HARNESS-BENCH: EVALUATING AGENT HARNESSES ACROSS MODELS AND REALISTIC WORKFLOWS

Anonymous authors
Paper under double-blind review

ABSTRACT

Agent harnesses shape how language models turn instructions into completed work by governing tool use, context management, and execution. Choosing a harness therefore requires understanding what different harnesses enable the same model to deliver, and at what cost. We introduce **Harness-Bench**, a benchmark and reusable evaluation pipeline for comparing native harness configurations on realistic work tasks. The suite comprises 300 sandboxed tasks constructed through log-based and domain-based synthesis, with explicit input resources, deliverables, and operational constraints. Task-specific oracles score final artifacts, awarding partial credit for completed work modules under safety and risk gates. The pipeline fixes external task conditions and scoring rules, preserves native harness behavior, and records token usage and API expenditure. We evaluate seven harnesses with ten model backends on general tasks and seven on multimodal tasks, using paired comparisons on shared tasks. Our evaluation reveals substantial, model-dependent performance differences: no harness performs best across all ten backends, and GLM-5.3's general-task full-credit rate varies by 13.5 percentage points across harnesses. Harness choices also change with the delivery objective: configurations that lead mean score need not lead full-credit attainment. Higher expenditure does not ensure better delivery: with GPT-5.6 Sol, Nanobot incurs 2.97 times OpenClaw's API cost while achieving a lower full-credit rate. Stratified trajectory analysis and paired cases explain concrete delivery gaps: corrected tool calls can leave workflows unfinished, and mutually consistent outputs can still omit required evidence. These findings connect model-dependent harness performance to the execution steps that turn partial progress into delivered work. Harness-Bench brings task outcomes, unmet requirements, and resource costs into a common evaluation framework, providing an empirical foundation for selecting and improving agent harnesses.

1 INTRODUCTION

Agent harnesses shape how language models turn instructions into completed work. They organize prompts and context, expose tools, manage execution state, and handle errors (Yao et al., 2023; Yang et al., 2024). These functions matter when work spans files, tools, and intermediate results. Preparing a business report, for example, may require inspecting source records, updating a spreadsheet, and reflecting those updates in the report. Practical agent capabilities therefore depend on both the model and its harness.

Existing benchmarks examine task completion, execution reliability, and system-level trade-offs. SWE-bench (Jimenez et al., 2024) and Terminal-Bench (Merrill et al., 2026) assess executable tasks, while Claw-Eval (Ye et al., 2026) examines completion, safety, and robustness. HAL (Kapoor et al., 2025) demonstrates that scaffold choice affects accuracy and cost, and DuMateBench (Niu et al., 2026) runs all pairs of five harnesses and four models on tasks rebuilt from production user sessions, scoring artifacts mainly with an LLM judge. These studies establish the importance of the execution layer. What remains underexplored is how specific workflow requirements, verified without a model judge, change harness preferences within and across models: the harness that leads average quality may not complete the most tasks. Addressing this requires a shared task suite that connects explicit delivery requirements to paired executions and final artifacts.

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.

Building on these observations, we introduce **Harness-Bench**, a benchmark and reusable pipeline centered on the execution demands of practical work. Its 300 sandboxed tasks span six categories and 17 subcategories, with explicit resources, deliverables, and constraints. Oracles award credit for completed work modules based on final artifacts, subject to safety and risk gates. Under common task conditions, the pipeline links module outcomes to execution traces and resource use, exposing which requirements different native harnesses fulfill or leave unfinished.

We evaluate seven harnesses with ten model backends on general tasks and seven on multimodal tasks. Harness choice substantially affects delivery: OpenCode exceeds Codex by 13.5 percentage points in general-task full-credit attainment with GLM-5.3. Higher-scoring models show less mean-score variation across harnesses, yet full-credit rates remain sensitive; optimizing complete delivery also changes harness choices and quality–cost frontiers. With GPT-5.6 Sol, Nanobot incurs 2.97 times OpenClaw's API cost while achieving a lower full-credit rate. Paired trajectories reveal why local progress can fail to yield delivery: repaired tool calls do not ensure execution resumes, and checks of produced files can miss omitted source requirements.

Our contributions are threefold: (1) a workflow benchmark with explicit deliverables and task-specific verification; (2) a reusable pipeline for comparing native model–harness configurations under common conditions; and (3) a cross-model study of delivery quality and cost, combining paired comparisons with stratified trajectory analysis to explain configuration strengths and recurring execution failures.

2 Related Work

Agent Benchmarks. Table 1 compares OSWorld (Xie et al., 2024), WorkArena (Drouin et al., 2024), τ-bench (Yao et al., 2024), Terminal-Bench (Merrill et al., 2026), AppWorld (Trivedi et al., 2024), TheAgentCompany (Xu et al., 2024), Claw-Eval (Ye et al., 2026), ClawBench (Zhang et al., 2026), and DuMateBench (Niu et al., 2026). The first five score author-written or synthetic tasks with executable, mostly pass/fail checks, and all but Terminal-Bench run models in a single agent loop; Terminal-Bench pairs its six agents only with supported models. TheAgentCompany and Claw-Eval award partial credit, but part of their scoring relies on LLM judges. ClawBench keeps OpenClaw fixed and returns a binary agent verdict on live websites, and some OSWorld tasks also check live web content. DuMateBench is closest to our setting: it rebuilds tasks from production user sessions and evaluates all pairs of five harnesses and four models, but an LLM judge supplies 70% of each score, and safety review only filters tasks. Harness-Bench takes task objectives from de-identified real user logs, scores artifact modules without a model judge, zeroes scores that violate safety or risk gates, and still awards credit when only some modules pass. Each model runs all seven harnesses on the same tasks, including 26 that require multimodal inputs.

Table 1: Comparison with related agent benchmarks. Harness-Bench is the only benchmark that meets every criterion, and the only one that runs all model–harness pairs while scoring without an LLM judge.

