CURRICULUMWORLDBENCH: BENCHMARKING AUTONOMOUS RECONSTRUCTION OF EXECUTABLE UNIVERSITY CURRICULA FROM THE WEB

Anonymous authors
Paper under double-blind review

ABSTRACT

University curricula are distributed rule systems: requirements, prerequisites, policies, and exceptions are scattered across institutional websites and must be composed correctly to support downstream decisions. We introduce CURRICULUMWORLDBENCH, a benchmark for autonomous reconstruction of source-grounded, executable curricula from the web. It is built from 735 rendered curriculum worlds at 111 institutions, of which 450 carry adjudicated executable references and form its instances, and it evaluates agents both on what they extract and on whether the recovered rules behave correctly and whether claims of completeness are justified. Every headline result is stated on the worlds whose reference is closure-complete, where reconstructing the whole rule system is well defined, and holds in direction on the larger set of scored worlds. Across models, scaffolds, controlled website layouts, and frozen institutional sites, we find a persistent gap between finding courses and reconstructing the rules that govern them. Even when course recovery is high, 190,296 of 252,108 decidable obligations are omitted entirely. Stronger models achieve higher whole program behavioral agreement, yet substantial coverage failures remain. Handing agents the pages that state the requirements improves a scaffold that cannot search for sources, while a browsing agent that can search gains little from the same pages. Agents also often report that reconstruction is complete even when the resulting curriculum still lacks a substantial share of its obligations. Together, these findings characterize curriculum reconstruction as a distinct agent capability. Success requires more than retrieving relevant pages or extracting individual rules. An agent must identify the governing sources, integrate their requirements into a coherent executable model, and recognize when important parts of that model are still missing or unresolved.

1 INTRODUCTION

University curricula are distributed rule systems. Degree requirements, prerequisites, and institutional policies appear across pages with different scopes and catalog versions, yet must be composed correctly to determine whether a student satisfies a program. Recovering the relevant courses is insufficient: a reconstruction may name the right courses while combining them incorrectly or missing a governing policy. This makes curricula a concrete setting for studying whether web agents can reconstruct an executable domain model from distributed sources.

The challenge involves both interpretation and completeness. An agent may encode every requirement it finds correctly while never discovering another obligation that governs the program. Conversely, some source statements do not support a unique executable interpretation and should remain unresolved. Evaluation must therefore distinguish what the agent recovered, whether its rules behave correctly, and whether its claim of completion is justified.

We introduce CURRICULUMWORLDBENCH, a benchmark for autonomous reconstruction of source-grounded, executable curricula. Each episode specifies an institutional entry point, catalog year, and curriculum scope. The agent discovers the governing sources and returns executable rules, supporting provenance, unresolved requirements, and a completion decision (Figure 1). Controlled website layouts vary access to the same governing content, while frozen institutional sites

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

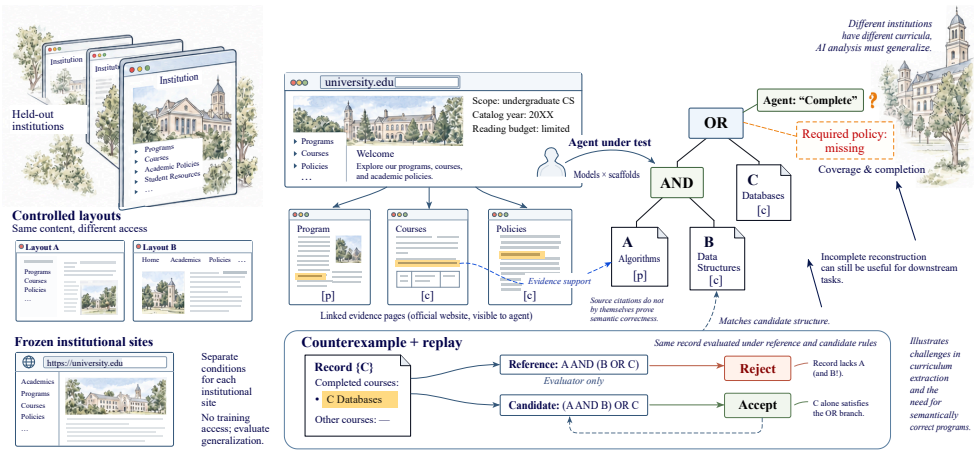

Figure 1: **CurriculumWorldBench: from official websites to executable curricula.** Under a reading budget, agents reconstruct curriculum rules with source evidence, explicit gaps, and a completion claim. Evaluation covers held-out institutions, content-preserving website layouts, and a separate frozen-site condition. The illustrated counterexample shows that recovering the same courses can still produce different decisions, while a missing governing policy can undermine an otherwise plausible claim of completeness.

provide a complementary evaluation condition. The benchmark distinguishes closure-complete references, which support evaluation against a closed obligation inventory, from diagnostic references that support only partial assessment.

The reported evaluations reveal a persistent gap between course recovery and executable obligation recall. They also show why completion must be assessed separately: a reconstruction can recover useful rules while still missing requirements. These observations motivate a benchmark that evaluates discovery, executable fidelity, and completeness within the same task.

Our contributions are:

- **A reconstruction task.** We formulate curriculum reconstruction from institutional websites as the production of executable rules with provenance, explicit unresolved content, and a completion decision.
- **An integrated evaluation protocol.** We combine controlled layouts and frozen sites with obligation-level coverage, semantic checks, program-level behavior, and assessment of completion claims.
- **An empirical diagnosis.** We evaluate multiple models and scaffolds, examining obligation omissions, semantic errors, source access, and the reliability of reported completion.

2 Related Work

University curriculum research has shown that academic requirements can be represented as structured constraints and used for downstream reasoning. KNOWPLAN reconstructs curricula from official university sources and uses a closure certificate before downstream degree planning (Cao et al., 2026). Earlier work has similarly encoded study regulations as executable logic programs (Hahn et al., 2024). **Our focus is a benchmark that evaluates obligation recovery, executable fidelity, and completion claims across controlled website layouts and frozen institutional sites.**

Rule extraction addresses part of this problem. Existing benchmarks study conditional structure, dependencies, and other relations expressed in documents (Yang et al., 2026; Holzenberger & Durme, 2021). Work on rules as code goes further by translating regulatory text into executable representations (Merigoux et al., 2021; Lorenzo et al., 2025), while rule reasoning benchmarks evaluate models on textual rule systems supplied directly in the input (Zhou et al., 2025; Servantez et al., 2026).

These tasks generally supply the rule-bearing material; curriculum reconstruction additionally requires identifying which distributed sources govern the target program and catalog year.

WebArena and Mind2Web evaluate agents that navigate websites and complete tasks (Zhou et al., 2024; Deng et al., 2023), and Mind2Web 2 extends this setting toward evidence-grounded information seeking (Gou et al., 2025). WideSearch evaluates the completeness of fact tables assembled from the open web (Wong et al., 2025), while τ-bench evaluates policy-guided interactions through the resulting database state (Yao et al., 2025). YAGO integrates heterogeneous sources into a structured knowledge base (Suchanek et al., 2007). **Building on these capabilities, curriculum reconstruction requires a persistent, executable representation of governing requirements that can determine whether previously unseen student records satisfy a program.**

This requirement connects curriculum reconstruction to executable evaluation. AppWorld and SWE-bench judge systems through the consequences of execution rather than surface similarity alone (Trivedi et al., 2024; Jimenez et al., 2024). Behavioral testing and semantic parsing evaluation likewise use test cases or execution to distinguish outputs that look similar from outputs that behave differently (Ribeiro et al., 2020; Zhong et al., 2020; He et al., 2024; Chu et al., 2017). **Execution can expose errors in recovered rules without establishing that all governing obligations have been recovered.** Benchmarks of abstention **evaluate recognition of unanswerable inputs** (Kirichenko et al., 2025); **here, the prediction concerns whether a reconstruction satisfies a declared closure criterion.**

3 What Should We Measure?

Curriculum reconstruction requires evaluating both the rules an agent produces and the requirements it fails to recover. Entity-level metrics measure vocabulary recovery (Yang et al., 2026), but cannot distinguish reconstructions that name the same courses while composing them differently (Figure 1). Conversely, correctly encoding every emitted rule does not establish completeness: a governing requirement may never have been discovered. Behavioral tests can expose semantic differences that surface comparisons miss (Ribeiro et al., 2020; Zhong et al., 2020), yet whole-program agreement depends on the test-record distribution and can conceal missing obligations. Evaluation must also distinguish prediction errors from uncertainty in the reference and assess whether an agent's claim of completion is justified.

These considerations motivate four evaluation requirements. **Coverage** measures recovery against a reference obligation inventory, including obligations with no executable counterpart in the prediction. **Behavioral fidelity** checks whether recovered rules preserve the reference's decisions within a declared domain and assesses the reconstructed program on student records. **Explicit uncertainty** excludes unresolved reference semantics from correctness judgments while reporting the resulting evaluable coverage. **Completion awareness** treats an agent's completion claim as a selective prediction, measuring how often agents certify completion and how often those certifications fail the stated closure criterion. Together, these requirements distinguish vocabulary recovery, obligation coverage, executable correctness, and justified completion. Section 4 specifies the reference boundaries and operationalizes these measurements.

4 Benchmark Design

4.1 Task Definition

An episode asks an agent to rebuild one curriculum from an institution's own website, and tells it only where to start and what to rebuild. It consists of an official entry page r, such as the home page of a university catalog, a catalog year y, which fixes the version of the rules that applies, and a scope s, such as "the B.S. in Computer Science with the general-education requirements and policies that govern it." We write the episode as $e = (r, y, s)$. None of the three points to the pages that actually state the rules. Starting from r, the agent has to find those pages, read the requirements they state, and compose the requirements into a curriculum that can be executed. The agent returns a reconstruction $\widehat{\mathcal{C}} = A(e)$ with four parts, namely the requirements it could express as executable rules, the source passage behind each rule, the requirements it found but could not resolve, and a decision on whether the reconstruction is complete. Executed on a student record, the rules decide

whether that student satisfies the program. The last two parts let an agent return an honest partial reconstruction and turn "I am done" into a checkable claim (Appendix E).

4.2 Instance Construction

Each episode is embedded in a *curriculum world* $w = (e, \mathcal{D}^\star, \mathcal{C}^\star, \mathcal{V})$ that fixes everything needed to score it. The governing sources $\mathcal{D}^\star$ are the pages that actually define the target, for example the program page, the course pages it links to, and the policy pages that change how those courses count. The reference curriculum $\mathcal{C}^\star$ is the executable curriculum derived from those pages (Section 4.3). The views $\mathcal{V}$ are the websites through which an agent can reach $\mathcal{D}^\star$.

The central design choice is to hold $\mathcal{D}^\star$ and $\mathcal{C}^\star$ fixed while the website $v \in \mathcal{V}$ changes. Controlled views render the same governing content as a few departmental pages, as paginated listings, or deep inside a navigation tree, and every rendering keeps all reference evidence reachable. Two views of the same world therefore differ only in how the information is organized, never in what has to be reconstructed. Frozen snapshots of the original institutional sites, available for a subset of worlds, add the size and link structure of real websites while the target and the reference stay the same.

4.3 Ground Truth

A reference curriculum is an interpretation of institutional text, and the benchmark keeps that interpretation open to inspection. Each governing requirement is stored three ways, as the source passage in the frozen snapshot, as an adjudicated statement of what the passage requires, and, when the rule language can express that statement, as an executable rule in $\mathcal{C}^\star$. The rule language covers required courses, alternatives, choose-k groups, credit pools, exclusions, ordered prerequisites, and conditions on facts such as class standing. We call a requirement *reference-decidable* when its adjudicated meaning has such an executable form. A requirement whose source is ambiguous, or whose construct lies outside the language, stays in the world as explicit `UNKNOWN` content with its source passage and is never scored against an agent.

References come from human annotators, who either corrected a pre-filled machine reconstruction or wrote the worksheet from the source alone. Requirements they returned in prose were translated into the rule language by a language model and kept only after a second model accepted the translation, and anything still awaiting a human ruling stays `UNKNOWN`. Worlds whose adjudication is finished carry a *closure-complete* reference, in which every obligation in scope is either executable or an explicit `UNKNOWN` with a recorded ruling, so that the inventory of obligations is closed and "every obligation recovered" has a meaning. The others carry a *diagnostic* reference that locates obligations without closing the inventory, so a reconstruction can be scored against the obligations it lists but cannot be judged complete (Appendix D).

4.4 Evaluation

The metrics follow the properties of Section 3, distinguishing entity recovery, obligation coverage, behavioral fidelity, and completion claims. Every obligation-level metric counts reference-decidable obligations only, and the share of obligations that are reference-decidable is reported as evaluable coverage. Before meaning is compared, each reference obligation is aligned with the predicted rule that expresses it, by the identifiers and content of the rule, with declared normalizations and reviewed mappings for a requirement that a reconstruction spreads over several rules (Appendix F).

Vocabulary. Entity precision, recall, and F_1 compare the courses named by the reconstruction with those of the reference. They measure recovery of the curriculum's vocabulary, but do not by themselves establish rule correctness or completeness.

Coverage. An obligation is *recovered* when an executable rule in the reconstruction aligns with it. *Omission* is a composite coverage shortfall: one minus the lower of course recall and the fraction of reference-decidable obligations recovered. Obligation-level absence, reported separately in the failure decomposition, counts reference-decidable obligations without aligned executable counterparts.

Behavioral equivalence. A recovered rule is compared with its reference by what it decides. Write $q(x)$ for the verdict that an obligation q, for instance "a grade of C or better in CS 101 before CS 201," returns on a student record x, and let $\mathcal{X}$ be the set of student records the rule language can describe. For a reference obligation q and its recovered rule $\hat{q}$, the evaluator searches $\mathcal{X}$ for a record on which the two disagree,

$$\exists\, x \in \mathcal{X} : \quad q(x) \neq \hat{q}(x). \tag{1}$$

A record found this way is a witness of non-equivalence, and an independent verifier replays it. A successfully replayed witness establishes non-equivalence; establishing equivalence requires proving that no such witness exists within the supported domain. Rules outside the supported fragment stay undecided.

Executable recall. *Executable Obligation Recall* (Exec. Recall) is the share of reference-decidable obligations that are recovered and proven equivalent. It is the quantity the archived scorer calls semantic accuracy, renamed because missing obligations stay in its denominator.

Whole-program agreement. Obligations also interact, for example when one course could count toward two requirements. Behavioral agreement (BA) runs the complete reconstructed and reference curricula on a frozen suite of student records and reports how often their verdicts match. We report BA alongside the empty-program baseline on the same record suite. This measures agreement on the specified suite rather than unrestricted program equivalence.

Completion. The completion decision is scored as a selective prediction (Geifman & El-Yaniv, 2017; 2019). Over N episodes, let $z_i = 1$ if the agent certified episode i as complete and $c_i = 1$ if the reconstructed world is judged correct. Then

$$\text{Coverage} = \frac{1}{N}\sum_i z_i, \qquad \text{Risk} = \frac{\sum_i z_i(1-c_i)}{\sum_i z_i}. \tag{2}$$

Coverage measures how often an agent claims to be done, and risk how often such a claim is wrong. An agent that never certifies is safe but says nothing, so the two are read together. World correctness is judged by obligation closure, which holds when every decidable obligation of the world is recovered and equivalent and is scored on closure-complete worlds only, and Section 6.3 also reports the certified reconstructions' own recall.

5 Benchmark Validation

5.1 Annotation Reliability

Obligation-level scores are agreements with a human-built reference, so the reference has to be consistent. Two passes that both start from a pre-filled reconstruction agree on 92.1% of fields, and a pre-filled pass and a pass written from the source alone agree on 87.2%, close to the 87.7% agreement between two source-only passes (Appendix G.1). Pre-filling appears to nudge annotators toward the proposal. These rates measure consistency rather than chance-corrected reliability, and adjudication itself is not separately measured.

5.2 Metric Validity

The evaluator must respond to changes in meaning and to nothing else, and we test this on the reference rules. From 8,181 executable reference rules we generated 22,914 mutants. The evaluator decides 81.7% of them and distinguishes 99.8% of those it decides from the original rule. The 36 decided mutants it does not distinguish preserve meaning on independent inspection, so no meaning-changing mutant was missed among the decided cases in this audit. In the opposite direction, none of 14,265 meaning-preserving rewrites is flagged as different, and 10,791 of them are proven equivalent. Mutants outside the fragment, and the semantics of general-education and policy blocks, lie outside this audit (Appendix R).

The alignment step is audited separately, because it decides whether an obligation counts as absent. Two auditors read 300 obligations that the scorer marked absent, drawn at random from the scored

cells of Table 1, and agreed on 95.0% of them before adjudication. 279 are absent. The other 21 (7.0%) have a rule the auditors read as expressing the obligation, and 13 of those rules are equivalent to it once aligned by hand, so alignment understates Exec. Recall by about 3 points at most, and the omitted share of failures in Section 7.1 stays above 77% in every run.

Reading the scorer's rules against the reference files alone also shows that a strict pass, in which every in-scope decidable obligation of a world is recovered, supported by evidence, and equivalent, is attainable in only 18 of 243 scored worlds, because most references carry world-level credit contracts that the strict path cannot resolve (Appendix R.1). Strict WORLDPASS is therefore bounded by the instrument rather than by the agents, and we report it only as a diagnostic.

5.3 CONTAMINATION AND LEAKAGE

Worlds are split by institution into 12 development and 48 test institutions, and every view of a world stays on its side, so conventions that recur within a university cannot inform both sides of the split. Exposure during pretraining is a separate question, since the pages are public and we cannot inspect what the evaluated models saw in training. A closed-book control bounds what memory contributes. Given only the institution, catalog year, and scope of each of the 243 scored worlds and no pages, Sonnet 5 names the reference courses at an entity F_1 of 12.6% and recovers 0.9% of the decidable obligations, and Opus 5 17.3% and 2.2%, against 60.0% and 41.1% for Sonnet 5 reading a single bundle of pages (Table 1). The results point the same way, since browsing course recovery under Sonnet 5 falls from 74% on the re-renderings to 39% on the original websites of the same curricula (Table 13), so what an agent reconstructs depends on what it reads.

5.4 SENSITIVITY ANALYSIS

Repeated runs. The serving path of the Claude pair offers no sampling control, so we execute every controlled-layout plan twice and treat model differences smaller than the difference between two executions of the same plan as unresolved (Table 19). Browsing agents are fairly stable across executions, one-shot compilation much less so, and the structured ledger is nearly deterministic.

Output format. Under the original contract, which requires every field line to be indented by exactly two spaces, all 756 GPT-6 one-shot cells were rejected although each returned complete, schema-valid JSON, and re-parsing the stored replies with any leading whitespace accepted rescues 639 of 2,097 failed one-shot cells, all in this row, with no other row changing (Appendix Z). Table 1 and Figure 2 therefore use the whitespace-tolerant contract.

Reference completeness. Adjudication is finished for 63 of the 243 scored worlds, and the other 180 carry diagnostic references. The share of obligations the reference can score is similar in both groups (71.5% and 69.4%), but a diagnostic reference lists fewer of the obligations that are hard to find, and it flatters the agents. Re-aggregating the scored cells over each group, Exec. Recall is 3 to 12 points higher on the diagnostic worlds than on the closure-complete ones for every model, while every comparison of Section 6 points the same way in both (Table 31). This is why the headline results are stated on the closure-complete worlds and the full set is reported as an extended tier.

6 BENCHMARKING CURRENT SYSTEMS

6.1 SETUP

Two tiers of worlds are reported. The *confirmatory tier* is the 63 closure-complete worlds from 18 institutions, fixed before scoring, and every headline claim of Sections 6 and 7 is stated on it. The *extended tier* adds the 180 diagnostic worlds and is used for breadth, for the paired Claude comparison with its repeated executions, and for the original-website condition. We evaluate six base models with five reconstruction scaffolds. Claude Sonnet 5 and Claude Opus 5 form a paired comparison with a shared serving path and repeated executions. GPT-5.6, GPT-6, Qwen3.8-27B, and Gemma-3-27B-it run on the same worlds, scaffolds, references, and scorer through their own serving paths and without repetitions (Table 26), so they show breadth across model families rather than a ranking. One-shot compilation reads a fixed bundle of pages, BM25+LLM a retrieved context, and

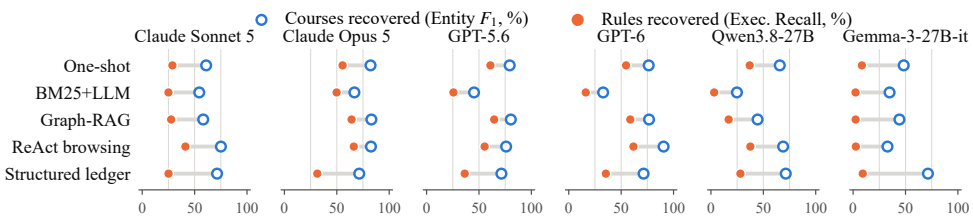

Figure 2: **Recovering the courses is not recovering the rules.** For every model and scaffold on the closure-complete worlds, the open circle marks Entity F_1 and the filled circle Exec. Recall, so the length of each segment is the gap between finding a curriculum's courses and reconstructing its rules. Values are those of Table 1.

