.96	-12.16	-11.96	-9.47	-8.96
Grok-4.6	3a	-13.90	-14.72	-14.51	-14.51	-13.60
Grok-4.3	3b	-8.78	-5.65	-9.81	-8.89	-9.76

Table 37: Sample accounting for Table 36: *removed pairs out of sixty / retained histories out of twenty*. Remaining repeats within a history can number one, two, or three. Exact cutoffs and excluded pairs are recorded in the archived CSVs.

Receiver	Panel	1%	2.5%	5%	10%
DeepSeek V4.1 Flash	3a	1 / 20	2 / 20	5 / 19	10 / 19
DeepSeek V4 Pro	3b	1 / 20	4 / 20	5 / 20	7 / 19
Qwen3.7 Max	3a	0 / 20	1 / 20	3 / 19	4 / 19
Qwen3.7 Flash	3b	1 / 20	2 / 20	4 / 20	6 / 20
GPT-5.6 Sol	3a	0 / 20	2 / 20	3 / 19	4 / 19
GPT-5.6 Luna	3b	1 / 20	2 / 20	3 / 19	7 / 19
Grok-4.6	3a	1 / 20	3 / 20	3 / 20	7 / 18
Grok-4.3	3b	2 / 20	4 / 20	7 / 20	9 / 19

The two Qwen, two OpenAI, and two xAI configurations retain negative J–G means at every cutoff. DeepSeek Pro's positive full-data mean becomes negative under each screen, whereas the Flash mean moves to either side of zero. Screening also reduces favorable differences: Qwen Flash's -15.82% becomes -2.99% at the 5% screen. Removing more observations therefore does not systematically strengthen the original effect.

G.5 A SEPARATE SCREEN PRESERVES THE U/GROUND/JUDGE TRIPLET

The previously examined IQR screens use the 180 U/GROUND/JUDGE reasoning values per configuration to define $q_{.75} + k(q_{.75} - q_{.25})$, for $k = 3$ and $k = 1.5$. If *any* of U, GROUND, or JUDGE exceeds this bound, the whole triplet is excluded. The same history-first averaging rule applies. Because the input pool, threshold rule, and matched set differ, these are not additional percentile columns in Table 36.

Table 38: Relative mean changes (%) under the two previously examined IQR screens. Across all configurations, the screens remove 26 and 56 of the 480 U/G/J triplets respectively. Cutoffs, excluded triplets, and retained-history counts are recorded in the archived CSVs.

Receiver	Panel	3× IQR			1.5× IQR		
		G–U	J–G	J–U	G–U	J–G	J–U
DeepSeek V4.1 Flash	3a	+13.88	-0.65	+13.14	+8.16	-1.85	+6.16
DeepSeek V4 Pro	3b	-13.37	-8.67	-20.88	-14.41	-7.88	-21.15
Qwen3.7 Max	3a	-4.68	-12.58	-16.67	-5.97	-11.91	-17.17
Qwen3.7 Flash	3b	-10.24	-8.60	-17.96	-6.89	-6.99	-13.41
GPT-5.6 Sol	3a	-7.66	-8.11	-15.15	-9.44	-7.48	-16.21
GPT-5.6 Luna	3b	-9.13	-9.54	-17.79	-10.44	-10.24	-19.61
Grok-4.6	3a	-9.64	-13.90	-22.20	-6.83	-13.78	-19.67
Grok-4.3	3b	-10.64	-8.78	-18.49	-11.11	-10.59	-20.52

Both screens leave all eight J–G means negative, but the full U/G/J ordering holds in only seven configurations. DeepSeek Flash becomes the exception: G–U and J–U turn positive because high-U executions contributed to the original savings. Some histories lose all three repeats, so these estimates no longer describe the complete twenty-history suite. Full-data estimates remain primary; screening describes how particular executions influence the means.

H SAVED ARTIFACT EVIDENCE

The cases are selected illustrations from original executions, not an estimate of how often a support condition helps. They use public rules, task packets, and final artifact files retained in the separate evidence archive described in Appendix I. Scores are read from the selected formal table; artifacts are inspected without rescoring.