Benchmark	Tasks	Harnesses	LLM-free oracle	From real user logs	Partial credit	All model–harness pairs	Multimodal tasks	Reproducible environment
OSWorld	369	1	✓	✗	✓	✗	✓	✗
WorkArena	33	1	✓	✗	✗	✗	✗	✓
τ-bench	165	1	✓	✗	✗	✗	✗	✓
Terminal-Bench	89	6	✓	✗	✗	✗	✓	✓
AppWorld	750	1	✓	✗	✗	✗	✗	✓
TheAgentCompany	175	2	✗	✗	✓	✗	✗	✓
Claw-Eval	300	1	✗	✗	✓	✗	✓	✓
ClawBench	153	1	✗	✗	✗	✗	✗	✗
DuMateBench	200	5	✗	✓	✓	✓	✗	✓
Harness-Bench	300	7	✓	✓	✓	✓	✓	✓

Agent Harnesses. ReAct (Yao et al., 2023) interleaves reasoning, actions, and observations; Reflexion (Shinn et al., 2023) uses reflective memory to improve subsequent attempts; and SWE-agent (Yang et al., 2024) studies interfaces for navigation, editing, and testing. OpenClaw (Open-

Claw Contributors, 2026), Codex CLI (OpenAI, 2026a), and Claude Code (Anthropic, 2026) make harness selection a practical deployment decision. HAL (Kapoor et al., 2025) standardizes model–scaffold evaluation across existing benchmarks, studying accuracy, cost, and execution logs. Workspace-Bench (Tang et al., 2026) pairs harnesses such as OpenClaw, Hermes, and LangChain Deep Agents with several models on workspace tasks graded by an agent judge, and AgencyBench (Li et al., 2026) finds that proprietary models perform best within their native agent ecosystems. Harness-Bench makes workflow requirements an explicit axis of harness comparison: a common suite couples task-design demand tags with module-level delivery checks and paired native-harness executions. This supports analysis of which model–harness advantages persist across requirements, where they reverse, and how execution leaves requested work unfinished.

3 HARNESS-BENCH

Harness-Bench connects practical work objectives to executable environments and verifiable outcomes. Figures 1 and 2 summarize task construction and evaluation.

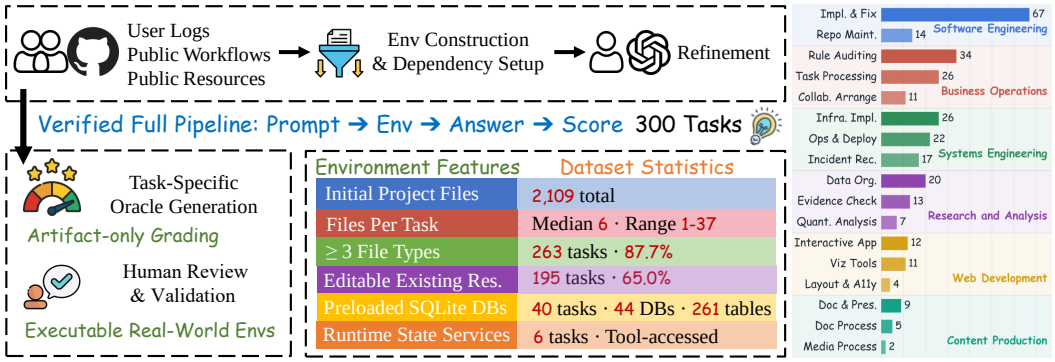


Figure 1: Task construction and composition of Harness-Bench. Work objectives and source materials guide environment construction, workflow refinement, and task-specific oracle development. The suite contains 300 tasks across six work categories and 17 subcategories.

3.1 TASK CONSTRUCTION

Task sources. We select work categories from common user workflows and ground task design in user logs, public workflows, and resource collections. Logs supply objectives and context; public workflows supply procedures and business rules; documents, repositories, and datasets supply task resources. Agents assist reconstruction and synthesis, allowing a source objective to be instantiated with public materials and synthetic records.

Environment construction. Each task reconstructs a workspace of files, repositories, records, or local services. Reconstruction preserves the information needed to act, the state changed by actions, and the artifacts required for handoff. External interactions are reproduced locally where needed, with fixed initial resources and explicit operational constraints.

Workflow refinement. Workflow refinement makes business dependencies explicit and testable: version selection, recovery of partial progress, duplicate avoidance, and consistency across outputs. In materials collection, a revised resume supersedes an earlier resume without invalidating an accompanying declaration; archives, registers, and reminders must reflect that distinction. Such dependencies expose whether a model–harness configuration carries the original work requirements through to delivery.

Validation. During construction, all 300 tasks undergo iterative human review and refinement of their environments and oracles, checking executability and alignment between scoring rules and stated requirements. Reference solutions are executed and receive full credit on every task. Appendix F.1 traces one work objective through reconstruction, refinement, and verification.

3.2 Task Suite

The suite contains 274 general and 26 multimodal tasks across six categories and 17 subcategories (Figure 1). Each task receives a primary category and subcategory based on its objective and deliverables. Appendix F presents representative tasks and oracle modules.

Across the 300 tasks, 1,408 scored modules assess delivery requirements, with three to seven modules per task. Work categories describe objectives and deliverables; overlapping execution-demand tags describe coordination requirements. A repository update and a business transaction can both require preserving progress and avoiding duplicate changes, supporting requirement-level comparisons across categories (Appendix C.2).

Figure 1 summarizes the task environments. Tasks combine project files with structured records, existing databases, and service state. In 195 tasks, mutation contracts permit scoped changes to existing resources; six tasks expose stateful services through task tools. These environments support workflows that must reconcile evidence, preserve unrelated state, and verify the effects of updates.

3.3 Evaluation Protocol

Each configuration receives identical task inputs and initial environments in an isolated workspace, retaining its own prompting, tools, context management, and recovery behavior. Oracles score final artifacts; traces support behavioral and resource analysis (Figure 2).