Table 1: Confirmatory results on the closure-complete worlds (%), three controlled layouts, 189 cells per row. **Exec. Recall** is the share of reference-decidable obligations recovered and proven equivalent, **Omission** the larger of the missing-course and missing-obligation shares (Section 4.4), and **Closed** the share of cells in which every decidable obligation of the world is recovered and equivalent. Answers are read under a contract that accepts any leading whitespace (†: 0.0 under the original contract, Section 5.4). ‡Course inventory from deterministic adapters, identical under every model. The same rows on all 243 scored worlds are in Appendix K.1.

Model	Scaffold	Entity F_1 ↑	Omission↓	Exec. Recall↑	Closed↑
Claude Sonnet 5	One-shot	61.0 [40.4, 78.6]	78.6 [58.3, 99.0]	28.8 [11.0, 44.8]	1.6
	BM25+LLM	54.4 [33.3, 74.4]	84.6 [66.1, 100.0]	25.0 [5.8, 40.8]	1.6
	Graph-RAG	58.2 [35.6, 79.3]	77.2 [49.1, 100.0]	27.6 [7.7, 50.5]	3.2
	ReAct browsing	75.1 [59.9, 89.6]	78.6 [63.6, 90.8]	41.2 [22.4, 63.1]	4.8
	Structured ledger‡	71.4 [59.4, 80.0]	74.0 [55.6, 93.9]	25.1 [4.4, 45.5]	0.0
Claude Opus 5	One-shot	82.2 [66.9, 96.6]	56.9 [36.6, 80.3]	55.5 [35.1, 74.9]	7.9
	BM25+LLM	66.8 [41.5, 86.0]	64.7 [37.5, 95.9]	49.8 [22.4, 71.4]	4.8
	Graph-RAG	83.0 [59.0, 100.0]	54.6 [29.8, 82.5]	64.1 [47.4, 79.0]	12.7
	ReAct browsing	82.6 [65.9, 96.2]	58.1 [43.7, 73.5]	66.2 [52.1, 82.0]	15.9
	Structured ledger‡	71.4 [59.4, 80.0]	68.9 [50.2, 89.2]	31.4 [10.0, 52.7]	0.0
GPT-5.6	One-shot	79.3 [58.9, 94.4]	54.8 [30.6, 83.3]	60.9 [35.6, 80.9]	7.9
	BM25+LLM	45.3 [26.0, 64.4]	82.2 [71.3, 94.8]	25.6 [10.2, 42.4]	1.6
	Graph-RAG	80.4 [60.3, 95.6]	52.6 [28.4, 80.2]	64.5 [40.7, 84.0]	11.1
	ReAct browsing	75.9 [57.2, 92.7]	65.7 [50.5, 80.1]	55.4 [36.2, 76.7]	9.5
	Structured ledger‡	71.4 [59.4, 80.0]	66.2 [49.6, 84.1]	36.4 [18.3, 55.6]	1.6
GPT-6	One-shot†	76.3 [54.4, 92.8]	59.0 [35.6, 86.7]	54.7 [29.1, 74.9]	6.3
	BM25+LLM	32.8 [10.1, 53.0]	92.4 [78.7, 100.0]	16.2 [0.0, 39.4]	0.0
	Graph-RAG	76.6 [53.2, 94.4]	58.0 [34.7, 86.5]	58.7 [33.9, 77.5]	9.5
	ReAct browsing	90.5 [79.7, 99.3]	52.5 [35.2, 72.0]	61.7 [45.4, 78.1]	12.7
	Structured ledger‡	71.4 [59.4, 80.0]	66.8 [50.0, 83.8]	35.4 [18.0, 54.6]	1.6
Qwen3.8-27B	One-shot	65.7 [38.2, 87.6]	66.6 [41.8, 91.2]	37.0 [15.7, 58.6]	3.2
	BM25+LLM	24.8 [9.4, 39.0]	100.0 [97.6, 100.0]	3.1 [0.0, 12.2]	0.0
	Graph-RAG	44.5 [27.6, 59.4]	86.4 [72.5, 98.1]	16.9 [4.1, 33.1]	1.6
	ReAct browsing	68.9 [57.4, 79.3]	72.8 [57.3, 87.0]	37.6 [18.1, 61.3]	4.8
	Structured ledger‡	71.4 [59.4, 80.0]	71.0 [55.1, 88.7]	28.1 [10.2, 47.3]	0.0
Gemma-3-27B-it	One-shot	48.3 [24.7, 73.5]	94.0 [83.2, 100.0]	8.4 [0.0, 25.1]	0.0
	BM25+LLM	34.8 [16.9, 51.0]	99.2 [92.3, 100.0]	2.3 [0.0, 13.6]	0.0
	Graph-RAG	44.2 [19.4, 65.3]	99.2 [94.1, 100.0]	2.4 [0.0, 11.5]	0.0
	ReAct browsing	32.8 [18.4, 49.2]	99.1 [96.9, 100.0]	2.6 [0.0, 7.9]	0.0
	Structured ledger‡	71.4 [59.4, 80.0]	86.3 [62.8, 100.0]	9.2 [0.0, 30.5]	0.0

Graph-RAG a context retrieved along the site's link graph, while ReAct browsing explores the site itself and the structured ledger also tracks evidence and unresolved obligations, extracting courses with deterministic adapters. Every configuration runs on the three controlled layouts under fixed page and call budgets (Appendix C), and the Claude pair also runs on frozen original websites of 234 worlds with two source-access interventions. The benchmark's instances are the 450 worlds with adjudicated references, drawn from 735 rendered worlds at 111 institutions and split by institution. Results cover 243 scored worlds among the 252 test worlds of 48 held-out institutions, with 95% intervals from a bootstrap over institutions and model comparisons paired by world.

6.2 Main Results

Figure 2 and Table 1 give an answer in two parts. Entity F_1 exceeds Exec. Recall in all thirty rows, by a median of 29 points and by as much as 62. At the same time, several configurations achieve ex-

ecutable recall above 50% on reference-decidable obligations. Opus 5 with ReAct browsing reaches 66.2% Exec. Recall. Decidable-obligation closure remains infrequent. The best configuration recovers every decidable obligation of a world in 15.9% of its episodes, Sonnet 5 in at most 4.8%, the structured ledger in none, and coverage shortfall exceeds 50% in every row.

Observed model differences vary across scaffolds. Opus 5 improves on Sonnet 5 by 25 to 37 points of Exec. Recall on four scaffolds and by 6 on the structured ledger, whose course inventory comes from deterministic adapters and whose Exec. Recall stays between 25% and 36% for every model except Gemma. GPT-5.6 and GPT-6 come close to Opus 5 on their best scaffolds, while the two open-weight models trail far behind. Most of their rejected answers are well-formed documents that break the worksheet contract (Appendix O), so part of their deficit is interface compliance rather than reading. Within the cells whose answers passed the contract, the Claude rows move by less than a point, the GPT rows by 2 to 6 points, and the two open-weight rows by 4 to 9 points of Exec. Recall, and no row changes its place.

6.3 Capability Breakdown

Vocabulary. Finding courses is the most solved stage. Browsing reaches 68.9% to 90.5% Entity F_1 for five of six models, and a deterministic extractor with no language model recovers 95.4% of course entities on two layouts while recovering no executable requirement (Appendix L). On the extended tier, Opus 5 and Sonnet 5 find courses about equally well when browsing, with a paired difference of +2.5 points within run-to-run variation on every layout (Table 10).

Coverage, fidelity, and executable recall. The rules are where reconstructions fall short, and they fall short mainly through omission. In every run we decompose, at least 83.7% of the reference-decidable obligations an agent fails on have no counterpart in its reconstruction, rather than a counterpart that is present and proven wrong (Section 7.1). When an agent does emit a rule that can be checked, the rule is usually right. Exec. Recall therefore tracks coverage, and it is where the stronger model makes its gains. Opus 5 raises browsing Exec. Recall from 41.2% to 66.2%, three times its gain in course recovery, so its advantage lies in recovering obligations more than in finding courses.

Whole-program behavior. Running the full programs on the frozen record suites of the extended tier gives BA of 0.503 for Sonnet 5 and 0.574 for Opus 5 over all scaffolds, a paired gain of +7.1 points [+4.6, +9.5] over 3,510 cells. Under Sonnet 5, 47% of cells do no better than the empty program, and reconstructions decide on average 10.5 of 15.3 decidable obligations per world incorrectly, against 7.4 under Opus 5 (Table 21).

Source access. A median original site has 189 pages, of which the page budget covers 14%, although the reference evidence is reachable from the entry page in every one of the 234 worlds (Appendix M). Handing the agent exactly the pages the reference cites raises the structured ledger's Exec. Recall by 18.6 points under Sonnet 5 and 16.8 points under Opus 5, but gives the search-capable browsing agent no gain that separates from zero under either model (Table 30). In-page search likewise lifts the browsing agent's course recovery on real sites from 39.2% to 51.8% without changing its Exec. Recall.

Completion awareness. Certification is frequent, and it is not a closure claim. On the closure-complete worlds under Sonnet 5, one-shot compilation certifies 72% of its episodes at an Exec. Recall of 28.8% (Table 16). Scored against obligation closure, 98% of these certifications are wrong, and the certified episodes' own Exec. Recall is 31.5%, against 21.8% for the episodes the same run left uncertified. Under Opus 5, which certifies 84% of one-shot episodes at 55.5%, the risk is 92% and the certified episodes reach 57.5% against 45.0%. A certified reconstruction is thus somewhat better than an uncertified one and still lacks more than two fifths of its decidable obligations. Certification also tracks the scaffold more than the reconstruction, and on the original websites it nearly disappears (Table 16).

7 Analysis

7.1 Failure Modes

The outcome decomposition covers 252,108 reference-decidable obligation observations (Appendix T). Among obligations not proven equivalent, 99.3% are classified as absent from the reconstruction, and 99.1% on closure-complete worlds alone; this share remains at least 83.7% in every decomposed run. Absent counterparts may reflect undiscovered sources, unencoded requirements, empty or contract-invalid outputs, or alignment failures. The audit in Section 5.2 found predicted counterparts for 7.0% of sampled obligations classified as absent, so this category should not be interpreted as confirmed source omission.

Present-but-wrong rules differ primarily in course sets or grade conditions rather than logical structure. For example, in `cod-service-technician-certificate`, the ReAct-style browsing agent on the paginated layout reconstructed `FIELD-OF-STUDY-CODE-HVACR-CER-HVAC`. On the record {`HVACR 1100`}, the reference returns True and the prediction returns False. An independent verifier replayed this evaluator-generated counterexample, confirming a decision error in an emitted rule (Appendix X).

Browsing traces provide a partial distinction between source access and subsequent reconstruction failures. On controlled layouts, 71% of obligations classified as omitted under Sonnet 5 and 68% under Opus 5 have an evidence page among those the agent opened. On original websites, the shares fall to 38% and 41%, respectively. Many omissions persist despite page access on controlled layouts, whereas failure to open an evidence page is more prevalent on original sites. Opening a page does not establish that the relevant passage was fully exposed to or processed by the model.

7.2 Implications

These findings motivate distinguishing source access, obligation recovery, and semantic fidelity. Page access does not ensure rule recovery, and correctness of recovered rules does not establish closure. Completion claims must therefore be evaluated against the declared obligation inventory, as in Section 6.3, rather than inferred from retrieval or extraction performance.

8 Limitations

CurriculumWorldBench measures reconstruction within a supported rule fragment. About 70% of enumerated obligations are reference-decidable, the rest stay `UNKNOWN`, and only 63 of the 243 scored worlds, from 18 institutions, are closure-complete, so the confirmatory tier rests on few institutions and its intervals are wide. Aligning reference obligations with predicted rules is a judgment step, audited on a sample of 300 obligations (Section 5.2). Annotation quality rests on consistency diagnostics rather than an independent reliability study, and strict WorldPass, attainable in only 18 of 243 worlds, is reported only as a diagnostic. Because the pages are public, pretraining exposure cannot be excluded, although the closed-book control recovers little. The original-website condition covers 234 worlds, the serving paths do not control sampling, repetitions are limited to one per controlled layout, and the source-access interventions cover two scaffolds and two models in one condition (Appendix B).

9 Conclusion

Reconstructing a rule system from the web is a different problem from retrieving information from it. CurriculumWorldBench makes its parts separately observable, with obligation-level coverage, semantic checks with independently replayed counterexamples, explicit reference uncertainty, and completion claims scored as predictions, and it validates its own evaluator before comparing systems. Current systems recover a curriculum's vocabulary more reliably than the rules that govern it, close a rule system in at most one episode in six, and obligations without aligned counterparts account for most of the remaining error. Closing the obligation inventory, and knowing when it is closed, is the central challenge ahead.

AI USE STATEMENT

Language-model agents assisted in drafting and formatting this manuscript and, as Appendix J describes, in translating annotated natural-language fields into the rule language under human review. They played no part in the evaluated runs beyond being the systems under test. The authors have reviewed every passage of this paper and are responsible for its content.

ETHICS STATEMENT

The task concerns published institutional rules, with synthetic student records used as semantic probes. Sources are held as frozen, hashed snapshots of published pages; redistribution remains subject to the permissions and restrictions of the originating institutions, and the authorization under which they were collected is not documented in the current evidence package, so nothing is asserted about it. The snapshots and the references derived from them are part of the intended release. Per-institution permission for redistribution was not sought, and the authors take responsibility for that decision. How the pages were fetched is documented, and Appendix D.1 reports it in full: every request went through one rate-limited function that sends a research User-Agent and no credential, `robots.txt` was fetched per host and had to permit a URL before it was requested, a host whose `robots.txt` could not be read was treated as closed, bot challenges were recorded and never bypassed, and hosts that disallowed the content were left alone—rules the records state were not relaxed under schedule pressure. One failure is disclosed with them: the one-second floor between requests holds per process and was not coordinated across processes, and on 2026-09-07 two overlapping queues put two runners on the same school, so that one institution's server received about two requests per second for that period; the duplicate runner was stopped and one runner per school became a standing rule. An incorrect reconstruction could mislead a student if treated as authoritative advising, so benchmark outputs do not constitute degree audits.

REPRODUCIBILITY STATEMENT

The construction and diagnostic artifacts associate source snapshots, reference versions, evaluation reports, and source hashes. Concretely: every cell pins the prompt, source snapshot, reference and scorer it ran against in a frozen execution plan; every attempt is preserved, including the aborted ones (Table 8); each table in this paper is generated from hashed inputs listed with their SHA-256 in the package's input manifest; the institution partitions, model and CLI versions, budgets, evaluator configuration and resampling seed ($B = 10{,}000$, school clusters) are recorded per run. The caveat of Section 6.1 stands: the serving path exposes no sampling control, so a rerun is not expected to reproduce a cell byte for byte. The authors commit to releasing the code and the annotation protocol, and an anonymized artifact of the code and the protocol, carrying no institutional page bytes, is provided as supplementary material. What the evidence package holds has been inventoried component by component: 18 components, 17 of them hashed, and those 17 hold 172,992 files and 5,438.88 MB; the component without a manifest hash is the pinned scorer root, which is not a release candidate. The release is what the authors decided it is: the code, the annotation protocol, and—by the authors' decision—the institutional snapshots and the references derived from them, which are 4 components and 1,686.06 MB of the hashed content; the evidence package records no per-institution authorization for that redistribution, and no legal basis for the decision is asserted here. Whether any other component of that inventory is released is not decided here, and this paper claims nothing about it.

REFERENCES

Shuheng Cao, Weijia Zhang, Jiaqi Wu, Xiyun Hu, Yat Yang, Juqy Chen, and Zhaoxiang Feng. Knowplan: Knowledge-driven ai agents for smart degree pathway planning. arXiv preprint, 2026. URL `https://arxiv.org/abs/2608.06530`.

Zijian Chen, Xueguang Ma, Shengyao Zhuang, Ping Nie, Kai Zou, Andrew Liu, Joshua Green, Kshama Patel, Ruoxi Meng, Mingyi Su, Sahel Sharifymoghaddam, Yanxi Li, Haoran Hong, Xinyu Shi, Xuye Liu, Nandan Thakur, Crystina Zhang, Luyu Gao, Wenhu Chen, and Jimmy

Lin. BrowseComp-Plus: A more fair and transparent evaluation benchmark of deep-research agent. arXiv preprint, 2025. URL https://arxiv.org/abs/2508.06600.

Shumo Chu, Chenglong Wang, Konstantin Weitz, and Alvin Cheung. Cosette: An automated prover for sql. In *CIDR*, 2017. URL https://www.cidrdb.org/cidr2017/papers/p51-chu-cidr17.pdf.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2web: Towards a generalist agent for the web. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. doi: 10.18653/v1/2023.emnlp-main.398. URL https://aclanthology.org/2023.emnlp-main.398/.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *International Conference on Machine Learning*, pp. 2151–2159, 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Boyu Gou, Zanming Huang, Yuting Ning, Yu Gu, Michael Lin, Weijian Qi, Andrei Kopanev, Botao Yu, Bernal Jimenez Gutierrez, Yiheng Shu, Chan Hee Song, Jiaman Wu, Shijie Chen, Hanane Nour Moussa, Tianshu Zhang, Jian Xie, Yifei Li, Tianci Xue, Zeyi Liao, Kai Zhang, Boyuan Zheng, Zhaowei Cai, Viktor Rozgic, Morteza Ziyadi, Huan Sun, and Yu Su. Mind2Web 2: Evaluating Agentic Search with Agent-as-a-Judge, 2025. URL https://arxiv.org/abs/2506.21506.

Susana Hahn, Torsten Schaub, Cedric Martens, Amade Nemes, Henry Otunuya, Javier Romero, and Sebastian Schellhorn. Reasoning about study regulations in answer set programming. *Theory and Practice of Logic Programming*, 24(4):790–804, 2024. doi: 10.1017/S1471068424000383. URL https://www.cambridge.org/core/journals/theory-and-practice-of-logic-programming/article/reasoning-about-study-regulations-in-answer-set-programming/8869C24FEB04630CA9358B7E02E265D8. Volume/issue year 2024; first published online 15 January 2025; publisher PDF bears 2025.

Yang He, Pinhan Zhao, Xinyu Wang, and Yuepeng Wang. VeriEQL: Bounded equivalence verification for complex SQL queries with integrity constraints. *Proceedings of the ACM on Programming Languages*, 8(OOPSLA1):1071–1099, 2024. doi: 10.1145/3649849.

Nils Holzenberger and Benjamin Van Durme. Factoring statutory reasoning as language understanding challenges. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 2742–2758, 2021. doi: 10.18653/v1/2021.acl-long.213. URL https://aclanthology.org/2021.acl-long.213/.

Wendan Jiang, Shiyuan Wang, Hiba Eltigani, Rukhshan Haroon, Abdullah Bin Faisal, and Fahad Dogar. AdvisingWise: Supporting academic advising in higher education settings through a human-in-the-loop multi-agent framework. arXiv preprint, 2025. URL https://arxiv.org/abs/2511.05706.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R. Narasimhan. SWE-bench: Can Language Models Resolve Real-World GitHub Issues? In *ICLR*, 2024. URL https://arxiv.org/abs/2310.06770v3.

Polina Kirichenko, Mark Ibrahim, Kamalika Chaudhuri, and Samuel J. Bell. AbstentionBench: Reasoning LLMs fail on unanswerable questions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://arxiv.org/abs/2506.09038.

Feng Li and Yoritaka Iwata. A locally deployed RAG-based academic advising system for course selection. arXiv preprint, 2026. URL https://arxiv.org/abs/2606.02983.

Gabriele Lorenzo, Aldo Pietromatera, and Nils Holzenberger. Translating tax law to code with llms: A benchmark and evaluation framework. In *Proceedings of the Natural Legal Language Processing Workshop 2025*, pp. 31–47, 2025. doi: 10.18653/v1/2025.nllp-1.4. URL https://aclanthology.org/2025.nllp-1.4/.

Lorena Amanda Quincoso Lugones, Christopher Kverne, Nityam Sharadkumar Bhimani, Ana Carolina Oliveira, Agoritsa Polyzou, Christine Lisetti, and Janki Bhimani. Aurora: Neuro-Symbolic AI driven advising agent. arXiv preprint, 2026. URL https://arxiv.org/abs/2602.17999.

Denis Merigoux, Nicolas Chataing, and Jonathan Protzenko. Catala: A programming language for the law. *Proceedings of the ACM on Programming Languages*, 5(ICFP), 2021. doi: 10.1145/3473582. URL https://arxiv.org/abs/2103.03198. Article 77.

Yixuan Mi, Yiduo Yu, and Yiyi Zhao. SmartCourse: A contextual ai-powered course advising system for undergraduates. arXiv preprint, 2025. URL https://arxiv.org/abs/2507.22946.

Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. GAIA: a benchmark for General AI assistants. arXiv preprint, 2023. URL https://arxiv.org/abs/2311.12983.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *ACL*, 2020. URL https://aclanthology.org/2020.acl-main.442/.

Sergio Servantez, Sarah B. Lawsky, Rajiv Jain, Daniel W. Linna Jr., and Kristian Hammond. Openexempt: A diagnostic benchmark for legal reasoning and a framework for creating custom benchmarks on demand. arXiv preprint, 2026. URL https://arxiv.org/abs/2601.13183.

Jaeho Shin, Sen Wu, Feiran Wang, Christopher De Sa, Ce Zhang, and Christopher Ré. Incremental knowledge base construction using DeepDive. *Proceedings of the VLDB Endowment*, 8(11): 1310–1321, 2015. URL https://arxiv.org/abs/1502.00731.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. YAGO: A Core of Semantic Knowledge Unifying WordNet and Wikipedia. In *WWW*, 2007. URL https://www.iw3c2.org/papers/2018-YAGO.pdf.