Table 39: Qwen Flash full-delivery transitions for the existing focal contrasts. Each row includes sixty matched executions. Improved means the comparison condition has $F_3 = 0$ and added support has $F_3 = 1$; worsened is the reverse. Neither full and both full describe ties.

Contrast	Neither full	Improved	Worsened	Both full
G–S	7	17	7	29
J–G	4	10	11	35
J–U	9	15	6	30

SS-05, Qwen Flash, repetition 1. At evaluation time 70, the suspension of job 2 is active. The rules prohibit postponing dispatch to evade that suspension. U includes job 2 in the route and window fields and leaves the blocked list empty. JUDGE removes that route/window, includes job 2 in the blocked list, and keeps both maintenance assignments at 50 minutes. The paired scores are $Q_3/C_3/F_3 = 0.25/1/0$ and $1/1/1$. Both interface-valid submissions encode actual plans; the difference is not a change in explanatory wording.

SS-14, Qwen Flash, repetition 2. The task links both the exception and withdrawal to E42. The general record B51 sets five years for originals and two for backups. Exception E42 extends them to nine and four years. Notice V71 withdraws only the backup extension. The original provision of E42 still applies, while backups revert to B51's two years. Both U and JUDGE retain nine years for original records, protect the original group, and mark migration as a backup-only plan. The differing backup fields are shown below. With creation day 100 and 360 days per year, $100 + 2 \times 360 = 820$; $100 + 3 \times 360 = 1{,}180$. The expiry is calculated from the supplied public convention, not calendar-year arithmetic.

Table 40: SS-14 saved retention artifacts. Only the backup artifact group changes score.

Field or outcome	U	JUDGE
Original retention / expiry	9 years / 3,340	9 years / 3,340
Backup retention / expiry	3 years / 1,180	2 years / 820
Q_3	0.75	1.00
C_3	1	1
F_3	0	1

The examples establish differences in submitted artifacts. They do not disclose internal reasoning, demonstrate which particular hidden operation failed, or show that JUDGE is uniformly more accurate than GROUND or every other support. The complete-data tables remain the basis for outcome frequency and resource comparisons.

I EVIDENCE ATTACHMENT AND OFFLINE INSPECTION

The submission candidates are this combined PDF and the anonymous evidence attachment. The editable manuscript bundle is retained separately by the author; the compilation instructions below describe that author-side source, not files expected in the evidence attachment. The manuscript bundle contains the LaTeX source, bibliography, conference style files, figure assets, numerical inputs, and figure-generation code. The companion attachment, `P1_ICLR_ANONYMOUS_EVIDENCE_SUPPLEMENT.zip`, extends the earlier compact package with existing runtime sources and non-selected-attempt records. It preserves the earlier scientific payloads and adds no experimental observations or formal analyses. Canonical task material remains in its task directories; other byte-identical files share one payload, usually under `data/`. The ZIP uses ordinary files, without links or remote dependencies. Sharing file bytes does not merge observations or change their identities.

Current files and historical paths. All current navigation CSV file paths, including Study 1's, are relative to the evidence-supplement root. In `FILE_PROVENANCE.csv`, `path` is the historical logical name; `compact_path` locates its stored bytes, and `compact_sha256` records their digest. For a relocated file cited elsewhere in this supplement, find its historical name in `path` and follow `compact_path`; `compact.resolve(logical_path)` performs this lookup and checks the original anonymous hash. Current indices should be opened directly: resolving their historical names instead retrieves the preserved source versions, whose path conventions may differ. Historical archive-member strings and symbolic roots remain provenance, not physical links.