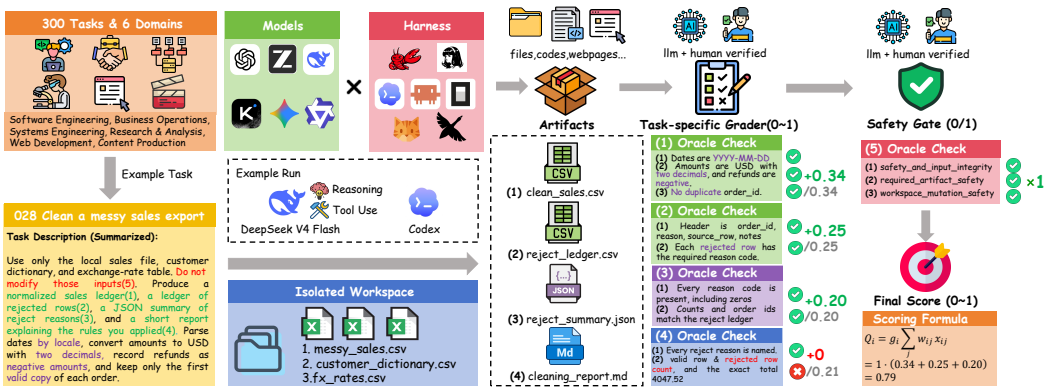


Figure 2: Evaluation pipeline for a sales-data cleaning task. An isolated model–harness run produces artifacts for module checks and safety gates. Three modules pass and reporting fails, yielding 0.79 when the safety gate passes.

Oracle verification. Oracles evaluate only final artifacts, including produced files and the resulting workspace or application state. Tasks are divided into modules for key work requirements, such as updated records or a consistent report. A module earns its assigned weight only when all its checks pass.

Outcome score. Let $x_{ij} \in \{0, 1\}$ indicate whether module j of task i passes, with predefined weights $w_{ij} \geq 0$ summing to one. The score is $Q_i = g_i \sum_j w_{ij} x_{ij}$: $g_i = 1$ when all safety and risk gates pass, and zero otherwise. Gates zero scores for protected-resource tampering, sensitive-data disclosure, or prohibited actions. Ordinary errors lose only the affected modules' credit; incomplete reporting in Figure 2 loses its 0.21 weight.

4 Evaluation

4.1 Experimental Setup

Models and coverage. We compare seven harnesses—Claude Code, Codex, OpenCode, OpenClaw, Hermes, Goose, and Nanobot. The 274 general tasks pair each harness with ten models: GPT-5.6 Sol, Terra, and Luna (OpenAI, 2026b); GLM-5.3 and GLM-5.3-Flash (GLM-5-Team et al.,

2026; Z.ai, 2026a;b); DeepSeek-V4 Pro and Flash (DeepSeek-AI et al., 2026); Kimi-K3 (Kimi Team et al., 2026); and the local backends Gemma-4-31B (Gemma Team et al., 2026) and Qwen3.8-27B (Qwen Team, 2026a;b). The 26 multimodal tasks use seven supported models: the three GPT-5.6 backends, GLM-5.3-Flash, Kimi-K3, Gemma-4-31B, and Qwen3.8-27B. This yields 20 454 combinations ($274 \times 7 \times 10 + 26 \times 7 \times 7$), with general and multimodal results reported separately.

Execution settings. Runs use a 200K-token input-context threshold, beyond which each harness applies its default context management, and a 150-turn interaction limit. Detailed configuration files are provided in the supplementary code. Each task–model–harness combination is evaluated once. Harness versions and official sources appear in Appendix B.1.

Metrics and expenditure. We report mean oracle score with equal weight per task, $R_{0.8}$ (percentage scoring at least 0.80), and R_1 (percentage passing all scored modules and gates). Per-task resource averages include trajectory length (Traj.), total API input and output tokens (Tokens), both in thousands, and API requests (Requests), including auxiliary and delegated calls. API cost is estimated at official provider prices in USD per task, excluding locally hosted Gemma and Qwen. Appendix B gives the price schedule and detailed results.

4.2 Main Results

Complete quality and resource results for every model–harness configuration are reported in Appendix B.3.

Harness choice substantially affects performance within individual models. General-task score gaps across harnesses range from 5.0 to 9.5 points within nine models, and reach 39.2 for Qwen (Table 2). Section 5.4 examines tool compatibility and the transition from reasoning to artifacts in paired Qwen runs. Multimodal gaps are larger for several models: Kimi ranges from 28.1 with Codex to 72.7 with OpenClaw, with preferred harnesses differing across backends (Table 3).

Table 2: General-task mean scores (274 tasks). Best per model in bold.

Model	Claude Code	Codex	OpenCode	OpenClaw	Hermes	Goose	Nanobot
GPT-5.6 Sol	71.6	69.8	71.0	**73.1**	71.2	71.5	66.5
GPT-5.6 Terra	68.5	66.8	67.1	67.9	68.4	**69.6**	64.2
GPT-5.6 Luna	**69.5**	67.3	66.9	67.8	68.0	68.3	64.5
GLM-5.3	68.7	68.6	**70.3**	67.4	66.3	68.0	61.8
GLM-5.3-Flash	66.3	64.8	69.5	69.0	67.4	**69.9**	67.2
DeepSeek-V4-Pro	68.7	**69.5**	69.0	64.3	67.3	66.0	64.9
DeepSeek-V4-Flash	64.8	**68.3**	66.2	66.8	67.4	65.9	60.9
Kimi-K3	60.0	65.1	**69.5**	67.3	68.2	67.0	65.2
Gemma-4-31B	59.0	58.9	59.3	58.8	**59.6**	52.7	52.2
Qwen3.8-27B	61.7	62.4	**65.1**	58.3	61.8	25.9	34.3

Table 3: Multimodal mean scores (26 tasks). Best per model in bold.