Qiang Sun, Yuanyi Luo, Wenxiao Zhang, Sirui Li, Jichunyang Li, Kai Niu, Xiangrui Kong, and Wei Liu. Docs2KG: Unified knowledge graph construction from heterogeneous documents assisted by large language models. arXiv preprint, 2024. URL https://arxiv.org/abs/2406.02962.

Harsh Trivedi, Tushar Khot, Mareike Hartmann, Ruskin Manku, Vinty Dong, Edward Li, Shashank Gupta, Ashish Sabharwal, and Niranjan Balasubramanian. AppWorld: A Controllable World of Apps and People for Benchmarking Interactive Coding Agents. In *ACL*, 2024. URL https://aclanthology.org/2024.acl-long.850/.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. BrowseComp: A simple yet challenging benchmark for browsing agents. arXiv preprint, 2025. URL https://arxiv.org/abs/2504.12516.

Chenxi Whitehouse, Clara Vania, Alham Fikri Aji, Christos Christodoulopoulos, and Andrea Pierleoni. WebIE: Faithful and robust information extraction on the web. arXiv preprint, 2023. URL https://arxiv.org/abs/2305.14293.

Ryan Wong, Jiawei Wang, Junjie Zhao, Li Chen, Yan Gao, Long Zhang, Xuan Zhou, Zuo Wang, Kai Xiang, Ge Zhang, Wenhao Huang, Yang Wang, and Ke Wang. WideSearch: Benchmarking agentic broad info-seeking, 2025. URL https://arxiv.org/abs/2508.07999.

Chen Yang, Ruping Xu, Ruizhe Li, Bin Cao, and Jing Fan. Business as rulesual: A benchmark and framework for business rule flow modeling with llms. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 35198–35220, 2026. doi: 10.18653/v1/2026.acl-long.1625. URL https://aclanthology.org/2026.acl-long.1625/.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2406.12045.

Ruiqi Zhong, Tao Yu, and Dan Klein. Semantic evaluation for text-to-SQL with distilled test suites. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 396–411, 2020. doi: 10.18653/v1/2020.emnlp-main.29.

Ruiwen Zhou, Wenyue Hua, Liangming Pan, Sitao Cheng, Xiaobao Wu, En Yu, and William Yang Wang. Rulearena: A benchmark for rule-guided reasoning with llms in real-world scenarios. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 550–572, 2025. doi: 10.18653/v1/2025.acl-long.27. URL https://aclanthology.org/2025.acl-long.27/.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. Webarena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations (ICLR)*, 2024.

Yichao Zhou, Ying Sheng, Nguyen Vo, Nick Edmonds, and Sandeep Tata. Simplified dom trees for transferable attribute extraction from the web. arXiv preprint, 2021. URL https://arxiv.org/abs/2101.02415.

Appendix map. A protocol freeze and denominators; B limitations of the measured sections; C conditions behind every table, including the enforced budgets and the cluster counts; D construction and reference inventories; J language-model use; K per-layout and cross-model results; P component ablations; R evaluator and reference diagnostics; N completion claims; O contract failures; S external cohorts; U related work, row by row; Y evaluable coverage; L deterministic baseline.

A Protocol Freeze, Denominators and Re-runs

One runtime snapshot serves the whole grid, and every cell pins the prompt, source, reference, and scorer it ran against. Denominators are reference-defined and fixed before scoring: a cell is scored against the obligations its reference declares, not against what the prediction happened to return. An output that violates the contract—an unparseable or truncated answer, an exhausted call or context budget—is scored as the failure it is and stays in the denominator rather than being dropped or resampled. No cell is ever re-run because of its outcome; a cell whose request ledger records a transport error is re-run because it produced no outcome at all, and a cell's official result is its first terminal attempt with a clean ledger, with every other attempt preserved (Appendix I).

One scorer behind every number. Each scoring run pins the hashes of 9 evaluator files, of which 8 are in the closure the scorer actually imports; that closure is 88 files, so 80 of them are not pinned and 3 of them changed after the runs. Cross-run comparability therefore does not rest on the pin. It rests on the measured drift: complete re-scorings under both evaluator builds, diffed field by field. Some runs completed after the evaluation code had been edited, which is a version switch and not a licence to mix results, so the two builds were compared by re-scoring complete views—not samples—under each of them and diffing every scored field: 5 such re-scorings over 5265 cells changed 0 of 1,326,429 scored values, and the per-cell records were byte-identical. One bookkeeping limit of the pinned re-scorings belongs here: the wrapper that scores a run under the pinned evaluator passes no manifest through, so those runs' scoring records name the frozen grid manifest rather than the execution plan of the lane that produced the predictions. The evaluator hashes, the reference set and the bootstrap settings are unaffected; what is missing is a pointer, and the lanes' own intake records carry it. Two readings of that zero must be separated. It does *not* say the two builds are equivalent programs—they differ by hundreds of lines—but that on every prediction these runs actually contain, no scored field is distinguishable between them, which is the only property a cross-run comparison needs. What that root pins, and what it does not, belongs beside those re-scorings. The score root pins the scorer code by real copy—9 scorer files whose hashes are recorded per run under `scorer_files_sha256` in the run's scoring record—but the reference set, the world roster, the page bundles and the prompt tree are *not* pinned by that root—the root reaches all four through junctions into the live tree, so the statement elsewhere in this appendix that every cell pins the prompt it ran against holds in the execution plan, which records that pin, and not in the root directory: they are read from absolute paths into the live evidence tree, and their traceability comes from input hashes instead, namely the candidate-summary hash of the reference build, which records that candidate as not complete, together with each run's experiment-manifest and runtime-snapshot hashes. The page bundles and the table-rendering script carry no hash in the *scoring record* at all; that is a real gap in that ledger, and this paper states it rather than closing it. It is a different ledger from the one the Reproducibility Statement names: this paper's own input manifest hashes every input each printed table is generated from, which is why a table input can be hashed there while the run's scoring record carries no hash for the bundle it was scored against. Comparability therefore rests on the measured zero drift above and on those input hashes, not on the isolation of the root directory. Runs re-scored under that root reproduced their published values field for field. The deviation log records the switch, the re-scorings and the affected runs.

B Limitations of the Measured Sections

The construction pool reflects accessible institutional sources and is not a representative sample of all universities. Controlled renderings preserve selected content but do not capture every property of original websites; the original-website condition covers 234 worlds and its snapshots are subtrees of the live sites, with outbound links that the snapshot does not always resolve; no unresolved-link rate is stored, so none is quoted (Appendix D.1). Exact equivalence is restricted to supported record

domains and trusted context facts, including the single-attempt assumption of temporal checks. Unresolved reference semantics and interactions such as global credit allocation limit full-world certification: in this scorer a reference carrying a world-credit contract can only return `FAIL` or `UNKNOWN` strictly, leaving strict certification attainable on 18 of 243 worlds; deciding those contracts is a scorer change for a later version, with its own deviation entry and a full re-scoring (Appendix R.1). Historical diagnostic scores remain specific to their source and evaluator configurations. The measured sections carry four further limitations—one execution per plan, diagnostic references, the floor under behavioral agreement, and the fact that the discovery-versus-interpretation contrast exists 4 times, on two scaffolds and two models in one condition, of which 2 improve anything—stated in full below.

Three of the limitations below are quantities rather than cautions, and they are stated once here. Every interval rests on at most 48 school clusters and in places on 3 (above). No token or call ceiling was enforced in any cell: the budget the runtime reads is the page budget plus one per-scaffold bound, and the manifest's 40-call and token budgets are preflight intent, null in all 1,170 receipts. The pinned scorer root covers 9 of the 88 files in the effective scoring closure, so cross-run comparability rests on the measured zero-difference re-scorings and not on the pin alone.

The measured sections carry four further limitations. Every plan was executed once, at one seed, through a serving path with no sampling control; each of the three layouts now carries a second execution of the same plan, and that repetition shows run-to-run variation of the same order as many of the differences reported, so every cross-model comparison is annotated with whether it exceeds its own layout's floor. One repetition per layout is one draw of that variation, not its distribution. Scores are computed over 243 of the 252 test worlds, and 180 of those carry diagnostic references whose residual fields are typed `UNKNOWN` and therefore excluded from the decidable denominator rather than charged to a prediction; the reference checkpoint itself contains author rulings, and fields blocked by the frozen worksheet's one-scalar-per-field shape remain non-executable. Behavioral agreement is a decision-level measure with a non-trivial floor, and a reconstruction whose verdict mix happens to match the reference's base rate can score above that floor without recovering the rules, which is why the decision composition accompanies it everywhere. Finally, the contrast between locating evidence and interpreting it is measured once and narrowly. The privileged-source arm declared by the original frozen manifest was found, on audit of its own receipts, to have handed the agent the site's full page list rather than the evidence pages; its results are reported nowhere. Later arms do hand over the declared evidence pages, and their receipts confirm it on every cell that produced one, but the contrast exists for two scaffolds on two base models in one condition—4 pairs, of which 2 improve anything (Appendix V), it is an oracle that depends on the reference by construction, its structured half is conditioned on the worlds that arm could finish, and it says nothing about the re-rendered layouts, where discovery is trivial by construction. The two scaffolds disagree, which is itself the result: the difference between them is read, not estimated, because they differ in capacity too.

C Conditions for the Measured Tables

Scales. A metric is printed on one scale per table: the comparison and per-layout tables in percent, the one-field contrast documents' differences in the metric's own proportion, and a difference between two percentages as percentage points where it is labelled *pp*. The same quantity therefore appears in more than one scale across body and appendix, and each place that reports one states the scale it uses.

What the intervals are, and are not. Every interval in this paper is a percentile bootstrap over school clusters with $B = 10{,}000$ resamples and seed 20260901. The number of clusters an interval rests on is recorded per estimate by the scorer: across the runs reported here it ranges from 3 to 48 clusters, the main comparison's own estimates rest on 6–48, and the one-declared-field contrasts on 39–45. The smallest counts are the Certification risk rows of the structured arms, where the metric is defined on a handful of schools (as few as 18 scored units). Percentile cluster bootstraps with so few clusters are anti-conservative, so those intervals are indicative rather than inferential. The beyond/within labels of the per-layout and cross-model tables are descriptive counts of uncorrected simultaneous 95% statements: no multiplicity adjustment is applied anywhere in this paper, and "beyond" is not a test at a family-wise level.

Table 2: Construction inventory for the archived September 13 snapshot. Pages and evidence placements are aggregated over admitted renderings.

Quantity	Count
Candidate curriculum worlds / institutions	735 / 111
Admitted / audited renderings	2,931 / 2,940
Worlds admitted under all four layouts	726
Pages across admitted renderings	109,905
Evidence-sentence placements	557,601

Conditions for the tables of Section 6. One seed and one execution per plan, except for the departmental layout under one model, which was executed twice so that run-to-run variation could be measured; the pre-registered minimum of three seeds for stochastic agents is not met, so no seed-sensitivity claim is made anywhere. Sampling temperature, seed and token budget are declared by the protocol but not enforced by the serving path. Each model was scored in its own run against the same references, prompts, budgets and scorer; layouts of one model and condition—and, for behavioural agreement, scaffolds of one model—may be pooled, while models, conditions and serving paths never are. Scored worlds: 243 of the 252 test worlds, the excluded world being the one whose reference declares itself outside this scorer's contract, excluded identically for every system. References: of the 252 built, 63 are complete and 189 diagnostic; of the 243 scored, 63 are complete and 180 diagnostic. A diagnostic reference bounds a world's obligations rather than closing them. Contract-violating answers are scored as the failures they are and stay in the denominator. Two scorer versions appear in this paper. Table 1, Figure 2, and the numbers of Sections 6 and 7 that derive from them use the current scorer, which aligns obligations by identifier and content and reads answers under the whitespace-tolerant contract of Appendix Z. The appendix tables were produced by the archived scorer with its earlier alignment and the strict contract, so their semantic accuracy and omission values are lower than the corresponding entries of Table 1 and are not interchangeable with them. Entity F_1, behavioural agreement, and certification coverage do not depend on obligation alignment and agree between the two. Two scoring-side corrections were applied after the execution freeze and changed no previously computed value: the table renderer's condition names were mapped onto the names the scored rows carry, and the scorer was made total so that one unparseable prediction can no longer remove the other cells of its world; a regression guard compares all 388,692 previously computed field values before and after, with none changed. Ninety-nine malformed relation items in forty-five browsing cells are dropped from the executable prediction and counted; treating those forty-five cells instead as wholly unusable answers moves only that row, by -5.1 points of entity F_1, $+0.6$ of omission and -0.2 of semantic accuracy.

D Versioned Construction and Reference Inventory

Each diagnostic is associated with a source report, reference configuration, and recorded hash. The 735-world construction pool, the 450-world reconstruction checkpoint, and the 708-world baseline sets describe different stages and episode selections. Comparisons across those stages require a matched manifest rather than substitution of one count for another.

Table 3: Reference-reconstruction inventory at the accepted but incomplete checkpoint reference. Counts describe reconstruction and validation operations.

Quantity	Count
Rebuilt worlds / machine-validated worlds	450 / 99
Retained / excluded objects	49,896 / 13,986
Original / subsequently accepted natural-language fields	8,586 / 36
Accepted natural-language fields / converted fields	8,622 / 7,434
Semantic residual items / overlapping field-contract errors	1,251 / 45
Reviewed policy records / executable extension fields	2,439 / 1,170
Requirement conditions / reviewed open-pool requirements	54 / 9

A condition such as standing or eligibility additionally depends on trusted context facts, rather than on an inferred course-number threshold. The collection pool, reconstruction subset, and evaluated episode sets therefore have distinct denominators; rendering a world does not establish its semantic reference validity. The reference checkpoint is accepted but not complete. It records `final_reference_complete` as false for three reasons: residual fields that await a human rul-

ing, fields blocked by the frozen worksheet's one-scalar-per-field shape—whose blocks are mostly governed by reviewed policy contracts already—and the enumerated grammar gaps listed below.

Entity reconciliation links repeated course references while preserving differences in policy year, audience, and requirement scope. These associations let a reviewer trace an executable rule back to the text that licenses its operands and conditions. Structural validation checks identifiers and field contracts, whereas the behavioral cases check particular interpretations.

The 99 machine-validated worlds establish field-level validity for those worlds, not independent semantic completeness of the 450-world checkpoint. The 1,251 residual items include issues in structured fields as well as unconverted natural-language content, so their count is not the $8{,}622 - 7{,}434 = 1{,}188$ unconverted fields. The 45 field-contract errors overlap that residual inventory. Policy records, extension fields, requirement conditions, and open-pool requirements likewise denote different objects rather than disjoint sets of validated rules.

How the checkpoint was produced. Human annotation returns are the only source of an annotation decision. Each field a return left in natural language is packaged as a job with its source receipts, and a language-model translator proposes a per-field translation into the executable grammar; every proposal carries the quoted span, its normalized offsets, and a SHA-256 of the file it was read from. A different language-model agent then reviews each proposed field independently and returns `ACCEPT`, `REJECT`, or `HOLD`; only `ACCEPT` is merged, and a `HOLD` is a question for a human rather than a value. The first translation pass made 4,149 proposals and was reviewed at 3,807 accept, 243 reject, and 99 hold; every later batch was reviewed the same way, with 2 repair batches the exception: they carry a batch report but no archived independent review, so no review disposition is claimed for them. The accepted set is then rebuilt across all 450 worlds, sealed, and put through seven pipeline replay gates, an exact-change root verification that confirms the worksheet fields changed are exactly those proposed, and 20 temporal scenarios, before the checkpoint is accepted. No language-model agent decides a human ruling: where a ruling is undecided the field is encoded as `UNKNOWN` and is excluded from the decidable denominator.

Why the checkpoint is not complete. `final_reference_complete` is false, and the three recorded reasons are distinct. First, 1,251 fields carry a residual that needs a human ruling—an ambiguous operand attachment, a course class the page defines twice, a number the source prints but no adjudicator has ruled on. Second, roughly 315 fields are blocked not by the grammar but by the frozen worksheet's one-scalar-per-field shape, which cannot hold a range, a set, or "no limit" in a field the contract vocabulary already supports; most of their blocks are governed by reviewed policy contracts. Third, an enumerated set of grammar gaps remains deliberately unexpressed, among them a course-set (pair-level) equivalence, an entry-type guard inside a rule, a band-triggered or applicability-guarded credit exclusion, a rule-level non-reuse condition, a guarded corequisite, a subject-scoped term limit, and attempt-count limits; each is recorded with the reviewer position that confirmed it, so a translator abstains deliberately rather than by oversight.

The 557,601 evidence-sentence placements count occurrences across admitted renderings, rather than unique policy facts. The 109,905-page total also includes alternative renderings of the same underlying content. The 726-world paired set retains only worlds admitted under every layout, making rendering admission part of the selection rule.

D.1 Data Collection Practices

Every network byte this benchmark ever requested from an institution went through one rate-limited request function, which sends a single header—a research User-Agent naming the study and pointing at the repository for contact—and no credential of any kind: there is no authorization header, no cookie jar and no login form anywhere in the fetch path, and the 72 hosts of the frozen original-website condition are public catalog, bulletin and departmental advising pages, with 1 exception the artifact enumerates: a vendor analytics endpoint serving 2 pages. The floor between two network fetches inside one process is one second, a host's own `Crawl-delay` or request-rate wins whenever it is larger, a request times out at thirty seconds, and a response over eight mebibytes raises instead of being stored. `robots.txt` is fetched once per host through the same rate limit and both scopes must permit a request—the study's own allow-list and the site owner's rules—with `Disallow` applied per URL before the request is made; a plain 4xx leaves a host open, and

any other failure to read `robots.txt` closes that host for the whole run. Two consequences are recorded rather than argued: a page already on disk is not re-fetched and therefore not re-judged, and no eviction pass exists, so a `robots.txt` that became stricter after a page was cached would not retroactively evict it.

Anything behind a bot challenge was recorded and never bypassed, anything a host disallowed was not fetched, and the recorded stance is that neither rule is relaxed under schedule pressure: a catalog platform family answering deep pages with a bot-challenge body was ruled out of scope, a host serving `Disallow: /` over its course text was ruled out, one host with a full ban was left alone, and the 3 *blocked* verdict surviving in the school ledger is a host whose root was allowed but whose platform signature stopped the crawl at the scout. Across processes the one-second floor was *not* coordinated: school-level runners execute three to four schools concurrently, each child process keeping its own one-second floor, which is one second per host only because different schools are different hosts. The repository records one failure of that assumption, and it is disclosed here rather than left in the logs: on 2026-09-07 two overlapping queues started two runners for the same school, and for that period two processes each requested that one institution's server about once per second. The duplicate runner was stopped and the recorded remedy is a standing rule—one runner per school, queues must not overlap.

Scale, as the ledger records it: 777 school records over 774 distinct catalog root hosts, of which 183 rows store a `robots.txt` verdict string—including 18 hosts recorded as having no `robots.txt` at all—while the remaining rows do not carry the check's result, which is a documentation gap and not evidence that the check was skipped. 234 rows carry a crawl start time, spanning 2026-09-06 to 2026-09-07; a later root-index seeding pass over further schools is documented in code rather than in a ledger field, and no single machine-readable record of the full crawl date range exists. What is stored is decoded text, not received bytes: each cached page is written under a hash of its canonical URL, decoded as UTF-8 with replacement, and no HTTP response header is kept. The frozen original-website condition is exactly such a cache—48 institution directories holding 13,053 files, all HTML, XML or JSON, with no PDF, image or other binary—whose manifests describe 24,165 page entries across 72 hosts, and whose scored part is the 234 worlds and 22,368 pages of Section 6. No unresolved-outbound-link rate is stored anywhere in that evidence, so this paper quotes none. The authorization under which these pages were collected is not documented in the current evidence package, and nothing is asserted about it.

E Output Contract and Worked Cases

Two websites can agree on every page an agent has visited while differing in an unvisited policy page, and an agent cannot distinguish them from its observed history alone; this is why a completion claim is evaluated against a fixed reference boundary. The scope identifies a target program or framework, applicable general-education obligations, prerequisite closure, policy families, and admissible source domains. A referenced course whose definition lies outside the permitted snapshot remains unresolved until the agent obtains admissible evidence for it. The interface permits an agent to return a useful partial reconstruction without equating partial progress with completeness. Student records enter the task as evaluation inputs for the recovered rules; an individually optimized course schedule is not the required output.

The reporting protocol includes world-level macro-averages as well as pooled counts, so that large curricula do not silently determine the aggregate.

The decision d distinguishes `CERTIFY_COMPLETE` from continuing exploration, requesting more budget, or declaring an obligation unresolvable. The same typed representation accommodates supported choose-k groups, explicit credit pools, exclusions, and prerequisite conditions, so that extracting operands and recovering their composition remain distinguishable operations.

The output contract separates identity, executable content, applicability, evidence, and the status of unresolved work. This separation allows the same reconstructed rule to be checked for structural validity, source support, and behavior without conflating those checks.

Logical grouping. The reference `ALL(A, ANY(B,C))` and prediction `ANY(ALL(A,B),C)` agree on a record containing both A and C. They disagree on the record containing only C, which

Table 4: Information retained by the output contract.