Inspect an actual intervention. After extracting the evidence supplement, start with `study3/task_packet_inventory.csv`. Each of the twenty histories has six complete support packets. For example, compare `S`, `GROUND`, and `JUDGE` under `study3/material/tasks/SS-05/packets/`; each directory contains the actual `TASK_PACKET.md`. The same layout exists for SS-14 and the other histories. Within each task, `initial_workspace/` contains the public sources, rules, output interface, validator and artifact template. The `author/` directory supplies the frozen semantic partition, base truth, reference results, rule/status judgments, counterfactual fixtures and task-local scorer. These separate author materials allow inspection of construction and scoring; they were not additional actor inputs. The mapping in `study3/actual_packet_binding_evidence.csv` verifies that the selected actor-input, captured and suite packets have the same frozen hash for all 2,880 selected executions.

Trace a selected execution. `study3/selected_evidence_index.csv` maps panel, history, support, repeat, requested model and selected epoch to the captured public workspace, submitted artifact, original saved `score.json`, trusted events and terminal, native usage, and selection/replacement provenance. Its file-path columns now locate the physical payloads. Open `score_path` to read the original saved scoring object without running a scorer. A shared payload's physical filename need not match its name in the captured workspace. Complete saved formal analyses and selected-row records remain under `study3/formal/3a/` and `study3/formal/3b/`. Panel identities, scientific timestamps, original hash identities and non-full outcomes are retained.

Reconstruct a capture without scoring. An index column ending in `_logical_root` names a historical directory inventory, not a physical directory. Its `capture_manifest_path` or `logical_files_manifest_path` points to `FILE_PROVENANCE.csv`. From the extracted package root, Python 3.10 or later can reconstruct the `capture_logical_root` for the DeepSeek-Pro example below:

```
python -B compact.py --materialize-prefix \
  study3/selected_evidence/3b/s3b-2aba992ef7584ed7e4cc401decf30bd1/capture \
  --output reconstructed-capture
```

The output directory must be new or empty. The helper copies every file under that logical prefix, checks its original anonymous SHA-256, and restores its relative name without overwriting packaged files. The same entry accepts indexed Study 1 input/evidence roots and Study 2 capture roots. Reconstruction alone does not invoke a scorer.

Inspect execution versions and non-selected attempts. `study3/runtime_sources_index.csv` connects the actual versioned sources, task interfaces, allowed artifact paths, checker commands, and selected segment ranges. `study3/nonselected_attempts_index.csv` covers all 38 non-selected attempts, including the two failed replacements, and links original events, request ledgers, minimal historical authorization excerpts, and final selections. Its coverage and residual-gap fields distinguish detailed transport closure, host-stop evidence, and classification-only evidence. The source and publication hashes in `FILE_PROVENANCE.csv`, together with the excerpt ranges in `ANONYMIZATION_CHANGES.csv`, identify each released copy. Shared source bytes can be located or reconstructed through the same compact map. These records permit offline source and event inspection; they do not make live provider execution or operating-system isolation independently reproducible.

Verify a frozen Study 3 score offline. From the evidence-supplement root, use Python 3.10 or later:

```
python -B study3/read_frozen_score.py --coordinate \
  s3b-2aba992ef7584ed7e4cc401decf30bd1
```

This example selects DeepSeek Pro, SS-03, repeat 2, JUDGE. The current helper resolves the indexed files, reconstructs the frozen task and capture in a temporary directory, and checks the complete capture against `capture_file_hashes_before_scoring` in the saved score. It also checks the frozen scorer hash, then invokes that unchanged scorer with the saved trusted events and terminal. Its complete JSON output is compared with the original `raw_score` object. The underlying command is:

```
python -B TASK/author/score.py CAPTURE --world W0 \
  --events trusted-events.json --terminal trusted-terminal.json
```

Here `TASK` and `CAPTURE` denote the temporary copies for the selected history. The scorer's exit code one is an evaluated non-full artifact; two is a scorer error. Submission completion is still read from C_3/F_3, not from the exit code alone. The helper does not overwrite archived scores or call a provider. The helper succeeds on exact equality, including non-full outcomes. The historical `study3/representative_frozen_score_readback.json`, now at `data/19588.json` through the provenance map, records exact readback agreement for Qwen Flash SS-05/repeat 1 and SS-14/repeat 2 under U and JUDGE, and the DeepSeek example. These are package checks, not new observations or replacement official scores.