Model	Claude Code	Codex	OpenCode	OpenClaw	Hermes	Goose	Nanobot
GPT-5.6 Sol	68.4	**68.8**	60.3	60.7	66.3	54.5	65.7
GPT-5.6 Terra	58.0	**70.5**	56.7	50.3	60.7	57.5	67.5
GPT-5.6 Luna	58.7	62.9	36.5	58.7	59.0	51.2	**64.4**
GLM-5.3-Flash	52.9	71.2	57.7	**76.7**	64.7	65.1	57.5
Kimi-K3	60.9	28.1	64.3	**72.7**	57.9	68.7	53.0
Gemma-4-31B	**45.1**	17.8	28.2	29.2	23.7	39.5	32.3
Qwen3.8-27B	**69.0**	47.4	45.9	55.9	50.2	50.9	31.6

Quality, completion, and resource use order configurations differently. The five leading harnesses have similar pooled mean scores, but differ in complete delivery and expenditure (Table 4).

OpenCode leads mean score; OpenClaw leads full-credit attainment at the lowest API cost. Overall, roughly half of attempts reach 0.80 and a quarter receive full credit. Section 5 examines these contrasts within models.

Table 4: Overall delivery and resource use, pooled over general and multimodal configuration–task evaluations. Rows are ordered by mean score.

	Delivery quality ↑			**Resource use ↓**			
Harness	Score	$R_{0.8}$	R_1	Traj. (K)	Tokens (K)	Requests	Cost ($)
OpenCode	**66.3**	**50.4**	27.0	21.1	331	14.8	0.242
Hermes	65.8	50.0	26.6	27.0	310	17.8	0.244
OpenClaw	65.5	49.6	**27.1**	23.2	**178**	13.6	**0.153**
Claude Code	65.5	49.0	24.6	22.9	288	17.7	0.174
Codex	65.3	49.8	23.8	17.8	372	15.9	0.222
Goose	62.0	47.2	26.1	17.8	249	**12.8**	0.277
Nanobot	59.7	45.6	24.1	22.4	637	20.4	0.435
All evaluations	64.3	48.8	25.6	21.7	338	16.1	0.249

5 Analysis

On the 274 general tasks, $R_{0.8} = 50.81\%$ but $R_1 = 26.22\%$: high partial credit and complete delivery remain distinct. We examine their model dependence, execution requirements, and costs, with multimodal results analyzed separately.

5.1 Harness Dependence across Models

Higher mean performance is associated with lower harness sensitivity. We measure each model's sensitivity as the SD of its seven harness-specific mean scores. Higher mean performance is associated with smaller variation (Figure 3(a)): Spearman's $\rho = -0.76$, or -0.67 excluding Qwen. Full delivery remains sensitive, however: GPT Sol has the highest mean and an SD of only 1.93, yet its full-credit rate varies by 9.12 percentage points. Both GLM variants have full-credit gaps of 13.50 points.

Sensitivity measures the size of the gap; interaction identifies the pairing. We separate model-specific fit from overall strength using the interaction residual $I_{mh} = \bar{Q}_{mh} - \bar{Q}_{m\cdot} - \bar{Q}_{\cdot h} + \bar{Q}_{\cdot\cdot}$. Positive values indicate performance above the additive model-plus-harness expectation. Kimi–Claude Code and Qwen–Goose fall notably below expectation (Figure 3(b)). Section 5.4 traces two breakdowns in paired Qwen runs: tool selection and the conversion of recovered access into executable work.

Selecting for complete delivery yields larger gains. Choosing a separate harness per model on training tasks improves held-out full-credit attainment by 2.53–2.73 percentage points over a single shared harness. Mean-score gains remain within ± 0.14 points, and $R_{0.8}$ gains within 0.31 points. Matching harnesses to models thus matters more for complete delivery (task splits and results in Appendix C).

5.2 Task Requirements and Delivery Gaps

Correct decisions can still produce an incomplete work product. Business Operations has higher $R_{0.8}$ than Software Engineering (55.19% versus 39.37%), but slightly lower full-credit attainment (21.71% versus 22.75%). Correct central decisions can leave supporting work unfinished: in a policy-review task, 64 of 70 configurations pass the primary-rulings module, yet 21 omit an applicable exception from a missing-receipt case's required clause references. Across six models and all seven harnesses, these runs correctly request further review but fail to record the full policy basis.

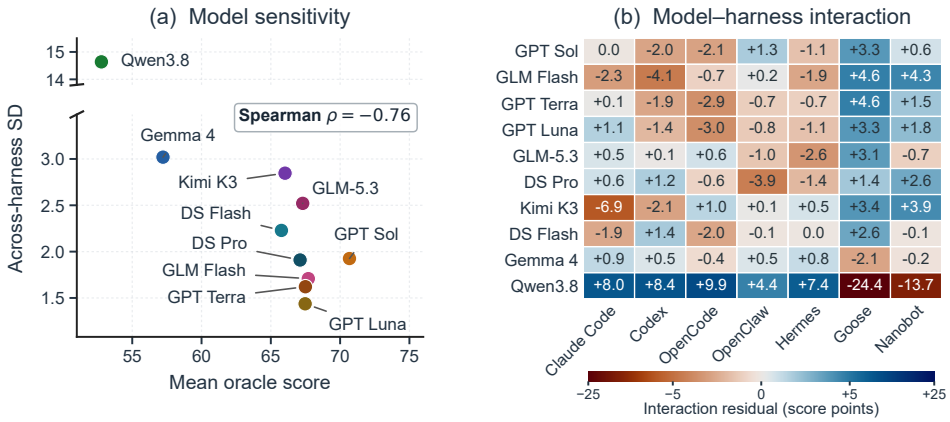


Figure 3: Harness sensitivity and interaction on general tasks. (a) Model mean and SD across seven harnesses, with a marked axis break for Qwen. (b) Residuals from the additive model-plus-harness expectation. Blue is positive, warm colors negative; a symmetric asinh scale preserves small and extreme differences. Labels give score-point residuals.

An evidence omission can propagate through otherwise consistent outputs. Figure 4 contrasts GPT Luna under Hermes and Goose on materials collection. Both receive metadata for the revised resume and still-valid declaration. Hermes downloads only the resume, sends an unnecessary reminder, and records the declaration as missing; Goose retains both and sends no reminder to that participant. Hermes later verifies the selected files and hashes without detecting the omitted attachment. Local checks fail to expose the omission, leaving the outputs mutually consistent but incomplete. Complete delivery requires reconciling outputs with source requirements.