Component	Information
Identity and scope	Institution, catalog year, target program/framework, source boundary
Entity	Stable identifier, title, source-stated attributes
Constraint	Operator tree, operands, thresholds, exception scope
Applicability	Program, cohort, grade, temporal or eligibility context
Evidence	Snapshot identity, source location, span, associated claim
Unresolved item	Original wording, missing fact or unsupported construct, reason
Completion	Decision, confidence, open obligations, resource usage

satisfies the prediction without satisfying the reference. The informative test is therefore a decision-disagreement witness, rather than an arbitrary record containing the shared course identifiers.

Temporal grouping. The temporal domain assigns each course one of `missing`, `completed`, `before`, `current`, or `future`, with the target course fixed in the current term. An ordinary prerequisite requires prior completion, co-enrollment permits prior or current enrollment, and strict same-term enrollment requires the current term. The implementation preserves an outer same-term restriction even when its child permits ordinary co-enrollment. Repeated attempts, cross-listed identities, and variable-credit temporal expressions fall outside the general equivalence claim.

Applicability. A majority-of-courses condition and a majority-of-credits condition can disagree on the same course set because they aggregate different quantities. Similarly, a rule triggered by a change of grading mode has a narrower scope than an unconditional restriction on the destination mode. Such conditions require explicit bindings and context facts; a successful execution on supplied facts establishes a conditional result.

F Semantic Outcomes and Behavioral Checks

The reference inventory and alignment policy are fixed independently of a system's semantic scores. Requirement omission and entity recovery are not independent evidence: omission is $1 - \min(\rho, \pi)$, so wherever course recall ρ binds it is a monotone function of the same coverage entity recovery reports, and a movement that appears in both is one measurement twice. Citation support is a contract-compliance check rather than a discriminating metric as reported here: only two of the five scaffolds have a claim-support audit in these runs, the other three record none, so its precision has no denominator there, and the two that do sit at the two ends of the scale. Citation support is a precision over the claims a prediction makes, not a recall over reference citations, and it records which of the four routes each claim took, because a claim supported only through an equivalence or a band is supported by a route the prediction does not say it used. It does not check the year, audience, or exception scope of the cited passage; a citation that resolves to a valid URL is insufficient when the cited passage does not name the course the rule requires.

Reference obligations are aligned to predictions before inspecting semantic outcomes. A declared normalization or reviewed mapping handles a reference requirement represented by several predicted rules. The outcome accounting in Table 5 then separates failures of the prediction from limits of the reference or evaluator.

Following the distinction between minimum-functionality, invariance, and directional tests in CheckList (Ribeiro et al., 2020), a minimal rule case targets one logical or applicability condition. An invariance check changes an irrelevant identifier or presentation feature while preserving the intended rule behavior. A directional check changes a condition for which the direction of the expected decision is established in advance. Adding course credits is not universally such a change: caps, exclusions, and double-counting restrictions can break monotonicity.

Sampled student records measure agreement on a specified test distribution, whereas exact checks search the supported domain for any disagreement. Neither establishes correctness outside its declared domain. Moreover, independent local checks do not determine whether a course can satisfy several requirements at once or how credits must be allocated globally. An unsupported interaction

Table 5: Outcome accounting for the reference-defined semantic denominator.

Outcome	Meaning	Treatment
Equivalent	No witness in the supported domain	Success within that domain
Counterexample	A witness passes independent replay	Incorrect; retain error direction
Predicted abstention	Evaluable reference; absent/unresolved prediction	Non-success in the fixed denominator
Prediction unsupported	Evaluable reference; unsupported prediction	Non-success in the fixed denominator
Reference unknown	Reference semantics or trusted facts unavailable	Explicit reference coverage limitation
Timeout	Search unresolved at the resource limit	Undecided; retain rate and handling
Evaluator disagreement	Solver result conflicts with verifier replay	Evaluator failure requiring investigation

therefore leaves the corresponding full-world claim unresolved unless the scope contract explicitly excludes it.

Completion precision and conditional risk are defined only among certified episodes with adjudicable correctness labels. Risk–coverage thresholds are selected on development data, with unadjudicated labels and zero-certification cases reported separately. An always-abstain control and the obligation ledger's own closure test provide comparisons for confidence-based certification. Resource use is reported as the pages opened and the model calls charged against each cell's enforced budget—12–43 pages per cell, median 17, and per scaffold a 200-call ceiling for the structured ledger, 40 steps for the browsing scaffold, 600,000 prompt characters for one-shot and one answering call for each retrieval scaffold—so budget exhaustion is observable as a spend against a declared allotment. The manifest's 40-call and token budgets are preflight intent that the runtime does not read, and no token ceiling was enforced in any cell. We do not report a spend-the-budget stopping baseline: a recorded stopping decision carries the pages and calls of the whole episode, not the fraction of budget already spent at the moment the decision was made, and a baseline that certifies on budget exhaustion would certify every episode under that encoding.

G Diagnostic Results

Table 6: Mutation and reverse-rewrite audit of the evaluator against the reference checkpoint's executable reference rules.

Checkpoint worlds audited / without a built reference	432 / 18
Executable reference rules / generated mutants	8,181 / 22,914
Caught / undecided mutants	18,675 / 4,203
Decided but not distinguished, independently verified meaning-preserving	36
Evaluator disagreements	0
Undecided: outside the exact fragment	4,203
Reverse rewrites: generated / decided equivalent / unknown	14,265 / 10,791 / 3,474
Worksheet requirements present in references / total	10,728 / 10,935

The three mutant classes partition the generated total: caught, undecided and decided-equivalent sum to the generated count in the table above. The evaluator decides $18711/22914 = 81.66\%$ of all generated mutants and catches $18675/18711 = 99.81\%$ of those decided, the critical mutation score quoted in Section 5.2. The decided denominator, 18,711, includes 36 mutants independently checked to preserve meaning, so the distinction rate on the decided, meaning-changing subset is 100%. That conditional rate neither resolves the 4,203 undecided cases, all of which fall outside the exact fragment, nor establishes coverage of all possible semantic errors. The battery evaluates program requirement rules only: the checkpoint's general-education blocks, policy blocks, and reference-extension fields are hashed and inventoried but are never executed by it, and unresolved requirements are untested by construction.

Engineering verification for the same checkpoint records 2,049 tests passed and 1 skipped with source hashes unchanged during the run, alongside 164 tests covering the reference grammar con-

structs added for this checkpoint. These are code-level checks; they establish neither reference completeness nor any model result.

Table 7: Worksheet agreement on ten matched worlds in the archived annotation diagnostic.

Pair type	Pairs	Mean agreement
Verify–verify	10	0.9208
Verify–blank	40	0.8722
Blank–blank	10	0.8773

Verification worksheets present pre-filled content, whereas source-only worksheets begin without that proposed reconstruction. Pairs can share a world or annotator worksheet, so they are not independent observations. The archived world-resampled interval for the verify–verify minus verify–blank difference is 0.0272–0.0707. Annotation assignment, segmentation choices, and other differences can contribute to that association, preventing a causal attribution to pre-filling alone.

G.1 Annotation Protocol, as the Records Establish It

As issued. Annotation was governed by a single written guideline, issued in one version with two dated in-document amendments, and one packet-format revision. The guideline fixes the frozen source pages and their evidence spans as the only admissible source of fact, forbids approximating a rule that cannot be written in the grammar—such a field is quoted verbatim instead, because an approximate rule makes the evaluator confidently wrong—requires a status value on every block from a five-value vocabulary enforced in code, and restricts editing to field lines, the status and the note. Worksheets were issued in two modes: a verification mode that pre-fills a machine reconstruction for the annotator to confirm or correct, and a source-only mode whose rule lines are empty. The guideline states its own reason for that split—verification is faster but leans toward the machine—and reserves source-only mode for the hidden split. Mode is recorded per worksheet: of the 1,071 returns that carry a provenance record, 891 are verification mode and 180 source-only, and 1,071 of 1,071 attest a human return. That attestation is a statement by the authors on a date; it does not establish who performed the labeling. Those records cover 131,364 worksheet blocks, whose recorded statuses are 60,192 corrected, 35,820 confirmed, 8,712 added, 26,352 removed and 288 uncertain; source-only slots record no confirmation at all, which is the expected signature of a worksheet with nothing pre-filled to confirm.

Prescribed process, and what the records show of its execution. The guideline prescribes a calibration session on a shared world before independent work, restricted to notation rather than page facts; independent work thereafter, with annotators of the same world instructed not to confer and not to read each other's files; a daily one-line log per annotator; a pre-submission self-check that no block is left without a status and every uncertain block carries a note; and adjudication by someone who did not annotate that world, who sees only the evidence spans and the frozen source and not who produced which label. What the records establish is the mode and the status distribution of every returned worksheet, and nothing about the sessions: 81 of the 1,863 worksheet slots carry a one-line log, none of which records hours, and the packet's per-person hour figure is a workload projection computed from a per-block minute rate, so this paper prints no annotation hours. For adjudication the record is split and both halves are needed: the procedure is written down in the guideline, while its execution and its products are not recorded—no adjudicated worksheet directory, no per-disagreement record, and no adjudicator field exists in the current evidence package.

Participants, as the records permit. Project assignment records contain 15 annotator identity tokens (A–O) over 1,863 worksheet slots in 450 worlds. Of those slots, 1,071 carry a provenance record attesting a human return; the remaining 792, falling into 378 within-world groups, are byte-identical to another slot of the same world once the annotator token in the header is neutralized—the property the packet ascribes to the copies as issued—so they are worksheets that were issued and not returned. The tokens are assignment slots rather than verified person identities, and they are reused across rounds: tokens A–E appear in 180 to 207 worlds each, more worlds than the first packet contained. Supplementary review packages—a 6,489-record residual review over 414 worlds, and a manual audit carrying 7,110 pending rows over 432 source worlds—carry no identity or slot binding in their recorded metadata, so they cannot be reconciled with the A–O tokens.

Agreement, and what it is not. Worksheet agreement is computed by a comparator that canonicalizes each rule before comparing, so two orderings of the same disjunction agree while two bracketings of the same operands do not, and a block present for one annotator and absent for the other counts as a compared disagreement. Every pair drawn from the slots without a provenance record compares an issued worksheet with a copy of itself, which is why 459 such pairs agree exactly; any agreement claim therefore excludes them, which removes 2,277 of the 3,231 within-world pairs and leaves 954. The 10 matched worlds of the table above already exclude them; beyond those worlds the verification–verification row extends to 504 pairs over 423 worlds at mean 0.9164. These are agreement rates over a fixed field inventory with no chance correction: they are not reliability coefficients, and this paper reports none. No field of the open-field residual set was annotated twice independently: a blind overlap of 99 of the 459 test-split fields was prepared, stratified one per field type, and no second annotator returned it, so no inter-annotator agreement exists for that set.

Where the machine proposed and a human decided. Over the 891 verification-mode returns, 35,820 of the 89,163 pre-filled blocks were kept unchanged (0.4017, or 0.3005 of all 119,196 status-bearing blocks); 53,343 were corrected, 6,606 blocks the pre-fill lacked were added, and 23,139 were removed as not being rules at all. This is a pre-fill acceptance rate and not an accuracy rate, because the guideline instructs the annotator to correct an enumerated list of known extractor defects on sight. At field level, over the 459 open residual fields of the test split, the proposed formalization was accepted in 432 cases on first return and in 396 after the same person's own final review, with 405 of 459 decisions identical across the two returns—that is one rater's test–retest stability, never agreement between two annotators. 1 prerequisite-grouping ruling is stored as a hash-bound acceptance receipt and applied in 18 worlds; the checkpoint's adjudication count counts those world-level applications of that one ruling rather than two rulings.

Annotation process and documentation. Annotators were recruited through a survey, with one annotator assigned to each annotation package, and the annotators are acknowledged; these three facts are confirmed by the authors. Project assignment records contain 15 main-reference annotator identifiers (A–O); supplementary audit reviewers have not yet been reconciled with that roster, so this is not a total participant count, and the current evidence package does not establish one. Humans performed the returned labeling; the authors report that GPT-6 Astra Pro checked for unfilled entries. The following are not documented in the current evidence package and are not asserted anywhere in this paper: annotator qualifications, annotator independence, annotator exposure to prefilled values, annotator training, annotation hours and compensation, research-use consent, and—as Appendix G.1 states of the adjudication procedure—the execution and products of adjudication, together with the identity of any adjudicator. The available assignment records therefore do not establish those properties or the accuracy of the final reference. One annotation–source conflict was adjudicated by the authors rather than by an annotator: a Cal Poly worksheet field read "letter grading required" where the frozen source only forbids selecting major, support and concentration courses as credit/no credit. The authors ruled for the source's narrower reading; the reviewed contract therefore fails credit/no credit, passes letter grading and audit, and returns `UNKNOWN` for any grading mode it does not name. That ruling is recorded in the binding layer and hash-bound to the frozen source page rather than to the decision: unlike the prerequisite-grouping ruling of Appendix G.1 it carries no acceptance receipt of its own. The original annotation is unchanged, and the ruling is not presented as a statement by the institution.

H Experimental Controls

Ledger ablation removes induced obligations and learned yields from exploration decisions while keeping source-access tools and the final closure audit fixed. Evidence ablation removes provenance-obligation induction during exploration; final source-support checks still apply to the returned claims. Semantic-feedback ablation masks source-validator errors in parser and repair requests while retaining retry state, validation, and budgets; it neither queries nor removes feedback from a held-out world scorer. Applicability ablation replaces supported executable scope conditions with source-backed unresolved representations, preserving the underlying requirements and evaluating unresolved applicability as unknown. Each run records observed unsupported scope forms and unavailable source spans; such a run cannot be silently counted as a completed applicability intervention.

Layout and identifier interventions pair runs on the same underlying world, with identifier mappings checked for consistent source and rule transformation. Page removal, link removal, and budget restriction are recorded separately because they change information availability or access in different ways. The reporting protocol retains failed runs, undecided evaluations, and missing cells, together with decoding, retries, context truncation, and cost. Principal comparisons and exploratory analyses remain separate, with uncertainty clustered by institution.

I Execution Log and Deviations

Frozen before any cell ran: the experiment manifest, its preflight result, the applied school split, the runtime snapshot with its own manifest, and a per-step execution plan that hashes every runtime file, prompt, packet, and cell configuration, so that a resume refuses to run if any input changed. Nothing in those files changed after the first cell ran. Table 8 lists everything that did not go as planned in the comparison grid's own execution, in order; the post-freeze extensions log their own deviations where they are reported (Appendices A, S, V and Z). Two entries bear on numbers printed here and are named where they bite rather than only in the table: the budget fields the runtime never enforced, and the nine original-site cells lost to a harness defect in HTML-to-text conversion. This is an execution record.

Table 8: Execution deviations for the comparison grid, in order of occurrence.

#	What happened	Disposition
1	The first launch stopped at 378 of 5,544 executions: concurrent repository edits changed hashed inputs and the plan correctly refused to continue.	Run abandoned and preserved; relaunched from an isolated byte snapshot of the runtime tree. No cell of the abandoned run is scored.
2	A five-hour subscription limit made every model call return HTTP 500 from the serving shim; with transport retries disabled, the grid marched through its schedule writing terminal failures.	3,420 cells whose failure was a refused call were quarantined and re-run on the same plan; 108 failures that carried HTTP-200 answers were kept as outcomes. Concurrency lowered.
3	A weekly subscription limit reproduced the same failure mode, with no work possible for about 32 hours.	2,970 cells quarantined and re-run; 99 kept as outcomes.
4	A stale shim process from event 2 still held the serving port, so the resumed grid was served by the old process at lower concurrency.	All processes stopped, leaving in-flight cells pending with no receipts written; one shim per port under a supervisor thereafter.
5	An audit of events 2 and 3 found fifty-four quarantined cells whose single request was an HTTP-200 answer that the runner rejected as invalid JSON or a worksheet-contract violation. They had already been re-run, and eighteen re-runs succeeded.	The scoring rule below was adopted before any final score existed. The fifty-four first attempts, all failures, are the official results; their re-runs are preserved and never scored.
6	The same audit found that the structured and ReAct runners skip a failed model call, so 261 cells straddling events 2 and 3 had finished "ok" with between 2 and 201 refused calls each. Earlier reviews had inspected only failed cells.	All 261 quarantined and re-run. The supervisors now sweep any terminal cell whose request ledger contains a transport error, at most twice per cell.
E1	*Post-freeze extension, not the grid.* The Gemma-3-27B-it cohort's run directory carries no record of the row map its published aggregate was rendered with: no applied-row-map document, and an aggregate that names the Graph-RAG condition by a key the cells spell differently.	The cells, scorer and reference are pinned and unchanged and no value in this paper was altered; what the directory alone does not reproduce is the mapping step behind that one published row (Appendix S).

Scoring rule. A cell's official result is its first terminal attempt whose request ledger contains no transport error. A rejected HTTP-200 answer is a benchmark outcome, not a transport failure: it is scored as loss 1, omission 1, and strict WorldPass `FAIL`, and it stays in the denominator. A transport failure never does, and never shares a denominator with a benchmark outcome.

Invariants. Through every event above, the model, prompts, budgets, renderers, packets, and scorer bytes of every scored cell are those of the frozen plan. No cell was ever re-run because

of its outcome. Every attempt ever made is preserved on disk, including the abandoned run and every quarantined attempt.

Limitation. Re-running a transport-failed cell is outcome-neutral only if provider refusals are independent of the cell's content. The refusals recorded here were account-level quota responses returned within about 1.5 seconds to every request, which supports that assumption; it cannot be proven from the logs.

J Language-Model Use in Building the Benchmark

Language-model agents were used in three separable roles, and the distinction between them matters for reading every number in this paper. *Drafting.* Agents assisted with the exposition and the LaTeX layout of this manuscript. Every passage was reviewed by the authors. *Reference construction.* As Appendix D records, verification-mode worksheets were pre-filled by a machine reconstruction that annotators corrected, a language-model translator proposed a translation into the executable grammar for every field that an annotator returned in natural language, and a different language-model agent reviewed each proposal and returned `ACCEPT`, `REJECT`, or `HOLD`. Only accepted proposals were merged, a hold is a question for a human, and no language model decides a human ruling. A language model also checked the returned worksheets for unfilled entries (Appendix G.1). These agents share no state with the evaluated systems. *Systems under test.* The models evaluated by the benchmark are separate runs executed under the frozen protocol of Section 6.1.

K Per-Layout and Cross-Model Results

BA must be read against a floor, because an empty program that certifies nothing already agrees with the reference on every record the reference rejects; that floor is 0.351 here. Beside every BA we therefore report the reference's own accept rate, each scaffold's predicted accept, reject and unknown shares, and the mean number of reference obligations the reconstruction decides wrongly (Figure 5b,c); BA alone can be earned without recovering any rule, so it is never reported on its own. A higher certification rate is beneficial only when read together with the error rate among those certifications (Table 16 in Appendix N reports each arm's own stopping decision under Equation 2); on the original websites certification nearly disappears—Opus 5 certifies no episode at all—the appropriate response to a site the budget cannot read, but one that leaves conditional risk undefined rather than low. An audit computed from the snapshot and the references alone, before any model result was consulted, finds the evidence present and reachable from the root in 234 of 234 worlds, from a median of 6 evidence-bearing pages reached by opening 7 (Appendix M). What the protocol does not grant is the reading: that traversal requires a median of 782,773 characters, and 135 worlds exceed the fixed character cap of the one-shot bundle outright. The in-page-search execution changed two things at once—mechanism and page budget—so it is not an ablation of in-page search.

Table 9 reports the three base models per layout. Nothing in it is pooled across layouts, and the model that was executed on two of the three layouts is absent from the third rather than imputed. Table 10 gives the paired difference between the two models that cover all three layouts, computed within each layout over the worlds both scored.

Whether a difference is larger than this grid's own variation is a separate question from whether its interval excludes zero. Each of the three layouts was executed a second time under an identical plan, and Table 19 reports what changed between the two executions of each. An effect exceeds that floor when the 95 % interval of the paired difference-of-differences—the cross-model difference minus the repetition difference, over the same worlds—excludes zero. Every cross-model comparison in Table 10 is judged by that rule on its own layout, so no cell of its last column is left unassigned. A single repetition is one draw of the noise; failing to separate an effect from it is a failure to separate, never evidence that the effect is absent. The pattern is not uniform across metrics: the browsing scaffold's course recovery stays inside the floor on all three layouts while its rule semantics exceed the floor on all three, and every strict world-success difference stays inside it.

Table 9: Three base models, reported per layout and never pooled across layouts (%). Claude Haiku 4.5 was executed on the departmental and paginated layouts only, so it has no deep-navigation column and never enters a layout-pooled row. Each column comes from that model's own scoring run; paired differences and intervals are in Table 10. Strict WorldPass is instrument-bounded: it is attainable on 18 of the 243 scored worlds (Appendix R.1).