Inspect delivery transitions and the retained tail. `study3/qwen_flash_delivery_transition_evidence.csv` supplies all sixty matched history–repeat coordinates for each of GROUND–S, JUDGE–GROUND and JUDGE–U. Its 180 rows include both outcomes, failed content groups, artifact and score paths, and saved observed/expected field differences. The table recovers the existing delivery-transition counts, including improvements and worsenings, without a new interpretation of ambiguous rules. `study3/deepseek_pro_ss03_repeat2_evidence.csv` connects the U/GROUND/JUDGE triplet to its original usage, terminal state, artifacts and scores. The influential JUDGE execution remains in the formal full-data result.

Study 1 and Study 2 coverage. `study1/README.md` leads to the original tasks and scorers, `selected_960.csv`, `attempts_966.csv`, all 32 zero-score evidence records, and the saved formal analysis. Historical "P1 Study 2" denotes this paper's Study 1; original scientific filenames are retained. The indices distinguish 954 selected originals, six selected replacements and six non-selected originals. From the evidence-supplement root, `python -B study1/read_saved_results.py` checks the retained selection and saved scores against the original formal analysis without executing a scorer. Its required execution inputs are included through the compact map. The study README explains reconstruction; the historical `study1/METHODS_SOURCE_NOTES.md` is located at `data/14765.md`. Original Study 1 indices reached through the map retain their study-relative paths, unlike the current package-root-relative indices.

`study2/README.md` leads to the frozen five-support task/scorer materials, three original formal result sets and their saved analysis inputs. `study2/selected_evidence_index.csv` retains all 2,400 originally selected coordinates and original missingness. The fifteen technical completions remain separately identified under `study2/technical_completion/`, with their source-to-completion mapping and saved sensitivity inputs. They do not replace the original formal sets.

This release does not extend Study 2's runtime provenance. Its archived configuration projections are distinct from original request envelopes and authorization text; a projection alone does not establish those additional evidence layers.

Author-side manuscript compilation. From the manuscript source directory, run `pdflatex main.tex`, `bibtex main`, and two further `pdflatex main.tex` calls. The supplied figure assets permit compilation without regenerating the figures.

Author-side figure regeneration. The existing entry `scripts/figure_generate.py` uses the saved numerical inputs. The family assets use `scripts/figure_inputs/family_saved_inputs.json`, which retains the unrounded condition means and original intervals. The accompanying README documents the supported arguments and dependencies. Figure references in the manuscript are assigned by their placement: the work-allocation example, the eight-configuration reasoning comparison, the Qwen-Flash decomposition, and the within-family cost–delivery comparison.

These operations update manuscript presentation. They do not call a provider, choose different experimental attempts, overwrite saved scores, or generate new inferential results. The evidence-inspection and frozen-score readback commands described above are separate from manuscript compilation.

Anonymization and deliberate exclusions. `FILE_PROVENANCE.csv` retains source archive/member identities and original/anonymous hashes alongside the compact mapping. It and `SHA256SUMS.txt` describe the current storage representation; `SHA256SUMS.txt` covers the other physical members and excludes itself. `ANONYMIZATION_CHANGES.csv` records path anonymization and the exact fields or line ranges excerpted from newly released records. Reviewer indices resolve within this supplement; symbolic archive-root paths in historical metadata preserve source references. Task content, scorers, artifacts, scores, usage, selection and scientific timestamps are unchanged. The curated supplement excludes full provider-response bodies, encrypted continuation, credentials and account records, private paths, full private conversations and unrelated engineering history. It includes only the minimal historical authorization text relevant to the indexed execution and completion scope, distinguishing original text from contemporaneous scope summaries. It is distinct from the complete internal preservation archive. This release introduces no new experimental observations, selection, official scores or intervals. Its new ZIP digest is supplied with the release, rather than reusing the earlier package's digest.