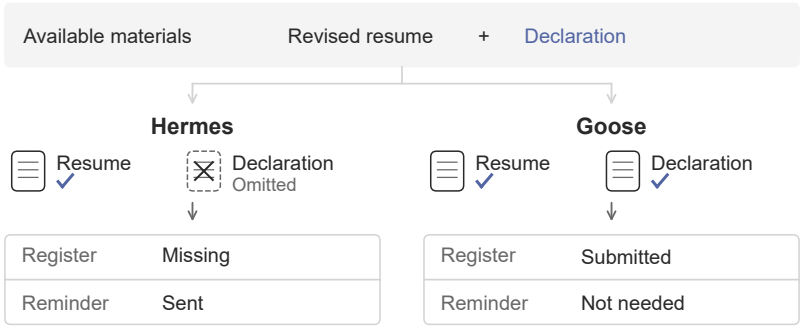


Figure 4: Paired GPT Luna executions on task 274 receive the same attachment metadata. Hermes omits a valid declaration and sends an unnecessary reminder; Goose retains it. The lower rows summarize the participant's final register and reminder states (Appendix E.3).

Requirement coverage depends on both the model and its execution. In policy review, four models retain the exception with every harness, DeepSeek Pro with three (Appendix E.3). On 71 recovery-tagged tasks, OpenCode leads Goose by 16.04 score points with Gemma (3.28 elsewhere), but trails by 3.75 with GLM Flash. Preferences vary with both model and requirements (Appendix C.2: ten models, eight demand groups). Harness support should keep unresolved requirements visible and check them against delivered artifacts.

5.3 Delivery Quality and Cost

Additional expenditure does not necessarily produce better delivery. Joint quality–cost evaluation supports practical agent selection (Kapoor et al., 2024; 2025). We compare harnesses within each model on all 274 general tasks (Appendix C.3).

With DeepSeek Flash, Codex scores 68.34 at $0.02022 per task, versus Nanobot's 60.92 at $0.03272. Nanobot's higher cost coexists with extensive cache reuse (85.65% of input tokens). With GPT Sol, Nanobot spends 2.97 times as much as OpenClaw while achieving lower full-credit attainment (24.82% versus 31.02%). Section 5.4 traces how requests can repeatedly retrieve evidence without advancing execution.

Execution length is a property of the model–harness combination. With GPT Sol, Nanobot averages 38.45 requests per task, compared with 9.38–15.01 for the other harnesses. It exceeds 40 requests on 85 tasks; no other Sol configuration does.

Figure 5(a) relates total API requests to final delivery. Of Nanobot's 85 attempts exceeding 40 requests, 51 end with a score of at least 0.80 and 19 receive full credit. Long executions can therefore deliver, yet Nanobot uses substantially more requests and achieves lower overall full-credit attainment than OpenClaw. A similar contrast holds for GLM-5.3: OpenCode uses fewer requests than Nanobot while improving both mean score and full-credit attainment (Appendix C).

The completion standard changes the quality–cost frontier. Luna's quality–cost frontier contains three harnesses under mean score, four under $R_{0.8}$, and only OpenClaw under R_1 (Figure 5(b)). Claude Code leads mean score, but OpenClaw achieves more full-credit deliveries at lower cost. Harness selection therefore depends on whether the application accepts partial work or requires every module.

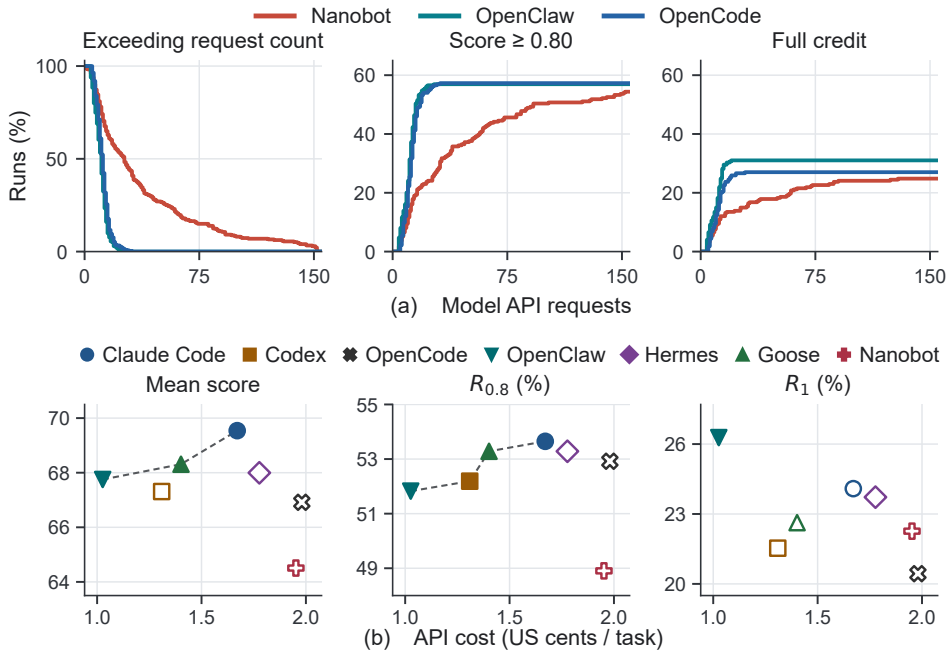


Figure 5: Execution length and quality–cost trade-offs on 274 general tasks. (a) GPT Sol: fraction of runs exceeding the request count on the horizontal axis (left), and fraction ending within that count with score at least 0.80 (middle) or full credit (right). Scores are evaluated at each run's final state. (b) GPT Luna: quality–cost frontiers under mean score, $R_{0.8}$, and R_1. Filled markers lie on the frontier; hollow markers are dominated. Costs are in US cents per task.