	Entity F_1 ↑			SemAcc↑		
Layout / scaffold	Haiku 4.5	Sonnet 5	Opus 5	Haiku 4.5	Sonnet 5	Opus 5
Department pages						
One-shot	44.9	63.9	75.6	15.7	26.3	56.5
BM25+LLM	42.2	53.2	61.4	17.6	19.5	40.3
Graph-RAG	44.5	64.8	75.7	16.6	33.9	67.8
ReAct-style browsing	62.6	76.5	76.8	18.7	23.7	61.3
Structured ledger	93.5	93.3	93.3	38.6	41.4	38.8
Paginated listing						
One-shot	49.6	51.4	72.7	6.7	26.1	51.0
BM25+LLM	34.4	51.3	71.2	6.3	32.4	53.5
Graph-RAG	44.6	49.4	75.8	15.6	32.7	57.9
ReAct-style browsing	70.3	70.5	80.2	18.9	20.8	39.9
Structured ledger	18.0	20.1	20.1	5.6	5.4	5.4
Deep navigation						
One-shot	–	64.8	79.7	–	42.3	56.3
BM25+LLM	–	55.8	52.6	–	30.8	50.0
Graph-RAG	–	57.5	78.5	–	34.8	62.2
ReAct-style browsing	–	74.5	71.9	–	25.9	61.5
Structured ledger	–	93.3	93.3	–	38.8	38.8

K.1 The Extended Tier: All Scored Worlds

Table 11 repeats the rows of Table 1 on all 243 scored worlds. The 180 diagnostic worlds raise Exec. Recall by 3 to 12 points per model and lower omission, because an unfinished reference lists fewer of the obligations that are hard to find (Table 31), and no ordering of Section 6 changes. The paired Claude comparison of Table 10, the repeated executions of Table 19, the behavioural agreement of Table 21 and the original-website condition are computed on this tier.

L The Deterministic Extraction Baseline

Table 12 is the deterministic extractor of Section 6 on the archived source-derived reference. It characterises that extractor; it establishes no ordering of layout difficulty for any other system. Across the associated deterministic runs, including diagnostic interventions, none of 2,970 stopping decisions certifies completion, so certification coverage is zero and conditional risk undefined. These observations characterise this extractor; they establish no model ranking and no general ordering of layout difficulty.

M The Original-Website Condition

Table 15 is the admission audit of the condition, computed from the frozen snapshot and the references alone and before any model result was consulted. Table 14 states the structural difference between that condition and the re-rendered layouts of the same worlds. The audit answers whether the condition is solvable in principle: the evidence exists in the snapshot, and a traversal that starts at the declared root reaches it in every world. It also states the cost of that traversal in characters, which is what the transplanted protocol does not pay for. A condition that is feasible on pages and infeasible on reading is a property of the budget rule, not of the websites, and the second execution of the browsing scaffold in Section 6 changes both at once, so neither change is isolated there.

Table 10: Opus 5 minus Sonnet 5, paired by world within each layout (percentage points, 95% school-clustered intervals). The last column says whether the gap is larger than what the *same* plan produces when it is simply run again on that layout: *beyond* iff the paired difference-of-differences against the repeated execution excludes zero, on the same worlds (Entity F_1 / SemAcc). Every layout now carries such a repetition, so no cell of this column is left unjudged; over the four metrics the floors report, 10 of 20 comparisons on department pages, 8 of 20 on paginated and 9 of 20 on deep navigation are beyond it. The structured row is the scaffold's deterministic parse rather than a model comparison, and its own repeat executions return byte-identical course lists in 252/252, 252/252, 216/252 of the worlds on the three layouts.

Layout / scaffold	Pairs	Δ Entity F_1	Δ SemAcc	Beyond the retest floor?
Department pages				
One-shot	243	+11.7 [-8.7, +35.0]	+30.2 [+6.3, +53.6]	beyond / beyond
BM25+LLM	243	+8.2 [-15.8, +29.3]	+20.8 [+1.3, +36.2]	beyond / within
Graph-RAG	243	+10.9 [+3.6, +21.1]	+33.8 [+20.4, +50.3]	beyond / beyond
ReAct-style browsing	243	+0.3 [-11.5, +11.9]	+37.7 [+17.5, +52.8]	within / beyond
Structured ledger	243	+0.0 [+0.0, +0.0]	-2.7 [-8.2, +0.0]	within / within
Paginated listing				
One-shot	243	+21.3 [+8.9, +38.8]	+24.9 [+6.2, +47.4]	beyond / beyond
BM25+LLM	243	+19.8 [+8.4, +36.9]	+21.1 [+7.6, +37.0]	within / within
Graph-RAG	243	+26.4 [+7.3, +48.7]	+25.2 [+11.4, +44.6]	beyond / beyond
ReAct-style browsing	243	+9.6 [-3.2, +19.9]	+19.0 [+5.2, +34.4]	within / beyond
Structured ledger	243	+0.0 [+0.0, +0.0]	+0.0 [+0.0, +0.0]	within / within
Deep navigation				
One-shot	243	+14.9 [+0.9, +29.6]	+14.0 [-3.5, +31.7]	within / beyond
BM25+LLM	243	-3.2 [-22.7, +11.1]	+19.2 [+3.0, +38.8]	within / beyond
Graph-RAG	243	+21.0 [+7.3, +36.7]	+27.4 [+8.1, +47.5]	beyond / beyond
ReAct-style browsing	243	-2.6 [-12.5, +5.3]	+35.6 [+15.1, +54.0]	within / beyond
Structured ledger	243	+0.0 [+0.0, +0.0]	+0.0 [+0.0, +0.0]	within / within

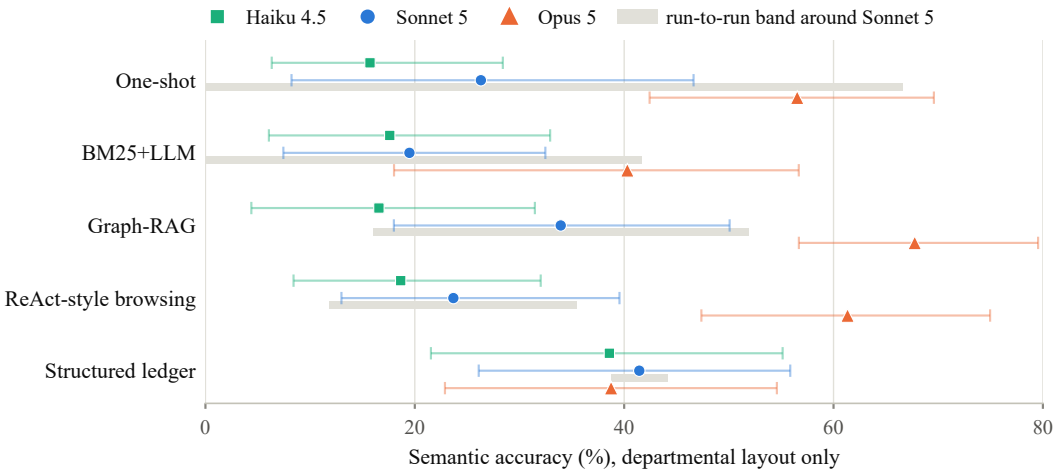

Figure 3: **Three base models on the one layout all three were run on.** Semantic accuracy on the departmental layout, 243 worlds shared by all three models, 95% bootstrap intervals clustered by institution. The grey band is the run-to-run floor measured by executing one plan twice on this layout: it is the mean absolute per-world change between the two executions, drawn around the Sonnet 5 value, and it is a dispersion over worlds rather than an interval on the mean. No ordering that places Haiku 4.5 against the other two survives it.

Table 11: The rows of Table 1 on the extended tier, all 243 scored worlds, 180 of which carry diagnostic references that locate obligations without closing the inventory (%). Recall is higher and omission lower than on the closure-complete tier, because an unfinished reference lists fewer of the obligations that are hard to find; the direction of every comparison in Section 6 is the same on both tiers. Brackets are 95% intervals from a bootstrap over institutions; † and ‡ as in Table 1.

Model	Scaffold	Entity F_1 ↑	Omission↓	Exec. Recall↑
Claude Sonnet 5	One-shot	60.0 [47.1, 71.0]	72.0 [59.3, 84.7]	41.1 [30.0, 51.1]
	BM25+LLM	53.5 [40.3, 66.0]	77.5 [65.9, 88.9]	35.7 [23.7, 45.6]
	Graph-RAG	57.2 [43.1, 70.4]	70.7 [53.1, 86.5]	39.4 [27.0, 53.7]
	ReAct browsing	73.8 [64.3, 82.9]	72.0 [62.6, 79.6]	58.9 [47.1, 72.6]
	Structured ledger‡	68.9 [61.4, 74.3]	67.8 [56.3, 80.2]	35.8 [22.9, 48.6]
Claude Opus 5	One-shot	76.0 [66.4, 85.0]	56.6 [43.9, 71.2]	61.1 [48.3, 73.2]
	BM25+LLM	61.7 [45.9, 73.7]	64.3 [47.3, 83.8]	54.8 [37.7, 68.3]
	Graph-RAG	76.7 [61.7, 88.6]	54.3 [38.8, 71.7]	70.5 [60.1, 79.8]
	ReAct browsing	76.3 [65.9, 84.8]	57.8 [48.8, 67.4]	72.8 [64.0, 82.7]
	Structured ledger‡	68.9 [61.4, 74.3]	68.5 [56.8, 81.2]	34.6 [21.2, 47.9]
GPT-5.6	One-shot	78.2 [65.4, 87.6]	54.2 [39.1, 72.0]	65.4 [49.6, 77.9]
	BM25+LLM	44.7 [32.6, 56.6]	81.3 [74.5, 89.2]	27.5 [17.9, 38.0]
	Graph-RAG	79.3 [66.7, 88.8]	52.0 [36.9, 69.3]	69.3 [54.4, 81.5]
	ReAct browsing	74.8 [63.1, 85.3]	65.0 [55.5, 74.0]	59.5 [47.5, 72.8]
	Structured ledger‡	68.9 [61.4, 74.3]	65.5 [55.1, 76.7]	39.1 [27.8, 51.1]
GPT-6	One-shot†	72.0 [58.3, 82.3]	56.4 [41.8, 73.7]	62.2 [46.2, 74.8]
	BM25+LLM	31.0 [16.8, 43.6]	88.4 [79.8, 96.5]	18.4 [5.2, 32.9]
	Graph-RAG	72.3 [57.7, 83.4]	55.5 [40.9, 73.3]	66.7 [51.2, 78.5]
	ReAct browsing	85.4 [78.7, 90.9]	50.2 [39.4, 62.4]	70.2 [60.0, 80.4]
	Structured ledger‡	68.9 [61.4, 74.3]	63.9 [53.4, 74.5]	40.3 [29.4, 52.3]
Qwen3.8-27B	One-shot	63.9 [46.7, 77.6]	65.5 [50.0, 80.9]	43.9 [30.6, 57.4]
	BM25+LLM	24.1 [14.5, 33.0]	99.0 [97.5, 100.0]	3.7 [0.0, 9.4]
	Graph-RAG	43.3 [32.7, 52.6]	85.0 [76.3, 92.3]	20.1 [12.1, 30.2]
	ReAct browsing	67.0 [59.8, 73.5]	71.6 [61.9, 80.5]	44.6 [32.4, 59.4]
	Structured ledger‡	68.9 [61.4, 74.3]	69.8 [59.9, 80.9]	33.4 [22.2, 45.4]
Gemma-3-27B-it	One-shot	47.5 [32.7, 63.2]	92.6 [85.8, 97.6]	13.8 [6.3, 24.2]
	BM25+LLM	34.2 [23.0, 44.3]	97.7 [93.4, 100.0]	3.7 [0.0, 10.8]
	Graph-RAG	43.4 [27.9, 56.6]	97.7 [94.5, 100.0]	3.9 [0.0, 9.6]
	ReAct browsing	32.2 [23.2, 42.5]	97.6 [96.2, 98.8]	4.2 [1.5, 7.5]
	Structured ledger‡	68.9 [61.4, 74.3]	85.0 [70.3, 97.1]	15.1 [3.5, 28.4]

Table 12: Deterministic baseline on the September 14 snapshot, evaluated against the source-derived reference. *Legacy rule loss* is the complement of requirement recovery under the legacy rule scorer: the value 1.000 is the same fact Section 6 states as requirement recovery of zero.

Layout	Worlds	Course recall (%)	Legacy rule loss	WorldPass
Department pages	708	95.41	1.000	0
Paginated listing	708	0.00	1.000	0
Deep navigation	708	95.41	1.000	0
One page/course	708	0.00	1.000	0

N Completion Claims

Table 16 gives certification coverage and conditional risk for every arm whose stop decisions were scored, under the definition of Equation 2 with strict WorldPass as the correctness criterion, and over main cells only. Section 6.3 rescores the one-shot certifications against obligation closure, which is weaker than the strict criterion, so its risks are at most the ones printed here.

O Contract Failures by Model and Condition

Table 17 classifies, post hoc, the answers that violated the output contract in the arms whose outputs were classified. The classification changes no score: a contract-violating answer is scored as the failure it is, keeps loss 1, omission 1, and a strict WorldPass of `FAIL`, and stays in its denominator. The distribution differs by model rather than only in magnitude: the smallest model fails most often by emitting text that does not parse as JSON at all, whereas the larger models fail predominantly by returning well-formed documents that violate the worksheet contract. The arms not listed have no classification pass and are therefore absent rather than empty.

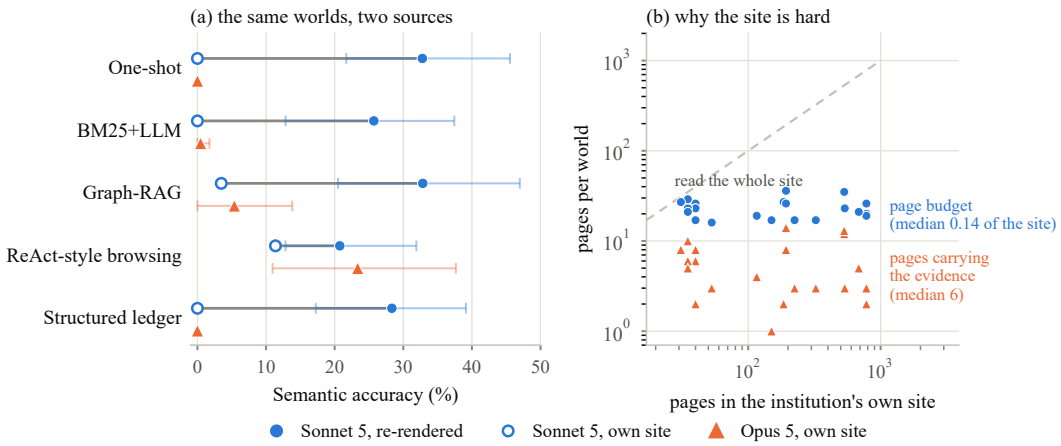

Figure 4: **The institution's own website is a different problem from a re-rendering of its content.** (a) Semantic accuracy on the same worlds, references, budgets and prompts, changing only the source; up to 225 worlds per row, 95% bootstrap intervals clustered by institution. Only the browsing row reads the site rather than being stopped by its own reading mechanism, and it is the one row whose interval does not separate. (b) Every world of that condition: the page budget and the smallest set of pages carrying the evidence, against the size of the site; the diagonal is a budget that would admit all of it.

Table 13: The frozen original institutional websites, reported as its own condition and never pooled with a re-rendered layout (%). Same worlds, references, prompts, budgets, seeds and snapshot on both sides; only the base model differs. Four of the five scaffolds are held below their own capability here by their reading mechanism rather than by the model: the one-shot bundle exceeds a fixed character cap and is refused, the two retrieval scaffolds are charged exactly one model call, and the structured scaffold's frontier order is independent of the program it was asked about. Only the browsing row is a reading of the site. One-shot reports nine worlds fewer than the other scaffolds: their cells died inside the HTML-to-text conversion because snapshot pages of those worlds carry binary PDF bytes while the manifest declares them as HTML. That is a harness defect, not a model failure; the cells are dropped rather than resampled, and the direction of the effect is unknown. Strict WorldPass is instrument-bounded: it is attainable on 18 of the 243 scored worlds (Appendix R.1).

		Entity F_1 ↑		SemAcc↑		WorldPass↑	
Scaffold	Worlds	Sonnet 5	Opus 5	Sonnet 5	Opus 5	Sonnet 5	Opus 5
One-shot	216	19.5	19.5	0.0	0.0	0.0	0.0
BM25+LLM	225	18.7	18.7	0.0	0.5	0.0	0.0
Graph-RAG	225	23.7	24.7	3.5	5.4	0.0	0.0
ReAct-style browsing	225	39.2	44.0	11.4	23.4	0.0	0.0
Structured ledger	225	31.2	31.2	0.0	0.0	0.0	0.0

P Component Ablations (Inconclusive)

We do not report the four component ablations as measured effects. All four flags are consumed by the frozen runtime and declared on every cell's receipt, and their arms were executed and scored; but the serving path offers no sampling control, each arm's request sequence agrees with the intact scaffold's about as often as the arms agree with one another, and three of the arms produce cell-for-cell identical semantic accuracy. Table 20 puts each arm against the repeated execution of the same plan by the rule stated in Appendix K: no arm's effect separates from run-to-run variation on any of the four metrics. Two readings survive this evidence and the data do not separate them: the flags may be inert on this serving path—three arms are cell-for-cell identical to each other, and all four score above the intact scaffold on semantic accuracy—or their effect may be smaller than run-to-run variation here. The arms are wired but their effect is unresolved at this signal-to-noise ratio; a receipt-level internal of the switch, or a repeated-seed serving path, is the observable that

Table 14: The frozen original institutional websites against the three re-rendered layouts of the same worlds. The page budget is matched to the largest budget the same world's re-renderings grant, so the difference is the site's structure and not the reading allowance.

median [min, max] over 234 worlds	original	deptpage	paginated	deepnav
pages in the site	189 [31, 781]	10 [7, 18]	10 [7, 19]	18 [11, 31]
page budget	23 [16, 36]	15 [12, 23]	16 [12, 24]	23 [16, 36]
share of site readable	0.14 [0.02, 0.87]	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]	1.00 [1.00, 1.00]
link depth of the deepest page	2 [1, 5]	2 [2, 2]	3 [2, 4]	4 [4, 4]
in-site links per page	52 [2, 192]	2 [1, 2]	1 [1, 1]	1 [1, 1]
pages carrying evidence	74 [3, 457]	6 [3, 14]	6 [3, 15]	8 [3, 16]
evidence pages / all pages	0.58 [0.06, 0.98]	0.60 [0.43, 0.78]	0.63 [0.43, 0.80]	0.43 [0.25, 0.57]
evidence sentences per such page	1.6 [0.1, 14.3]	12.7 [4.4, 40.3]	12.7 [4.4, 25.1]	12.2 [4.4, 30.2]
minimum link depth of evidence	1 [0, 2]	2 [1, 2]	2 [1, 2]	2 [1, 3]
depth of evidence (median)	1 [1, 3]	2 [2, 2]	2 [2, 2]	4 [3, 4]
worlds whose budget covers the whole site	0/234	234/234	234/234	234/234
pages left unread at the matched budget (median)	158	0	0	0

Table 15: Feasibility audit of the original-website condition, computed from the frozen snapshot and the references alone, before any model result is consulted. Every world's evidence is present and reachable; what the protocol's budgets do not grant is the reading.

Quantity	Value
Worlds in the condition	234
Worlds whose evidence is covered by some set of snapshot pages	234
Worlds whose evidence pages are all reachable from the root	234
Minimum evidence-bearing pages per world (median)	6
Pages a root-first traversal must open to reach them (median)	7
Page budget the protocol grants (median)	23
Characters that traversal must read (median)	782,773
Worlds whose evidence exceeds the one-shot character cap	135

would separate the two readings. Separating the switch from run noise requires a deterministic or repeated-seed serving path, or an internal observable of the switch recorded on the receipt, which we leave to future work. Table 18 is therefore reported as **inconclusive** and no claim in this paper depends on it.

An independent ceiling applies to these rows in particular. The structured scaffold is deterministic-first, and its course inventories are byte-identical across two different base models in 225/252 of the departmental worlds. A switch that changes only the model's behavior has a small observable effect on that scaffold by construction, so resolving these arms needs a higher signal-to-noise ratio rather than more cells.

Q Error Amplification

A program that abstains rarely and whose accept/reject mix happens to sit near the reference's own accept rate scores above the floor while still deciding most obligations incorrectly.

Table 22 and Figure 6 report how much behavioral error a unit of fact error buys, in bands of fact error, under two definitions of fact error that disagree with one another. $\mathrm{FE}(\mathrm{course})$ is one minus the world's course-entity F_1: what fraction of the course inventory is wrong. $\mathrm{FE}(\mathrm{req})$ is one minus the exact course-set agreement of the aligned requirements: what fraction of the requirements name the wrong set of courses. The first is the coarser and more forgiving measure, and amplification read against it is correspondingly larger. Both agree on the direction at low fact error: a reconstruction that gets most of its facts right still loses a larger share of its decisions. Neither supports a reading of the top band, where behavioral error is already saturated and a ratio below one is a ceiling artefact rather than damping.

Table 16: Completion claims: certification coverage / conditional risk of Equation 2, in percent, for each arm's own CERTIFY_COMPLETE decision. Risk is undefined (−−) when nothing was certified. World correctness is judged here by strict WORLDPASS; Section 6.3 reports the one-shot risk under obligation closure. Main cells only; the intervention arms of a run never enter a row.

Model and condition	One-shot	BM25	Graph-RAG	ReAct	Structured
Sonnet 5, re-rendered layouts	72 / 97	57 / 96	67 / 94	35 / 100	10 / 100
Opus 5, re-rendered layouts	84 / 96	40 / 91	80 / 95	65 / 94	10 / 100
Haiku 4.5, department pages	26 / 86	22 / 83	26 / 100	30 / 100	15 / 100
Haiku 4.5, paginated	22 / 100	11 / 100	22 / 100	26 / 100	0 / –
Sonnet 5, original sites	0 / –	4 / 100	8 / 100	0 / –	0 / –
Opus 5, original sites	0 / –	0 / –	0 / –	0 / –	0 / –

Table 17: Contract failures by model and condition, counted post hoc over the arms whose outputs were classified. Every cell counted here keeps its frozen-protocol score and stays in the denominator. *Contract* is an answer that parses but violates the worksheet contract, *schema* a JSON document that violates the output schema, *not JSON* an answer that does not parse, *truncated* an answer cut off at the output limit. A −− in the truncated column marks a row whose classifier has no such class; for the corrected rows the columns sum to the failed-cell total. †Corrected classification: cells whose raw status string names the worksheet/decision contract had been counted as schema violations by the frozen classifier; they are contract failures and are counted as such here. ‡Frozen classification, in which that mislabel is still present, so its *schema* column is an upper bound on genuine schema violations.