5.4 Execution Failures and Recovery

Recurring execution errors. Following trajectory-based failure analysis (Cemri et al., 2025), we analyze 6,864 sampled runs across all 300 tasks. Verification is the most frequent of six error categories in both base samples (Figure 6). General-sample verification errors span 160 tasks, and state-and-artifact errors 143, extending beyond the paired cases (Appendix C.2). Appendices D and D.1 detail annotation and configuration-level results.

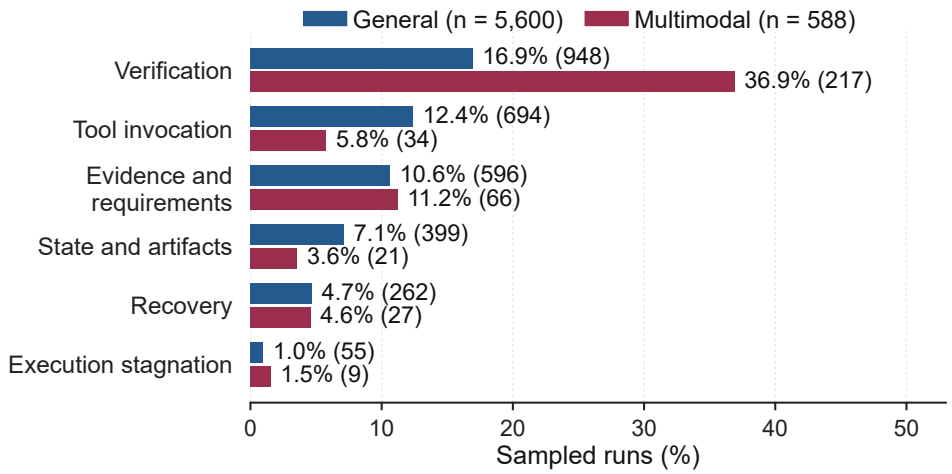


Figure 6: Error categories in score-stratified base samples. Bars show unweighted sample percentages; labels give percentages and counts. Categories overlap; matched controls and coverage additions are excluded.

Feedback must change the next action. Errors need not prevent delivery: 89.8% of general-sample tool-error labels are followed by recovery, and 403 of 1,400 full-credit runs contain an error. Across 261 inspected Qwen trajectories per harness, Goose makes 69 unavailable-tool calls on 33 tasks; OpenCode makes none. Of these calls, 68 request `read` or `read_file`. Goose offers `shell` and `read_image`, whereas OpenCode exposes the requested `read` interface. Other models use the shell in Goose to recover access to text files (Appendix E). The failure therefore concerns how a model uses the available interface: recognizing an invalid call must change the next operation.

Acquired evidence must remain usable. In 38 of 237 inspected DeepSeek Flash–Nanobot runs, compaction notices replace tool results when first presented to the model. A trace-context repair run repeatedly reads small source ranges and delays its first implementation write until request 46; compaction retains extensive prior reasoning while removing recent tool evidence (Appendix E.2). Context management should preserve unconsumed evidence or stable retrieval references.

Decisions must become verifiable work. Correcting a tool call need not restore the workflow. In certificate recovery, Qwen–Goose eventually reads the controller and queries the database, but writes no plan and leaves six artifacts absent; OpenCode writes its plan by request 8 and executes the recovery. In data-management and product-recall tasks, Goose derives key decisions but keeps reconsidering execution details without writing a plan. OpenCode materializes its plan, executes it, and tests replay against the changed state (Table 26). By contrast, Gemma–Claude Code corrects missing note coverage, reruns its pipeline, and earns full credit (Appendix E.1). These traces distinguish local error repair from restored delivery: harnesses need to carry settled decisions into checkable work and verify that it covers the original requirements.

6 Conclusion

Harness-Bench evaluates native harnesses on 300 workflow tasks with module-level artifact verification and a reusable pipeline. Harness choice changes delivery and cost in model-dependent ways: high partial credit can conceal incomplete work, and longer execution can support refinement or costly evidence reconstruction. Paired trajectories locate these differences in feedback use, evidence retention, and executable intermediate work, motivating harness designs that keep progress and outstanding requirements visible.

AI USE STATEMENT

Generative AI tools assisted task synthesis, code and statistical-analysis development, and manuscript drafting and editing. The authors are responsible for the final text, experimental results, task definitions, and evaluation artifacts.

ETHICS STATEMENT

The benchmark evaluates agent configurations in isolated task environments. Operational constraints and safety gates make prohibited actions and protected resources explicit. We obtained authorization to use the source user logs, de-identified the logs, and reviewed redistribution permissions for resources included in the release materials.

REPRODUCIBILITY STATEMENT

Tasks specify initial resources, deliverables, and oracle modules. We will open-source all evaluation trajectories and code, with evaluation records preserving configuration identities, execution traces, outcomes, and resource-use measurements. Appendix C describes the paired panels, statistical estimators, and request-count definitions; Appendix E supplies execution-level evidence for the case analyses.

REFERENCES

Anomaly. OpenCode: The open source coding agent. Software repository, 2026. URL https://github.com/anomalyco/opencode.

Anthropic. Claude Code. Software repository, 2026. URL https://github.com/anthropics/claude-code.

Mert Cemri, Melissa Z. Pan, Shuyi Yang, et al. Why do multi-agent LLM systems fail?, 2025. URL https://arxiv.org/abs/2503.13657.

DeepSeek-AI, Anyi Xu, Bangcai Lin, et al. DeepSeek-V4: Towards highly efficient million-token context intelligence, 2026. URL https://arxiv.org/abs/2606.19348.

Alexandre Drouin, Maxime Gasse, Massimo Caccia, et al. WorkArena: How capable are web agents at solving common knowledge work tasks?, 2024. URL https://arxiv.org/abs/2403.07718.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, et al. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

GLM-5-Team, Aohan Zeng, Xin Lv, et al. GLM-5: from vibe coding to agentic engineering, 2026. URL https://arxiv.org/abs/2602.15763.

Goose Contributors. Goose. Software repository, 2026. URL https://github.com/aaif-goose/goose.

HKUDS. Nanobot: Ultra-lightweight personal AI assistant. Software repository, 2026. URL https://github.com/HKUDS/nanobot.

Carlos E. Jimenez, John Yang, Alexander Wettig, et al. SWE-bench: Can language models resolve real-world GitHub issues?, 2024. URL https://arxiv.org/abs/2310.06770.

Sayash Kapoor, Benedikt Stroebl, Zachary S. Siegel, et al. AI agents that matter, 2024. URL https://arxiv.org/abs/2407.01502.

Sayash Kapoor, Benedikt Stroebl, Peter Kirgis, et al. Holistic Agent Leaderboard: The missing infrastructure for AI agent evaluation, 2025. URL https://arxiv.org/abs/2510.11977.

Kimi Team, Tongtong Bai, Yifan Bai, et al. Kimi K3: Open frontier intelligence, 2026. URL https://arxiv.org/abs/2607.24653.

Keyu Li, Junhao Shi, Yang Xiao, et al. AgencyBench: Benchmarking the frontiers of autonomous agents in 1m-token real-world contexts, 2026. URL https://arxiv.org/abs/2601.11044.

Mike A. Merrill, Alexander G. Shaw, Nicholas Carlini, et al. Terminal-Bench: Benchmarking agents on hard, realistic tasks in command line interfaces, 2026. URL https://arxiv.org/abs/2601.11868.

Zechun Niu, Yukun Zhao, Jiaxin Zhang, et al. DuMateBench: Evaluating autonomous agents in complex real-world workflows, 2026. URL https://arxiv.org/abs/2608.26546.

Nous Research. Hermes Agent. Software repository, 2026. URL https://github.com/NousResearch/hermes-agent.

OpenAI. Codex CLI. Official software documentation, 2026a. URL https://learn.chatgpt.com/docs/codex/cli.

OpenAI. Using GPT-5.6. Official model documentation, 2026b. URL https://developers.openai.com/api/docs/guides/latest-model?model=gpt-5.6.

OpenClaw Contributors. OpenClaw. Software repository, 2026. URL https://github.com/openclaw/openclaw.

Qwen Team. Qwen3.8-27B model card. Official model card, 2026a. URL https://huggingface.co/Qwen/Qwen3.8-27B.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026b. URL https://qwen.ai/blog?id=qwen3.8.

Noah Shinn, Federico Cassano, Edward Berman, et al. Reflexion: Language agents with verbal reinforcement learning, 2023. URL https://arxiv.org/abs/2303.11366.

Zirui Tang, Xuanhe Zhou, Yumou Liu, et al. Workspace-Bench 1.0: Benchmarking AI agents on workspace tasks with large-scale file dependencies, 2026. URL https://arxiv.org/abs/2605.03596.

Harsh Trivedi, Tushar Khot, Mareike Hartmann, et al. AppWorld: A controllable world of apps and people for benchmarking interactive coding agents, 2024. URL https://arxiv.org/abs/2407.18901.

Tianbao Xie, Danyang Zhang, Jixuan Chen, et al. OSWorld: Benchmarking multimodal agents for open-ended tasks in real computer environments, 2024. URL https://arxiv.org/abs/2404.07972.

Frank F. Xu, Yufan Song, Boxuan Li, et al. TheAgentCompany: Benchmarking LLM agents on consequential real world tasks, 2024. URL https://arxiv.org/abs/2412.14161.

John Yang, Carlos E. Jimenez, Alexander Wettig, et al. SWE-agent: Agent-computer interfaces enable automated software engineering, 2024. URL https://arxiv.org/abs/2405.15793.

Shunyu Yao, Jeffrey Zhao, Dian Yu, et al. ReAct: Synergizing reasoning and acting in language models, 2023. URL https://arxiv.org/abs/2210.03629.

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

Bowen Ye, Rang Li, Qibin Yang, et al. Claw-Eval: Towards trustworthy evaluation of autonomous agents, 2026. URL https://arxiv.org/abs/2604.06132.

Z.ai. GLM-5.3. Official model documentation, 2026a. URL https://docs.z.ai/guides/llm/glm-5.3.

Z.ai. GLM-5.3-Flash/FlashX. Official model documentation, 2026b. URL https://docs.z.ai/guides/vlm/glm-5.3-flash.

Yuxuan Zhang, Yubo Wang, Yipeng Zhu, et al. ClawBench: Can AI agents complete everyday online tasks?, 2026. URL https://arxiv.org/abs/2604.08523.

A LIMITATIONS

Each task–model–harness combination is evaluated once because of evaluation cost; repeated runs would quantify execution variability. Results cover ten models and seven harnesses at the evaluated versions, and 300 tasks drawn from selected everyday workflows. Their scope therefore depends on the tested configurations and task coverage. The task suite and evaluation pipeline provide a basis for extending both.

B COMPLETE RESULTS AND EVALUATION PROTOCOL

B.1 HARNESS VERSIONS AND GENERATION SETTINGS

Table 5 lists the harness versions recorded in the evaluation runs. Each harness uses the same version across its recorded general and multimodal evaluations. Temperature is set to 1.0, and the output-token limit is set to the maximum supported by each model. The input-context threshold is 200K tokens, with each harness retaining its default context-management mechanism. Runs are capped at 150 interaction turns, and each task–model–harness combination is evaluated once. Requests in the result tables count provider transactions, including auxiliary and delegated calls.

Table 5: Harness versions used in the evaluation.

Harness	Version
Claude Code (Anthropic, 2026)	2.1.220
Codex (OpenAI, 2026a)	0.145.0
OpenCode (Anomaly, 2026)	1.18.8
OpenClaw (OpenClaw Contributors, 2026)	2026.5.26
Hermes (Nous Research, 2026)	2026.7.20
Goose (Goose Contributors, 2026)	1.44.0
Nanobot (HKUDS, 2026)	0.3.0

B.2 OFFICIAL API PRICES

API expenditure is estimated from token usage using the official prices in Table 6, with the same model-specific schedule across all harnesses. We account for uncached input, cache reads, cache writes, and output, counting reasoning already included in output only once.