Model and condition	Failed cells	contract	schema	not JSON	other	truncated
Claude rows						
Opus 5, department pages	171	117	54	0	0	–
Opus 5, paginated	117	63	45	0	9	–
Opus 5, deep navigation	153	126	27	0	0	–
Haiku 4.5, department pages	486	225	108	144	9	–
Haiku 4.5, paginated	531	288	72	153	18	–
Sonnet 5, original sites	324	90	0	0	234	–
Opus 5, original sites	315	81	0	0	234	–
external cohorts, each its own transport (Table 26)						
Gemma-3-27B-it†	1,233	1,188	18	0	9	18
GPT-5.6†	576	531	0	0	45	0
GPT-6†	1,440	1,404	0	0	36	0
Qwen3.8-27B (thinking)†	1,395	1,062	0	0	324	9

Q.1 BEHAVIOURAL AGREEMENT ON THE FROZEN ORIGINAL WEBSITES

Table 23 reports program-level behavioural agreement for the 8 real-site v2 slices and for the two paired axes over them, on the same references, suites, seed and school-clustered bootstrap as the main grid. *No repetition slice exists on this condition*, so an interval that excludes zero is separable from sampling over worlds and schools and never evidence of a difference beyond run-to-run variation; none of the 3 model-axis pairs and none of the 3 source-access pairs excludes zero in any case (0 separate). BA is read against its own empty-program floor, printed beside every value, and *both* structured arms at root only sit below it (0.316 and 0.325 against 0.365): $BA = 0.316$ against a floor of 0.365, $\Delta BA = -4.9$ pp, with a share of 1.00 of its 216 cells no better than the empty program.

The structured/Sonnet 5 source-access pair separates on all three rule-level metrics (Appendix V) and does *not* separate on program-level behavioural agreement: $+0.193$ $[-0.077, +0.395]$ over 189 paired worlds. These are two different quantities, and the behavioural reading neither confirms nor refutes the rule-level result. The same pair under the stronger model does separate at the program level—$+0.207$ $[+0.022, +0.364]$ over 216 paired worlds with 0 dropped—and it is the only paired row of the whole behavioural-agreement square that excludes zero, so the source-access axis is separable at the program level on one model and not on the other, which is a difference between two

Table 18: Component ablations of the structured scaffold, Sonnet 5 (%). **Inconclusive**: see the text above for why these rows are not reported as measured component effects.

Condition	SemAcc↑	Citation↑	Cert. coverage	Cert. risk↓
Full structured scaffold	28.5 [18.8, 38.5]	99.5 [98.0, 100.0]	9.9 [0.0, 24.0]	100.0 [100.0, 100.0]
Without obligation ledger	38.8 [22.9, 54.6]	–	14.8 [0.0, 36.0]	100.0 [100.0, 100.0]
Without evidence obligations	38.8 [22.9, 54.6]	–	14.8 [0.0, 36.0]	100.0 [100.0, 100.0]
Without semantic feedback	41.0 [25.8, 55.6]	–	14.8 [0.0, 36.0]	100.0 [100.0, 100.0]
Without applicability handling	38.8 [22.9, 54.6]	–	14.8 [0.0, 36.0]	100.0 [100.0, 100.0]

Table 19: Run-to-run noise: the same plan, same worlds, same references, same scorer, executed a second time on each of the three layouts. For every scaffold we report the mean absolute per-world change in semantic accuracy and the test–retest correlation over the 234 worlds both executions scored. This is the floor against which every cross-model difference in Table 10 is judged. The structured scaffold is the only one whose course inventory is close to deterministic across executions (byte-identical course lists in 252/252, 252/252, 216/252 of the worlds on the three layouts), and even it is not byte-stable on deep navigation.

	Department pages		Paginated		Deep navigation	
Scaffold	mean $\lvert\Delta\rvert$	r	mean $\lvert\Delta\rvert$	r	mean $\lvert\Delta\rvert$	r
One-shot	0.403	-0.35	0.272	0.24	0.337	0.08
BM25+LLM	0.222	0.44	0.313	0.18	0.146	0.65
Graph-RAG	0.179	0.55	0.323	0.03	0.160	0.55
ReAct-style browsing	0.118	0.70	0.073	0.76	0.123	0.68
Structured ledger	0.027	0.93	0.000	1.00	0.000	1.00

arms and not an estimate of the model's effect. The same pair is conditioned on a world intersection: three Sonnet 5 structured oracle-evidence cells never produced an answer (wall-clock timeouts), so their worlds are absent from both sides and their absence flatters the arm.

Three further pairings exist on paper and are not computed. The Opus 5 structured root-only arm exists, so the fourth cross-model cell of the scaffold×source-access square and the Opus 5 structured source-access contrast are both reported above, the latter with the axis's only separable row; the model axis on that pinned arm is $+0.009$ $[+0.000, +0.026]$, not separable, as an arm fixed by its extraction mechanism should be. A v2 arm against a main-grid or v1 real-site arm moves two declared fields at once (a different environment and a different harness), and a cross-model pair that also moves source access moves two as well; neither is attributable to either field. A scaffold comparison inside one model and one source access is a per-method row of one arm, not an arm pair. Behavioural agreement also counts a predicted `UNKNOWN` as a disagreement, so a sparser program can score higher while being more wrong at the rule level; every difference here is therefore printed beside the rule-level metrics of the same arm.

R Evaluator and Reference Diagnostics in Full

A mutation audit on the reference checkpoint tests the evaluator on 22,914 mutants of 8,181 executable reference rules, over 432 of the 450 checkpoint worlds; the eighteen worlds without a built reference are untested. It catches 18,675 mutants and leaves 4,203 undecided because they fall outside the exact fragment, with no evaluator disagreements. The 36 mutants it decides but does not distinguish are independently verified to preserve meaning, so failing to distinguish them is the appropriate outcome. The critical mutation score—the caught fraction of the 18,711 decided mutants—is 0.998, while the decided fraction of all generated mutants, 0.817, bounds the scope of that discrimination. General-education, policy, and reference-extension semantics are inventoried but lie outside this mutation battery.

Reverse tests decide 10,791 of 14,265 meaning-preserving rewrites as equivalent, leave 3,474 unknown, and produce no false non-equivalence. A separate reference-construction diagnostic compares pre-filled verification and source-only annotation on ten matched worlds. Mean agreement is

Table 20: The four scaffold ablations against that noise floor (DEPTPAGE, structured ledger). An arm exceeds the floor iff the 95 % interval of the paired difference-of-differences (arm−baseline minus repetition−baseline) excludes zero.

Arm	Metric	Arm−base	Rep−base	(Arm)−(Rep) [95 % CI]
Without obligation ledger	Entity F1	0.000	0.000	0.000 [+0.000, +0.000]
Without obligation ledger	Requirement omission	0.013	0.013	0.000 [+0.000, +0.000]
Without obligation ledger	Semantic accuracy	-0.027	-0.027	0.000 [+0.000, +0.000]
Without obligation ledger	Strict WorldPass	0.000	0.000	0.000 [+0.000, +0.000]
Without evidence obligations	Entity F1	0.000	0.000	0.000 [+0.000, +0.000]
Without evidence obligations	Requirement omission	0.013	0.013	0.000 [+0.000, +0.000]
Without evidence obligations	Semantic accuracy	-0.027	-0.027	0.000 [+0.000, +0.000]
Without evidence obligations	Strict WorldPass	0.000	0.000	0.000 [+0.000, +0.000]
Without semantic feedback	Entity F1	0.000	0.000	0.000 [+0.000, +0.000]
Without semantic feedback	Requirement omission	-0.002	0.013	-0.015 [-0.048, +0.000]
Without semantic feedback	Semantic accuracy	-0.004	-0.027	0.022 [+0.000, +0.069]
Without semantic feedback	Strict WorldPass	0.000	0.000	0.000 [+0.000, +0.000]
Without applicability handling	Entity F1	0.000	0.000	0.000 [+0.000, +0.000]
Without applicability handling	Requirement omission	0.013	0.013	0.000 [+0.000, +0.000]
Without applicability handling	Semantic accuracy	-0.027	-0.027	0.000 [+0.000, +0.000]
Without applicability handling	Strict WorldPass	0.000	0.000	0.000 [+0.000, +0.000]

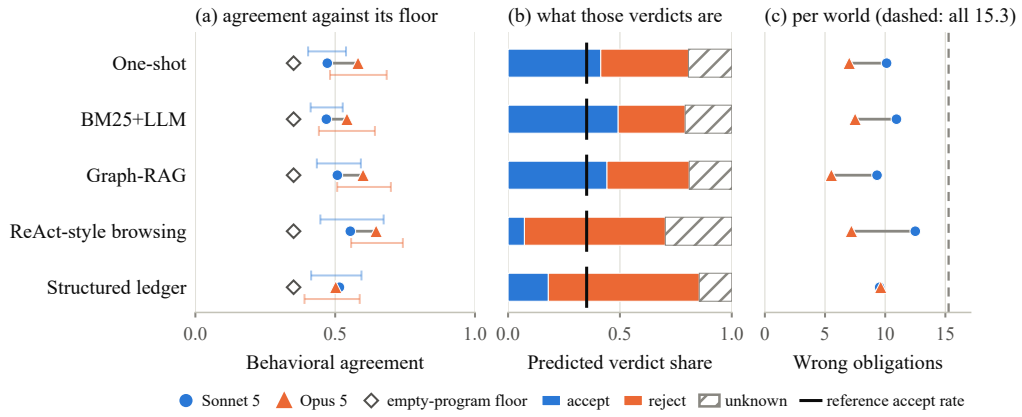

Figure 5: **What the reconstructed programs decide, and what that agreement is made of.** 3,510 world–layout cells over 45 institutions, three re-rendered layouts, 95% school-clustered bootstrap intervals. (a) Agreement against the same world's empty-program floor, which certifies nothing yet already agrees with every record the reference rejects. (b) Predicted accept, reject and unknown shares, Sonnet 5, against the reference's own accept rate. (c) Obligations decided incorrectly per world. Full numbers: Table 21.

0.9208 for verify–verify pairs and 0.8722 for verify–blank pairs, with a reported paired difference of 0.0487. This descriptive difference motivates independent reference review, but does not isolate the causal effect of pre-filling or certify the current reference collection.

R.1 Is a Strict WorldPass Attainable at All?

Strict WorldPass is the one metric in this paper whose small counts are a property of the reference instrument before they are a property of any agent, and Table 24 says so world by world. The rule was read from the scorer this paper pins (`pilot.py`, `04abd360`) and applied to the reference files alone, before any model output was consulted. Under it, a strict pass is attainable on 18 of the 243 scored worlds. Two things block the rest, and neither is an agent's doing. On 216 worlds the reference carries loader-validated world-credit contracts, and the strict path over them can only return `FAIL` or `UNKNOWN` (`pilot.py`:1273–1284); on 9 worlds no obligation is decidable at all, so the strict conjunction is never built (`pilot.py`:1168, 1191, 1239). The third path, `REFERENCE-EXTENSIONS` bundles, blocks 0 worlds here.

Table 21: **Upper panel:** program-level behavioural agreement on the three re-rendered layouts—the share of a world's frozen record suite on which the reconstructed program and the reference return the same verdict. BA_{empty} is the same world's empty-program floor, which certifies nothing and therefore already agrees with the reference on every record the reference rejects; BA_{norm} rescales the Sonnet 5 result between that floor and 1, and the two share columns are computed on the Sonnet 5 run. **Lower panel:** what that agreement is made of. A program that abstains rarely and whose accept/reject mix sits near the reference's own accept rate scores above the floor without recovering the rules, so the two right-hand columns give the mean number of reference obligations each reconstruction decides incorrectly, out of 15.3 decidable obligations per world.

Scaffold	BA Sonnet 5	BA Opus 5	BA_{empty}	$\leq$ empty	$BA{=}1$	BA_{norm}
One-shot	0.472	0.582	0.351	0.55	0.06	0.186
BM25+LLM	0.468	0.542	0.351	0.62	0.05	0.181
Graph-RAG	0.508	0.600	0.351	0.51	0.06	0.241
ReAct-style browsing	0.554	0.646	0.351	0.29	0.06	0.313
Structured ledger	0.515	0.502	0.351	0.37	0.00	0.253
All scaffolds pooled	0.503	0.574	0.351	0.47	0.05	0.235

	reference	predicted verdict share			wrong obligations / world	
Scaffold	accept rate	accept	reject	unknown	Sonnet 5	Opus 5
One-shot	0.35	0.41	0.39	0.19	10.1	7.0
BM25+LLM	0.35	0.49	0.30	0.21	10.9	7.5
Graph-RAG	0.35	0.44	0.37	0.19	9.3	5.5
ReAct-style browsing	0.35	0.07	0.63	0.30	12.5	7.2
Structured ledger	0.35	0.18	0.67	0.15	9.5	9.6
All scaffolds pooled	0.35	0.32	0.47	0.21	10.5	7.4

Table 22: Error amplification: behavioural error divided by fact error, in bands of fact error, Sonnet 5. Two definitions of fact error are reported because they disagree. FE(course) is one minus the world's course-entity F_1; FE(req) is one minus the exact course-set agreement of the aligned requirements. Amplification above 1 means a reconstruction loses more decisions than facts. In the top band the behavioural error saturates, so a ratio below 1 there is a ceiling artefact and not evidence of damping.

Fact-error band	Cells	median ratio	pooled ratio	share > 1
FE(course)				
(0.0, 0.2]	1,161	5.93	5.51	0.95
(0.2, 0.4]	414	1.97	2.03	0.76
(0.4, 0.7]	648	0.93	1.03	0.39
(0.7, 1.01]	855	0.73	0.72	0.08
FE(req)				
(0.0, 0.2]	270	2.68	2.42	0.90
(0.2, 0.4]	198	1.61	1.10	0.59
(0.4, 0.7]	657	0.86	0.85	0.32
(0.7, 1.01]	2,322	0.58	0.58	0.05

A residual `UNKNOWN` obligation does *not* block a strict pass. The scorer fills its strict conjunction only from reference-decidable obligations, so an undecidable one leaves the conjunction instead of failing it; what fails a world is an obligation the *prediction* leaves unresolved (`pilot.py`:1261). The consequence for this paper is specific. All 540 strict passes recorded across the 15 scored runs read here fall in 9 of the 18 attainable worlds, each a complete-reference world with about a dozen decidable obligations and no contracts. The other 9 attainable worlds are passed by no system in any of their 1,206 cells, and they are therefore the only worlds in this study whose strict zero is a statement about the systems. On the 180 diagnostic worlds, strict WORLDPASS is 0 by construction, so the zeros in Table 31's diagnostic column record the shape of those references and not the behaviour of the arms; 45 of the 63 complete-reference worlds are unattainable for the same contract reason.

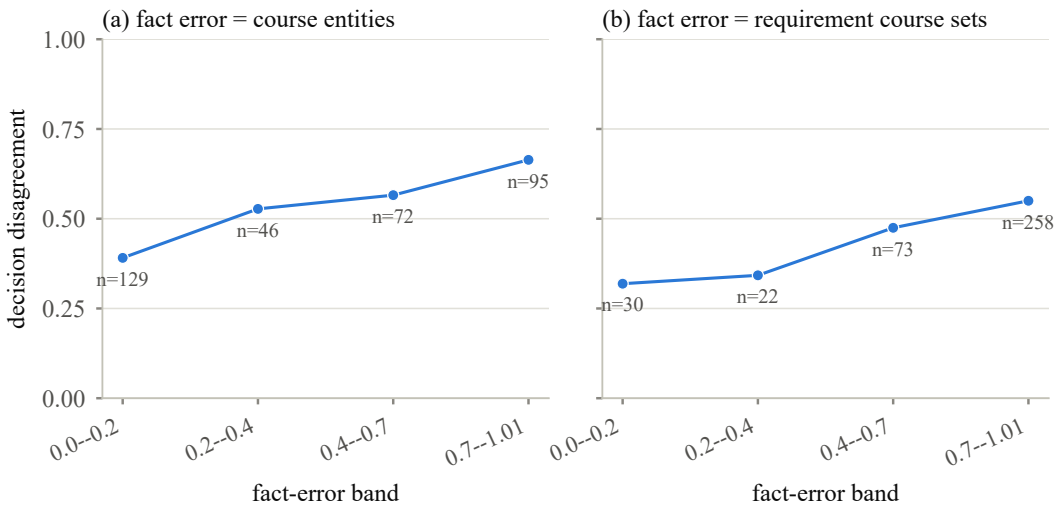

Figure 6: **Fact errors turn into decision errors at a rate above one.** The share of record decisions a reconstruction gets wrong, in bands of fact error, Sonnet 5 on the three re-rendered layouts; n is the number of cells in each band and both panels share the vertical scale. The two definitions of fact error are drawn separately because they disagree about which cells are accurate. In the top band the decision error saturates, so the ratio there is a ceiling artifact rather than damping.

Table 23: Program-level behavioural agreement on the frozen original websites, the square completed to 6 arms over 8 conditions, with both paired axes. Same references, record suites, seed and school-clustered bootstrap as the main grid. *This condition carries no repetition slice*, so an interval excluding zero is separable under resampling of worlds and schools and is never evidence of a difference beyond run-to-run variation. BA is read against the empty-program floor printed beside it. Exactly 1 of the 4+4 paired rows excludes zero.

Condition	Scaffold	Cells	BA	BA_{empty}	ΔBA (pp)	$\leq$ floor
S5 search root	react-search	216	0.529 [0.397, 0.659]	0.365	+16.4 [-1.7, +34.4]	0.46
S5 ledger root	structured	216	0.316 [0.259, 0.395]	0.365	-4.9 [-11.5, +0.0]	1.00
S5 search oracle	react-search	216	0.596 [0.505, 0.694]	0.365	+23.1 [+9.9, +37.2]	0.38
S5 ledger oracle	structured	189	0.502 [0.293, 0.673]	0.364	+13.8 [-11.3, +35.8]	0.33
O5 search root	react-search	216	0.570 [0.458, 0.669]	0.365	+20.5 [+3.4, +35.6]	0.38
O5 search oracle	react-search	216	0.608 [0.511, 0.702]	0.365	+24.4 [+7.7, +39.6]	0.33
O5 ledger oracle	structured	216	0.532 [0.391, 0.660]	0.365	+16.7 [-1.6, +33.7]	0.29
O5 ledger root	structured	216	0.325 [0.275, 0.400]	0.365	-4.0 [-9.1, +0.0]	1.00
Model axis: Opus 5 − Sonnet 5, same scaffold and source access						
search root	216 pairs, 0 dropped		0.570	0.529	+0.042 [-0.074, +0.161]	
search oracle	216 pairs, 0 dropped		0.608	0.596	+0.013 [-0.076, +0.083]	
ledger oracle	189 pairs, 27 dropped		0.538	0.502	+0.036 [+0.000, +0.085]	
ledger root	216 pairs, 0 dropped		0.325	0.316	+0.009 [+0.000, +0.026]	
Source-access axis: oracle-evidence − root-only, same model and scaffold						
S5 search, oracle – root	216 pairs, 0 dropped		0.596	0.529	+0.067 [-0.026, +0.172]	
S5 ledger, oracle – root	189 pairs, 27 dropped		0.502	0.309	+0.193 [-0.077, +0.395]	
O5 search, oracle – root	216 pairs, 0 dropped		0.608	0.570	+0.038 [-0.001, +0.120]	
O5 ledger, oracle minus root	216 pairs, 0 dropped		0.532	0.325	+0.207 [+0.022, +0.364]	

The counts of contracts read from the references agree with the `reference_contract_count` the scorer itself wrote into the cells in all 243 worlds. One limit travels with the table: *attainable* means not constructively excluded by the reference, and a prediction that emits its own credit contracts or reference extensions enters the same code path and blocks itself.

S External Models, Each Its Own Cohort

One provenance note belongs to this cohort table. The Gemma row's run directory carries no record of the row map its published aggregate was rendered with: the directory has no applied-row-map

Table 24: Is a strict WORLDPASS attainable at all? Read from the pinned scorer's own rules (`pilot.py`, `04abd360`) and the reference files, before any model output is consulted, over the 243 scored worlds. Residual `UNKNOWN` obligations do not block a strict pass, because the scorer builds its strict conjunction only from reference-decidable obligations (`pilot.py`:1168, 1191, 1239). What blocks is the reference's own world-credit contract, whose strict path can only return `FAIL` or `UNKNOWN` (`pilot.py`:1273–1284), and having no decidable obligation at all. Strict passes are cells, over all 15 scored runs read here. The contract counts agree with the `reference_contract_count` the scorer itself wrote in all 243 worlds.

	Complete reference	Diagnostic reference	All
Scored worlds	63	180	243
Blocked by a world-credit contract	36	180	216
Blocked by having no decidable obligation	9	0	9
Blocked by reference-extension bundles	0	0	0
Strict pass attainable	18	0	18
Cells of attainable worlds	2,412	0	2,412
Strict passes recorded	540	0	540
Attainable worlds passed by some system	9	0	9

document, and its aggregate names the Graph-RAG condition by a key that the cells themselves spell differently. The cells, the scorer and the reference are pinned and unchanged, and no value in this paper was altered; what is not reproducible from the directory alone is the mapping step that produced that one published row.

Three further models were executed on the same worlds, layouts, scaffolds, references, budgets and scorer as the runs of Section 6: one open-weight model served by the authors on an OpenAI-compatible server, and two hosted models reached through a vendor command-line client. Table 25 reports them, and Table 26 states why they appear as three separate cohorts rather than as additional rows of Table 1. The three transports differ in what a client may control and in what precedes the frozen prompt. On the self-hosted server a temperature of 0 and a top-p of 1 are applied on the wire, and the portable bundle imposes an output ceiling on two scaffolds that the other rows do not carry. On the vendor client neither sampling nor seed nor an output cap is honoured, and every request additionally carries that client's own harness system prompt—measured at roughly eleven thousand prompt tokens—which this benchmark did not write and cannot remove. The scored cells of those two rows are moreover distributed over several request-timeout conditions rather than one. For these reasons no number in this paper differences one cohort against another, averages them, or orders the models by them; the rows are reported so that the task's difficulty can be seen outside a single vendor, not to rank vendors. None of these three models exposes a weight hash and a build identifier simultaneously: the open-weight row is pinned by the hashes of its weight index, and the two hosted rows have no revision string at all, which is recorded rather than filled in.

One row needs an explicit reading. The GPT-6 one-shot cells score zero semantic accuracy, and this is a formatting-compliance failure rather than an inability to read the sources. All 756 of those cells returned a complete, well-formed JSON document in a single attempt, with 0 non-200 responses, 0 transport errors and 0 truncated replies; the JSON schema raised no issue on any of them. Every one of them failed at the same later stage and in the same place: the field lines of the course worksheet were written flush left instead of indented by two spaces, so the worksheet parser rejected the document before any content was compared. The same worlds under the other vendor model, which is otherwise structured identically, wrote the two spaces. The cells keep their frozen-protocol scores, as every contract failure does; what they measure is instruction compliance with the output contract, and they are not evidence about reconstruction ability.

S.1 A SECOND CONTROLLED MODEL AXIS, ON ANOTHER TRANSPORT

The cohorts above are never differenced. One pair of them, however, differs in exactly one declared field and can therefore be read as an axis in its own right: GPT-5.6 and GPT-6 ran over the same codex exec transport, the same layouts, scaffolds, worlds, references, prompts, budgets and pinned scorer (`04abd360`), with the model as the only difference. Table 27 pairs their 3,645 common cells by (world, layout, scaffold, intervention). Pooled over scaffolds, the newer model is

Table 25: Four external models, each reported as its own cohort (%). These rows are *not* comparable with each other or with the Claude tables as a capability ranking: they were reached over three different transports with four different degrees of decoding control, and none of them carries a repetition, so no difference here may be called beyond run-to-run noise, and one of them carries a vendor harness prompt in front of every frozen prompt (Table 26). Scaffolds, worlds, layouts, references, budgets and scorer are identical to the Claude runs. Brackets are 95% school-clustered bootstrap intervals; $--$ marks a metric with no denominator in that run. The GPT-6 one-shot row is a worksheet-format compliance failure, not a reading failure; see Appendix S. Strict WorldPass is instrument-bounded: it is attainable on 18 of the 243 scored worlds (Appendix R.1).

Model / scaffold	Entity F_1 ↑	Omission↓	SemAcc↑	Citation↑	WorldPass↑	Cells
Gemma-3-27B-it (self-hosted server, T=0 on the wire)						
One-shot	47.5 [32.7, 63.2]	95.5 [91.2, 98.7]	8.9 [2.4, 17.7]	–	1.2 [0.0, 4.5]	729
BM25+LLM	34.2 [23.0, 44.3]	98.5 [95.2, 100.0]	3.3 [0.0, 9.9]	–	0.0 [0.0, 0.0]	729
Graph-RAG	43.4 [27.9, 56.6]	98.8 [96.2, 100.0]	2.9 [0.0, 8.7]	–	0.0 [0.0, 0.0]	729
ReAct-style browsing	32.2 [23.2, 42.5]	98.2 [96.9, 99.3]	3.8 [1.3, 7.1]	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]	729
Structured ledger	68.9 [61.4, 74.3]	89.3 [77.4, 99.0]	14.8 [3.3, 28.1]	99.9 [99.4, 100.0]	0.0 [0.0, 0.0]	729
GPT-5.6 (subscription CLI)						
One-shot	78.2 [65.4, 87.6]	68.6 [57.9, 79.7]	49.8 [36.7, 64.2]	–	3.7 [0.0, 13.6]	729
BM25+LLM	44.7 [32.6, 56.6]	86.4 [81.1, 92.7]	24.1 [14.8, 34.5]	–	1.2 [0.0, 4.5]	729
Graph-RAG	79.3 [66.7, 88.8]	61.2 [47.4, 76.4]	60.9 [45.6, 74.8]	–	3.7 [0.0, 13.6]	729
ReAct-style browsing	74.8 [63.1, 85.3]	75.6 [66.6, 83.0]	42.5 [30.1, 59.0]	0.0 [0.0, 0.0]	3.7 [0.0, 13.6]	729
Structured ledger	68.9 [61.4, 74.3]	77.3 [68.0, 86.3]	31.9 [22.9, 41.0]	99.6 [98.3, 100.0]	0.0 [0.0, 0.0]	729
GPT-6 (subscription CLI)						
One-shot	20.1 [9.8, 30.2]	100.0 [100.0, 100.0]	0.0 [0.0, 0.0]	–	0.0 [0.0, 0.0]	729
BM25+LLM	31.0 [16.8, 43.6]	91.2 [85.0, 97.0]	15.7 [5.1, 29.2]	–	0.0 [0.0, 0.0]	729
Graph-RAG	72.3 [57.7, 83.4]	65.3 [52.4, 78.4]	54.8 [40.2, 67.8]	–	3.7 [0.0, 13.6]	729
ReAct-style browsing	85.4 [78.7, 90.9]	63.3 [55.0, 72.3]	53.7 [43.8, 64.2]	0.0 [0.0, 0.0]	2.5 [0.0, 9.1]	729
Structured ledger	68.9 [61.4, 74.3]	75.5 [66.0, 84.7]	33.3 [24.5, 42.4]	99.6 [98.4, 100.0]	0.0 [0.0, 0.0]	729
Qwen3.8-27B, thinking enabled (self-hosted server, sampling pending verification)						
One-shot	63.9 [46.7, 77.6]	73.6 [61.0, 85.5]	38.5 [25.9, 51.8]	–	3.7 [0.0, 13.6]	729
BM25+LLM	24.1 [14.5, 33.0]	99.1 [97.6, 100.0]	3.7 [0.0, 9.4]	–	0.0 [0.0, 0.0]	729
Graph-RAG	43.3 [32.7, 52.6]	88.6 [80.1, 95.1]	17.4 [9.8, 27.4]	–	3.7 [0.0, 13.6]	729
ReAct-style browsing	67.0 [59.8, 73.5]	79.9 [71.1, 87.6]	31.2 [19.1, 48.2]	0.0 [0.0, 0.0]	1.2 [0.0, 4.5]	729
Structured ledger	68.9 [61.4, 74.3]	81.1 [73.3, 89.0]	26.5 [18.4, 34.1]	99.5 [98.0, 100.0]	0.0 [0.0, 0.0]	729

separably worse on this transport: course recovery -0.136 $[-0.197, -0.085]$, requirement omission $+0.052$ $[+0.020, +0.079]$ and semantic accuracy -0.104 $[-0.167, -0.050]$. The browsing scaffold moves the other way, and it is the only row where all three move together and separably: $+0.106$ $[+0.008, +0.212]$, -0.124 $[-0.185, -0.051]$ and $+0.112$ $[+0.018, +0.188]$ over 729 paired cells.

The one-shot row collapses—-0.581 of course recovery, $+0.314$ of omission, -0.498 of semantic accuracy—and that row is a contract effect rather than a capability reading: all 756 of the newer model's scored one-shot cells fail the strict v1.0 worksheet contract by writing the DSL field lines flush left instead of with the required two-space indent, one call each, HTTP 200, complete JSON, no schema problem (Appendix Z). Read it as DSL surface compliance under strict v1.0. The structured scaffold's course recovery is identical cell by cell across the two models on all 729 pairs ($+0.000$); the artefact reports it and no mechanism is claimed here. Strict WORLDPASS separates on 0 of the six rows, which is what an instrument attainable on 18 of 243 worlds can show (Appendix R.1).

This paper now carries two controlled model axes, and they may be stated side by side but never subtracted from or ranked against each other: they differ in transport, model family and condition. In the Claude family, with source access held at root only on the frozen real site, a stronger base model is ahead on every mean and separably on none—a level comparison between two arms (Appendix V). In the GPT family, on the codex transport over the re-rendered layouts, a stronger base model separates on the browsing scaffold in all three metrics and loses the one-shot row to the worksheet contract. Each statement holds inside its own condition. Five further limits apply to this axis: no repetition slice and therefore no run-to-run floor, no enforced temperature or seed, harness preambles of 10,628–10,634 against 11,842–11,850 tokens measured on near-empty probes, cells spread over 4 transport conditions rather than randomised across them, and one paired cell that rests on a batch-order choice in the union index.

T What Happens to a Reference Obligation

Table 28 decomposes every cell × obligation observation of every reported run into the outcome the evaluator assigned it. The counts are micro-counts over obligations and are not the school-macro-averaged semantic accuracy of the main tables; they answer a different question, namely what the residual error is made of rather than how large it is per world. Two readings are supported and a third is not. First, of the 252,108 decidable obligations, 190,296 are absent from the reconstruction altogether, which is 99.3% of everything not shown equivalent: at this capability level the dominant failure is that an obligation never appears, not that it appears with the wrong meaning. Second, where a rule does appear and is proven wrong, the disagreement is concentrated in the course set (603 of 837 counterexamples) and in grade conditions (189), not in exotic logical structure. What the table does *not* support is a causal reading of the omissions: an omission is equally produced by an empty answer, a contract failure or a genuine miss, and these counts pool conditions, models and repeated rules. The grouping of counterexamples is produced automatically from the rule difference and is not a human root-cause annotation.

U The Closest Work, Row by Row

Table 29 places CurriculumWorldBench against the 8 lines of work of the original comparison and the 12 further works verified for this version, along the four axes this paper measures. The comparison is deliberately narrow. It reports, for each work, the task and metric definitions we read in a specific version of that paper; "not reported" therefore means that those definitions do not define such a measurement, not that the area lacks the capability or that a later version does not add it. Self-reported scales are given in each work's own units and do not rank the benchmarks against one another; no experiment of any other work was reproduced here.

Three claims that a reader might expect are deliberately absent. We do not claim to be first to produce executable rules from institutional text, first to measure rule omission or symbolic gold, or first to apply risk–coverage analysis to a system that may decline; each of those has established prior work, cited in the related-work section. We also do not claim that the discovery half and the interpretation half of this task have been separated on a single scaffold: the arm that would have established it was disqualified by an audit of its own receipts. What exists instead is 4 single-field handover contrasts (Table 30); they disagree, 2 of them improves anything, and the difference between them is read rather than estimated, because they differ in agent capacity as well as in the field that was moved. Specific here is the selective event—an agent declaring a bounded rule scope completely reconstructed—and equivalence exact only inside a supported fragment and record domain. Neither an executable degree rule nor a source-grounded curriculum representation is new here. The increment we do claim is in the protocol: the unit of evaluation is the obligation, the denominator excludes obligations the reference itself leaves undecided, equivalence is decided by counterexample inside a declared fragment, and a completion claim is scored with a conditional risk rather than assumed.

The 3 verified families. Browsing and deep-research benchmarks score answers, sometimes with citations, against reference strings; none of the four verified here asks for executable rules or coverage certificates (Wei et al., 2025; Mialon et al., 2023; Chen et al., 2025; Gao et al., 2023). Web knowledge-base construction turns pages into triples or graphs, scored by extraction quality where the fetched page reports a metric at all; the four verified here deliver data, not a program to run (Shin et al., 2015; Whitehouse et al., 2023; Sun et al., 2024; Zhou et al., 2021). Advising and degree-planning systems draw on curated institutional resources; in the four verified works rules are authored upstream, not reconstructed from official pages (Mi et al., 2025; Jiang et al., 2025; Lugones et al., 2026; Li & Iwata, 2026).

Aurora is the boundary case and the one place where this paper's novelty must be stated precisely (Lugones et al., 2026). It is the only work among the 12 whose output is an executable rule artifact, so the difference is not symbolic execution, which it already does: its input is a normalised BCNF curricular database that somebody else authored, whereas an episode here starts at a root URL over frozen official pages, and its reported behaviour is abstention on unanswerable prompts rather than a coverage certificate over a claimed scope. Nothing beyond the fetched abstract is asserted about it.

Table 29: The 20 closest lines of work, by what each one evaluates; the last 12 were verified for this table against their own landing pages, and every cell in them is a reading of the page that was fetched, not a measurement made here. Self-reported scales use incompatible units and do not rank these benchmarks against one another; "not reported" means the task and metric definitions we read do not define such a measurement, not that the area lacks the capability. Sources, page locations and the version of each paper that was read are recorded with the citation audit for this table.

Work	Input	Output	How behaviour is checked	Completion claim
KNOWPLAN	Cao et al. (2026) frozen official sites *scale (self-reported)*: two suites (self-reported); not a CWB-measured scale	course JSON, constraint graph, plan	AST and plan-gold checks	closure certificate; no risk–coverage report
BREX	Yang et al. (2026) given business documents *scale (self-reported)*: 409 documents, 2855 rules	condition–action and dependency structure	field and dependency F_1	not reported
Tax law	Lorenzo et al. (2025) given statute fragments *scale (self-reported)*: 591 examples (416/86/89)	Catala code, not always syntactically valid	similarity and AST metrics	not reported
RuleArena	Zhou et al. (2025) policy text supplied directly *scale (self-reported)*: 95 rules, 816 problems, 3 domains	answers and traces	answer and rule-application diagnostics	not reported
OpenExempt	Servantez et al. (2026) given cases and statutes *scale (self-reported)*: 9765 examples, 9 suites	structured answers	set F_1, numeric tolerance	not reported
WebArena	Zhou et al. (2024) self-hosted copies of real software *scale (self-reported)*: 812 tasks	web actions and answers	functional task success	task success only
Mind2Web 2	Gou et al. (2025) the live web *scale (self-reported)*: 130 tasks	cited answers; the executable part is the judge	rubric correctness and attribution	answer completeness, not reconstruction
AppWorld	Trivedi et al. (2024) controlled simulated apps and APIs *scale (self-reported)*: 750 tasks, 9 apps, ≈457 APIs	interactive code	database-state tests and collateral damage	task/scenario completion
BrowseComp	Wei et al. (2025) the live web *scale (self-reported)*: 1,266 questions	short answers	match against reference answers	not reported
ALCE	Gao et al. (2023) fixed retrieval corpora and questions *scale (self-reported)*: not stated on the fetched page	text with inline citations	fluency, correctness and citation quality	citation support measured per answer; no system verdict
GAIA	Mialon et al. (2023) real-world questions needing browsing and tools *scale (self-reported)*: 466 questions, 300 answers withheld	short answers	answer match, held-back leaderboard split	not reported
BrowseComp-Plus	Chen et al. (2025) a fixed curated corpus with human-verified supporting documents *scale (self-reported)*: derived from BrowseComp	answers and search traces	accuracy plus retrieval and citation analysis	not reported
DeepDive	Shin et al. (2015) given unstructured documents *scale (self-reported)*: five KBC systems	a populated database	extraction quality and inference cost	not reported
WebIE	Whitehouse et al. (2023) Common Crawl sentences, including ones with no facts *scale (self-reported)*: 1.6M sentences, ≈21K crowd-annotated triples, 5 languages	entity-linked fact triples	in-domain, out-of-domain, and zero-shot cross-lingual performance of generative IE models	not reported; negatives make "no fact" scorable
Docs2KG	Sun et al. (2024) heterogeneous documents, web pages included	one queryable knowledge graph	no held-out metric on the fetched page	not reported

Table 29, continued from the previous page.

Work	Input	Output	How behaviour is checked	Completion claim
	scale (self-reported): not stated on the fetched page			
SimpDOM	Zhou et al. (2021) rendered DOM trees of web pages *scale (self-reported)*: SWDE, +1.44 F1 over the prior best	typed attribute values bound to nodes	node-tagging F1, few-shot and cross-vertical	not reported
SmartCourse	Mi et al. (2025) a given transcript and four-year plan *scale (self-reported)*: 25 advising queries	prose course recommendations	custom PlanScore, PersonalScore, Lift, Recall	not reported
AdvisingWise	Jiang et al. (2025) authoritative institutional resources and student background *scale (self-reported)*: 20 queries, 8 advisors	drafted advising responses	expert review, LLM-as-judge, advisor user study	not reported; a human advisor validates every response
Aurora	Lugones et al. (2026) a normalised BCNF curricular database *scale (self-reported)*: 0.68 to 0.93 alignment, 20 queries timed	Prolog rules plus LLM explanations	structured suite over in-scope and out-of-scope cases	abstention on unanswerable prompts, not completeness
RAG advising	Li & Iwata (2026) structured syllabus data *scale (self-reported)*: not stated on the fetched page	course selection, prerequisite understanding, and personalized study planning	the fetched abstract reports no evaluation	not reported
CURRICULUM-WORLD-BENCH	frozen official subtrees and controlled re-renderings	executable rule fragments, `UNKNOWN` retained	per-obligation semantics plus program behaviour on frozen records	`CERTIFY_COMPLETE` with conditional risk and coverage

V DISCOVERY AGAINST INTERPRETATION

Table 30 reports the executions of the real-site condition that can be placed beside one another, and the two pairs among them that isolate source discovery. Each pair holds one scaffold fixed and moves a single declared field: whether the agent starts at the institution's root page or is handed the pages the reference itself cites as evidence for that world. Everything else—agent, prompts, page and call budgets, base model, snapshot, references, scorer and seed—is identical inside a pair, and the difference is paired by world over the worlds both of its arms scored. The two pairs answer differently. For the in-page-search browsing agent, handing the evidence over does not improve course recovery and makes requirement omission measurably worse. For the structured ledger, which reads a fixed page set in a single pass and cannot go looking, the same change lifts course recovery, semantic accuracy and omission together, all three separable from zero over 198 worlds in 42 school clusters. The reading is therefore not that discovery does or does not matter on this benchmark, but that the value of being handed the evidence depends on what the agent can do with an open site. The same one-field change on the stronger base model moves nothing separably at all—course recovery -0.018 [-0.046, +0.011], omission +0.018 [+0.002, +0.044], semantic accuracy +0.017 [-0.040, +0.093] over 225 worlds in 45 clusters—and that pair drops 0 worlds, so it is not conditioned on what its oracle arm could finish. Of the 4 single-field contrasts, 4 contain a separable movement and 2 contain a separable improvement: the structured pair under each of the two base models. The fourth contrast, added with the stronger model's structured ledger, moves all three metrics separably in its favour—course recovery $+0.161$ $[+0.063, +0.282]$, requirement omission -0.105 $[-0.234, -0.029]$ and semantic accuracy $+0.168$ $[+0.044, +0.381]$ over 225 paired worlds in 45 clusters, with 0 worlds dropped, so unlike its Sonnet 5 twin it is not conditioned on what its oracle arm could finish. The remaining separable movement is the weaker model's search agent getting worse at omission when the evidence is handed to it. The statement that survives all four is therefore the conditional one: what handing over the evidence buys depends on the scaffold, and it repeats across two models in both directions—separably for the ledger that cannot go looking, separably for neither search arm.

One of the four arms must not be read as a model comparison. With the structured ledger at root only, the model axis is a difference of exactly zero on all 4 reported metrics, with intervals of zero width over 225 paired worlds, and we checked the cells: all 225 worlds agree on every reported metric, on the course list and on the model-call count. That is not two models performing equally well; it is one arm pinned by its mechanism, because the structured ledger's course inventory at root only comes from deterministic extraction over a fixed page budget rather than from the model reading the site (the same behaviour the re-rendered layouts record). The oracle side is not pinned: there the two models differ on every metric. Handing over the evidence pages is where the model's reading begins to matter—a reading of these two columns, not a mechanism this paper establishes.

Six qualifications belong to that reading and are not negotiable. First, the oracle arms depend on the reference by construction—their evidence pages are the ones the reference declares—so they are an upper bound on what discovery could contribute, not a system anyone could deploy; that is what an oracle arm is for. Second, the step from the first row to the second changes the agent's reading mechanism *and* the page budget at the same time, so it compares a capacity-adequate condition against the protocol transplanted unchanged and is never an ablation of the search action. Third, the two pairs differ in agent capacity as well as in the field that was moved, so the difference between them supports the statement that the value of handing over evidence depends on the agent, and does not estimate an effect of either. Fourth, the structured pair is conditioned on the worlds its oracle arm could finish: the 225 root-only cells shrink to 198 on the oracle side, because 27 worlds exhausted the declared per-cell time limit on all three attempts, and those are the most expensive worlds in the set. Fifth, this condition carries no repetition, so it has no retest floor of its own, and the result is one base model and one condition: it does not generalise to the re-rendered layouts, where discovery is trivial by construction, nor to a stronger base model. Sixth, and most important for how the result is worded, the field that moves is whether the evidence was handed over, not whether the agent could have found it: the structured root-only arm ends at an omission of 1.000 and a semantic accuracy of 0.000, which is equally consistent with never locating the governing pages and with reaching them and being unable to use them, and no arm here separates the two. The gap is attributed to the handover, not to a measured difference in search ability. Strict world success is zero on every arm of this condition, so that column separates nothing. The last block of Table 30 moves a different field: the base model, with source access held at root only, over 225 worlds in 45 school clusters. The stronger model is ahead on every mean—course recovery 0.518 to 0.619, omission 0.939 to 0.902, semantic accuracy 0.120 to 0.135—and not one of the three intervals excludes zero: +0.101 [+0.042, +0.156], -0.037 [-0.060, -0.013] and +0.015 [-0.062, +0.058]. The lower bound on course recovery is a hair below zero and is not to be read as separation. Neither model passes a single world strictly, on worlds where a strict pass is attainable at all only under the conditions of Appendix R.1. Both arms hold source access fixed, so the only declared difference is the model: this is a level comparison between two arms, not an ablation, and this condition carries no repetition, so there is no retest floor to check it against. The second execution reaches 0.518 course recovery, about half the distance to the 0.744 the same worlds reach on the re-rendered layouts, with semantic accuracy unchanged (0.114 to 0.120) and strict world success 0.000 in all three settings. Its telemetry locates the constraint from the other side: every one of its 234 cells opened at least one of the pages its own manifest declares as evidence, 0 cells spent the page budget and 99 stopped at the step limit, so the agent was reaching the evidence before it was handed to it.

The fourth cohort: an open-weight model with its reasoning trace enabled. Table 25 carries a fourth external row, scored through the same pinned evaluator as every other row, over 3,645 cells—the same one hundred and thirty-five cells are outside the denominator here as everywhere else. It is its own cohort and is pooled with nothing: its serving conditions differ from the other self-hosted row in three ways at once (Table 26), it carries no repetition, so no difference involving it may be called beyond run-to-run noise, and its sampling is reported as pending verification rather than as a number. Of its 1,395 contract failures, 1,062 violate the worksheet contract, 324 fall in the residual class and 9 is a truncation; the corrected classification leaves 0 schema violations where the frozen classifier counted 189, with 0 cells not checkable. The dominant failure of this row is therefore the output contract rather than what it read from the pages, and no number of its one-shot, retrieval or graph rows should be read as an extraction result without that sentence attached. One per-layout pattern is worth naming because it is not a noise story: on the structured scaffold this row reaches 93.3% and 93.3% course recovery on the department-page and deep-navigation layouts and 20.1% on the paginated one, while its requirement omission on that layout is 100.0%—every requirement

of every one of that layout's 243 cells is omitted—against 73.2% and 70.2% on the other two, and none of the structured cells is a contract failure. A scaffold can therefore collect courses and still emit nothing a rule checker can use, which is the separation this benchmark is built to see. The per-layout values are the frozen renderer's own, and 20 of its pooled values were checked against the run's own scored table before being quoted.

An always-abstaining system therefore establishes a zero-coverage control, rather than a zero-risk solution. Episodes whose world correctness cannot be adjudicated remain unadjudicated in the completion analysis.

W THE COMPLETE-REFERENCE SUBSET

Of the 252 test worlds whose reference was built, 243 are scored, and of those 63 carry a complete reference while 180 carry a diagnostic one. The complete-reference worlds are the confirmatory tier of Section 6. Table 31 re-aggregates the already-scored cells of every arm over the three world sets. No cell was re-scored and no estimate was recomputed: the aggregation uses this paper's own renderer, and on the full set it reproduces the published tables value for value.

Three readings travel with that table. First, the complete/diagnostic split is how far reference adjudication got, not a randomisation, so a difference between two columns describes two world sets and estimates no effect of reference completeness. Second, the complete-reference columns rest on 18 school clusters against 48 for the full set, and an interval on fewer than 8 clusters is indicative rather than inferential here as everywhere else in this paper. Third, evaluable coverage is 71.54% on the complete-reference worlds against 69.41% on the diagnostic ones and 69.89% over all of them (71.54%, 68.79% and 69.44% on the real-site in-page-search arm): the obligations the reference cannot decide are therefore not an artefact of diagnostic references, and the paper does not claim a difficulty bound for them. A fourth reading was added once the strict metric's attainability was derived (Appendix R.1): the table's strict WORLDPASS column is 0 on every diagnostic world because a strict pass is not attainable there under the scorer's rules, so that column describes the references and not the arms, and 45 of the complete-reference worlds are blocked in the same way.

Over the 62 non-pooled arm rows, moving from all 243 worlds to the 63 complete-reference ones leaves entity recovery better on 45 rows and worse on 17; requirement omission better on 22, worse on 37 and unchanged on 3; semantic accuracy better on 16, worse on 44 and unchanged on 2; and strict WORLDPASS better on 27 rows, worse on none and unchanged on 35. These are counts of directions between two world sets, and no mechanism is claimed for any of them. The two columns being compared describe the 243 scored worlds and the 63 worlds whose references are complete, which are not two assignments of one population: the split records adjudication progress, so a direction here estimates no effect of reference completeness on anything, and in particular a majority moving one way is not evidence that cleaner references are harder.

X A PROVEN COUNTEREXAMPLE, EXPORTED

Figure 1 draws a schematic rule pair. The scored cells hold real ones: in the main comparison alone, 198 obligations are present in a reconstruction and proven to disagree with their reference on a student record. The smallest of them, chosen by record size and then by name so that it is the same witness on every regeneration, is the requirement `FIELD-OF-STUDY-CODE-HVACR-CER-HVAC` of `cod-service-technician-certificate` under ReAct-style browsing on the paginated layout: on the record {`HVACR 1100`} the reference evaluates to True and the reconstruction to False. A single course separates a rule a student would be advised by from the rule the institution published, which is the unit this benchmark scores.

Y EVALUABLE COVERAGE OF THE REFERENCES

Section 4.4 defines semantic accuracy over the obligations the reference can decide, and promises the share those are of all enumerated obligations. Table 32 reports it, per condition, from the scored cells themselves: between 69.4% and 69.9% of the obligations the scorer enumerates are decidable. The complement is not a model failure and is charged to no prediction; it is the reference-coverage

limitation the protocol declares, and it is why a diagnostic reference cannot be read as a weaker agent. The quantity is a property of the references and the supported fragment, not of an agent, so every model run of the same condition carries the identical value—the generator refuses to print the table if two runs of a condition disagree.

Z How Much of the Contract Failure Is Whitespace?

The worksheet contract requires a field line to begin with exactly two spaces. Because one reported row failed almost all of its one-shot cells on that rule alone, we measured what a relaxed contract would change. Contract v1.1 accepts any leading whitespace on a field line and changes nothing else; it was installed identically for every reported row, no stored answer was modified, and no model was called. The previously rejected replies are simply re-parsed.

2,097 failed one-shot cells across 14 reported rows still hold the reply that was rejected; 639 of them parse under v1.1, and every one belongs to the same model row. 1 of the 14 rows changes at all: on every other row the two contracts are identical to the value, verified by re-scoring whole runs and comparing every scored field (1,162,944 values compared, zero changed). Table 33 shows the one row that moves: its one-shot semantic accuracy goes from 0.00% to 51.36% and its course recovery from 20.15% to 72.05%.

Only the main-text Table 1 and Figure 2 report the relaxed contract. There the row above is the only one that changes, and the multi-call rows equal their strict values because none of their rejected replies is stored. Every other table keeps the strict contract, for two reasons. Only a subset of the reported rows was re-scored under v1.1, and a metric must not be a mixture of two contracts; and the multi-call scaffolds—structured, browsing and the two retrieval scaffolds—never persist a rejected reply, so whether v1.1 would have rescued any of their failures is not measured and is not guessed here. What the measurement does establish is the size of the formatting component: outside one model's one-shot cells, none of the contract failures in this paper is an indentation artefact.

Table 26: Serving conditions, per reported row. No row of this benchmark controls decoding: two of the three transports discard a requested temperature and seed outright, and the third applies them at the server. One transport prepends a vendor harness system prompt this benchmark did not write, so on those rows the effective prompt is not the frozen prompt alone. A number in this table is read from the run's own cells or from the recorded endpoint identity; nothing is inferred. Transport-condition counts are over each run's scored cells. One row's sampling is marked pending verification: the configuration file its own cells name is on the compute node and is not in evidence here, so the value the bundle records is reported as such and not as the wire value. This is why the external rows of Table 25 are four separate cohorts and not a model ranking.

Claude Sonnet 5 / Opus 5 / Haiku 4.5	
transport	subscription CLI behind a local OpenAI-compatible shim, one CLI invocation per request
decoding control	none: a requested temperature, seed or token limit is recorded requested-and-ignored
output cap	none
prompt overhead	none: the frozen scaffold prompt replaces the CLI's default system prompt
transport conditions	one condition per plan
reasoning effort	n/a
Gemma-3-27B-it	
transport	OpenAI-compatible HTTP to a self-hosted server
decoding control	temperature 0.0 / top_p 1.0 applied on the wire
output cap	65536 tokens on BM25 and Graph-RAG
prompt overhead	none
transport conditions	one condition
reasoning effort	n/a
GPT-5.6	
transport	`codex-cli 0.155.1` (subscription CLI) behind a local shim
decoding control	none: 44,325 of 72,882 shim-recorded requests carry `temperature_ignored`, every request records the seed as absent
output cap	none the client can set
prompt overhead	10628–10634 prompt tokens of vendor harness preamble (n=21 near-empty probes)
transport conditions	original (3,177), timeout1800 (450), timeout3600 (18)
reasoning effort	`--reasoning-effort medium`
GPT-6	
transport	`codex-cli 0.155.1` (subscription CLI) behind a local shim
decoding control	none: 44,712 of 73,728 shim-recorded requests carry `temperature_ignored`, every request records the seed as absent
output cap	none the client can set
prompt overhead	11842–11850 prompt tokens of vendor harness preamble (n=24 near-empty probes)
transport conditions	original (2,772), timeout1800 (819), timeout1800-https-only-v1 (54)
reasoning effort	`--reasoning-effort medium`
Qwen3.8-27B (thinking enabled)	
transport	OpenAI-compatible HTTP to a self-hosted server
decoding control	pending verification: the bundle config records temperature 0.0 / top_p 1.0, and this row's own `llm_config` is on the compute node, so what reached the wire is not established here
output cap	a 28,672-token reasoning allowance added to the answer cap, not a cap the server enforces on the trace
prompt overhead	none
transport conditions	one condition: request timeout 7,200 s, per cell 28,800 s
reasoning effort	thinking enabled

Table 27: The second controlled model axis: GPT-6 − GPT-5.6 on the codex exec transport, where the model is the only declared field that differs. Paired by (world, layout, scaffold, intervention) over the 3,645 cells both arms scored, with the same school-clustered bootstrap as the rest of the paper ($B = 10{,}000$, seed 20260901, 48 clusters); neither direction is conditioned on the other's completion. *This transport carries no repetition slice*, so an interval excluding zero is separable under resampling of worlds and schools and is never evidence of a difference beyond run-to-run variation; the transport also enforces no temperature or seed, the two arms' harness preambles differ by roughly 1.2k tokens, and their cells are spread over 4 transport conditions. The one-shot row is a strict worksheet-contract effect and not a capability reading (Appendix S.1). This axis is never subtracted from or ranked against the Sonnet 5 → Opus 5 axis: different transport, family and condition.

Scaffold	Δ Entity F_1	Δ Omission	Δ SemAcc	Δ WorldPass
BM25+LLM	-0.137 $[-0.270, -0.031]$	$+0.047$ $[-0.019, +0.109]$	-0.084 $[-0.213, +0.033]$	-0.012 $[-0.045, +0.000]$
Graph-RAG	-0.070 $[-0.154, -0.001]$	$+0.041$ $[-0.016, +0.094]$	-0.062 $[-0.182, +0.037]$	$+0.000$ $[+0.000, +0.000]$
One-shot	-0.581 $[-0.719, -0.428]$	$+0.314$ $[+0.203, +0.421]$	-0.498 $[-0.642, -0.368]$	-0.037 $[-0.136, +0.000]$
ReAct-style browsing	$+0.106$ $[+0.008, +0.212]$	-0.124 $[-0.185, -0.051]$	$+0.112$ $[+0.018, +0.188]$	-0.012 $[-0.045, +0.000]$
Structured ledger	$+0.000$ $[+0.000, +0.000]$	-0.018 $[-0.080, +0.016]$	$+0.013$ $[-0.026, +0.081]$	$+0.000$ $[+0.000, +0.000]$
Pooled (all scaffolds)	-0.136 $[-0.197, -0.085]$	$+0.052$ $[+0.020, +0.079]$	-0.104 $[-0.167, -0.050]$	-0.012 $[-0.045, +0.000]$

Table 28: What happens to a reference obligation, counted over every scored cell of every reported run (cell × obligation observations; micro-counts, not the school-macro-averaged semantic accuracy of the main tables). 361,089 observations carry a reference obligation, of which 252,108 are decidable by the reference; the shares below are over that decidable denominator. Omission dominates, but an omission may equally come from an empty or contract-failing output, so this table decomposes outcomes and does not attribute causes. The lower block splits the proven counterexamples by their leading difference; the grouping is automatic, not a human root-cause label.

Outcome of a decidable reference obligation	Count	% of decidable
Equivalent to the reference obligation	60,507	24.00
Whole obligation absent from the reconstruction	190,296	75.48
Present but proven non-equivalent (counterexample found)	837	0.33
Present but the reconstruction abstains	432	0.17
Present but outside the supported fragment	9	0.00
Undecidable at evaluation time	27	0.01
Evaluator disagreement	0	0.00
the 837 proven counterexamples, by leading difference		
course set	603	72.0
grade condition	189	22.6
credits or units	27	3.2
logic or gate count	18	2.2

Table 30: Discovery against interpretation on the frozen institutional websites (%). Upper block: each arm's own result. Lower block: the two pairs that differ in *exactly one* declared field—whether the agent starts at the institution's root page or is handed the pages the reference itself cites as evidence for that world—each paired by world over the worlds both of its arms scored, 95% school-clustered intervals. Handing the evidence over lifts the capacity-limited structured scaffold on every metric that moves; it does not lift the search-capable browsing agent, whose requirement omission separably worsens on the weaker model and moves nothing separably on the stronger one, where no world is dropped from the pairing either. Of the three single-field contrasts only the structured one improves anything: on the weaker model's search agent the single separable movement is a worsening of omission, and on the stronger model nothing separates at all. The two contrasts also differ in agent capacity, so the difference between them is a reading, not an estimated effect. The oracle arms depend on the reference by construction and are an upper bound on discovery, not a deployable system; the step from the first row to the second changes the agent's capacity *and* the budget rule at once and is never an ablation of the search action. Strict world success is zero on every arm here, so that column carries no information and is not contrasted. The last block moves the base model instead, with source access held at root only: the stronger model is ahead on every mean and not one of the three intervals excludes zero, so it is a level comparison between two arms, not an ablation, and this condition carries no repetition against which to read it.

Arm	Cells	Entity F_1 ↑	Omission↓	SemAcc↑	WorldPass↑
Browsing, root only (protocol as transferred)	225	39.2	96.4	11.4	0.0
In-page search, root only	225	51.8	93.9	12.0	0.0
In-page search, evidence handed over	225	44.8	95.2	10.1	0.0
In-page search, root only (stronger model)	225	61.9	90.2	13.5	0.0
In-page search, evidence handed over (stronger model)	225	60.1	92.0	15.2	0.0
Structured ledger, root only (as transferred)	225	31.2	100.0	0.0	0.0
Structured ledger, root only (contrast baseline)	225	31.3	100.0	0.0	0.0
Structured ledger, evidence handed over	198	53.3	89.4	18.6	0.0
Structured ledger, root only (stronger model)	225	31.3	100.0	0.0	0.0
Structured ledger, evidence handed over (stronger model)	225	47.5	89.5	16.8	0.0
evidence handed over, in-page search (225 paired worlds)					
Entity F_1	-7.0 [-14.8, +1.1] pp			not separable from zero	
Requirement omission	+1.3 [+0.5, +2.2] pp			separable from zero	
Semantic accuracy	-1.9 [-14.3, +6.0] pp			not separable from zero	
evidence handed over, structured ledger (198 paired worlds)					
Entity F_1	+17.7 [+6.9, +30.6] pp			separable from zero	
Requirement omission	-10.6 [-25.3, -2.1] pp			separable from zero	
Semantic accuracy	+18.6 [+4.4, +43.6] pp			separable from zero	
evidence handed over, in-page search, stronger model (225 worlds, none dropped)					
Entity F_1	-1.8 [-6.7, +3.3] pp			not separable from zero	
Requirement omission	+1.8 [-0.9, +6.3] pp			not separable from zero	
Semantic accuracy	+1.7 [-8.2, +14.8] pp			not separable from zero	
evidence handed over, structured ledger, stronger model (225 worlds, none dropped)					
Entity F_1	+16.1 [+6.3, +28.2] pp			separable from zero	
Requirement omission	-10.5 [-23.4, -2.9] pp			separable from zero	
Semantic accuracy	+16.8 [+4.4, +38.1] pp			separable from zero	
the model axis, source access held at root only (225 paired worlds)					
Entity F_1	+10.1 [-0.1, +19.7] pp			not separable from zero	
Requirement omission	-3.7 [-7.6, +0.5] pp			not separable from zero	
Semantic accuracy	+1.5 [-11.8, +8.9] pp			not separable from zero	

Table 31: The same arms on the 243 scored worlds, on the 63 worlds whose reference is complete, and on the 180 whose reference is diagnostic (%). Values are re-aggregations of already-scored cells by the paper's own renderer; on the full set they reproduce the published tables value for value. The split is how far reference adjudication got, not a randomisation: a difference between two columns describes two world sets and estimates no effect. The complete-reference columns rest on 18 school clusters against 48, so they are indicative rather than inferential and are marked $*$. The WorldPass columns are instrument-bounded: a strict pass is attainable on 18 of the 243 scored worlds and on none of the diagnostic ones, so their zeros describe the references (Appendix R.1).

	Entity F_1 ↑			Omission↓			SemAcc↑			WorldPass↑		
Arm	243	63$*$	180	243	63$*$	180	243	63$*$	180	243	63$*$	180
Sonnet 5, main grid	69.1	70.3	68.7	79.2	86.5	76.6	31.0	21.7	33.8	1.2	4.5	0.0
Opus 5, main grid	71.9	77.8	69.9	69.9	70.3	69.8	49.4	44.9	50.8	3.0	11.4	0.0
Gemma-3-27B-it	45.2	46.0	45.0	96.0	97.5	95.5	6.7	4.1	7.5	0.2	1.0	0.0
GPT-5.6	69.2	70.2	68.9	73.8	74.6	73.6	41.9	39.0	42.7	2.5	9.5	0.0
GPT-6	55.6	58.9	54.4	79.0	82.6	77.8	31.5	27.7	32.6	1.2	4.8	0.0
Qwen3.8-27B (thinking)	53.4	54.9	52.9	84.5	85.9	84.0	23.5	19.8	24.5	1.7	6.7	0.0
Sonnet 5, real site, in-page search	51.8	47.8	53.3	93.9	98.6	92.1	12.0	7.1	13.6	0.0	0.0	0.0
Opus 5, real site, in-page search	61.9	67.9	59.6	90.2	89.5	90.5	13.5	17.2	12.3	0.0	0.0	0.0

Table 32: Evaluable coverage $|Q_S|/|Q|$ per reported run: of the reference obligations the scorer enumerated over that run's own cells, the share it could decide. This is a property of the references and the supported fragment, not of the agent: an obligation the reference leaves undecidable is outside every denominator, so no prediction is charged for it. The complement is the reference-coverage limitation the protocol declares rather than a model failure. Oracle and ablation cells are excluded, exactly as in every other table.

Condition	Runs	Cells	Obligations	Decidable	$\lvert Q_S\rvert/\lvert Q\rvert$ (%)
Three re-rendered layouts	6	3,645	76,680	53,595	69.9
One re-rendered layout	2	1,206	25,488	17,802	69.8
Frozen original websites	2	1,116	24,768	17,199	69.4

Table 33: Worksheet contract v1.1: a field line may carry any leading whitespace instead of exactly two spaces. The contract was installed identically for every reported row and no stored answer was changed; 2,097 previously failed one-shot cells with a stored reply are re-parsed, of which 639 are accepted, and all of them belong to a single model row. Every other row is value-identical under the two contracts, verified by re-scoring whole runs and comparing every scored field: 1,162,944 values compared, zero changed. The main-text Table 1 reports the relaxed contract, where this is the only row that changes. Every other table keeps the strict contract, because only a subset of the reported rows was re-scored under v1.1.

	Entity F_1 ↑		SemAcc↑		WorldPass↑	
GPT-6 scaffold	strict	lenient	strict	lenient	strict	lenient
One-shot	20.15	72.05	0.00	51.36	0.00	3.70
BM25+LLM	31.01	31.01	15.71	15.71	0.00	0.00
Graph-RAG	72.31	72.31	54.75	54.75	3.70	3.70
ReAct-style browsing	85.40	85.40	53.71	53.71	2.47	2.47
Structured ledger	68.94	68.94	33.26	33.26	0.00	0.00