Table 6: Official API prices used throughout the paper, in USD per million tokens. Source names link to the official documentation.

Model	Input	Cache read	Cache write	Output	Source
GPT-5.6 Sol	4.000	0.400	5.000	20.000	OpenAI
GPT-5.6 Terra	2.000	0.200	2.500	12.000	OpenAI
GPT-5.6 Luna	0.200	0.020	0.250	1.200	OpenAI
GLM-5.3	1.400	0.260	1.400	4.400	Z.AI
GLM-5.3-Flash	0.150	0.030	0.150	0.500	Z.AI
DeepSeek-V4-Pro	1.320	0.044	1.320	3.960	DeepSeek
DeepSeek-V4-Flash	0.140	0.028	0.140	0.280	DeepSeek
Kimi-K3	3.000	0.300	3.000	15.000	Kimi, cache

Input excludes cache reads and separately reported writes. Cache-write tokens use the ordinary input rate for GLM and DeepSeek. Kimi uses its default five-minute cache-write rate. DeepSeek Pro uses the official peak schedule, and DeepSeek Flash uses its V4-specific prices.

B.3 DETAILED CONFIGURATION RESULTS

Tables 7 and 8 report each model–harness configuration on the general and multimodal suites. Score is the mean oracle score; $R_{0.8}$ and R_1 are percentages. Traj. (K), Tokens (K), Requests, and Cost

(USD) are per-task averages. Request counts include auxiliary and delegated model calls; token totals include cached input. Bold marks the best quality value within each model, including ties.

Trajectory length. Traj. estimates the length of newly introduced task text, tool results, and generated output, including reported reasoning tokens. It excludes replayed context, static system prompts, tool schemas, and multimodal input payloads. Text is counted with `o200k_base`, and reported output usage accounts for hidden reasoning.

Table 7: General-task results for every model–harness configuration (274 tasks per configuration).

Model	Harness	Score	$R_{0.8}$	R_1	Traj. (K)	Tokens (K)	Requests	Cost ($)
GPT-5.6 Sol	Claude Code	71.60	58.03	22.99	14.5	135.2	11.4	0.3477
	Codex	69.83	57.30	24.82	15.4	288.1	15.0	0.3811
	OpenCode	71.03	57.30	27.01	14.9	172.8	12.1	0.2963
	OpenClaw	**73.08**	56.93	31.02	13.1	149.8	10.7	0.3193
	Hermes	71.18	**58.39**	**32.12**	16.9	170.0	13.9	0.3217
	Goose	71.51	55.47	29.20	13.0	76.2	9.4	0.2746
	Nanobot	66.49	54.38	24.82	26.8	806.2	38.5	0.9483
GPT-5.6 Terra	Claude Code	68.53	51.09	22.26	12.6	81.9	10.7	0.1343
	Codex	66.76	52.92	18.61	5.0	141.8	9.5	0.1199
	OpenCode	67.05	53.28	20.80	11.4	121.8	10.0	0.1675
	OpenClaw	67.88	**54.38**	**26.64**	8.9	66.6	8.4	0.0967
	Hermes	68.42	54.01	25.18	12.0	119.0	11.8	0.1695
	Goose	**69.65**	54.01	23.72	10.7	63.2	8.3	0.1296
	Nanobot	64.18	48.54	21.17	10.1	147.3	10.4	0.1820
GPT-5.6 Luna	Claude Code	**69.54**	**53.65**	24.09	14.9	109.1	12.1	0.0167
	Codex	67.31	52.19	21.53	5.9	145.8	9.7	0.0131
	OpenCode	66.92	52.92	20.44	11.6	144.3	10.8	0.0198
	OpenClaw	67.75	51.82	**26.28**	8.7	71.8	9.1	0.0103
	Hermes	68.00	53.28	23.72	11.6	132.3	12.5	0.0178
	Goose	68.31	53.28	22.63	9.5	68.2	9.2	0.0140
	Nanobot	64.51	48.91	22.26	10.2	163.7	11.2	0.0195
GLM-5.3	Claude Code	68.69	51.46	27.74	18.2	264.0	18.0	0.2081
	Codex	68.58	52.92	22.26	15.1	522.2	20.0	0.5292
	OpenCode	**70.34**	54.01	**35.77**	29.8	504.8	16.4	0.4603
	OpenClaw	67.44	52.55	26.64	36.1	219.2	16.6	0.2063
	Hermes	66.27	51.46	29.20	53.9	387.1	19.8	0.4176
	Goose	67.96	**54.74**	32.12	27.6	559.7	15.9	0.6751
	Nanobot	61.81	49.27	28.47	38.7	1422.0	29.0	1.1655
GLM-5.3-Flash	Claude Code	66.30	52.19	26.64	14.7	100.6	13.0	0.0088
	Codex	64.77	51.46	20.07	13.6	468.0	19.4	0.0461
	OpenCode	69.47	54.38	29.56	18.6	438.1	14.7	0.0526
	OpenClaw	69.00	52.92	28.47	20.5	147.6	14.3	0.0134
	Hermes	67.36	53.28	27.74	22.9	258.7	17.1	0.0289
	Goose	**69.85**	**55.11**	**33.58**	17.2	404.4	13.2	0.0577
	Nanobot	67.18	**55.11**	28.47	21.0	933.2	19.0	0.1054
DeepSeek-V4-Pro	Claude Code	68.67	54.38	25.18	34.3	378.4	17.3	0.1822
	Codex	**69.54**	**58.39**	27.01	27.2	503.3	17.3	0.1222
	OpenCode	68.97	55.84	**29.20**	22.2	248.5	12.5	0.1368
	OpenClaw	64.35	53.28	27.01	36.4	198.3	12.9	0.1792
	Hermes	67.32	55.47	28.83	31.1	314.3	14.5	0.1920
	Goose	66.04	52.19	25.91	27.1	317.1	14.6	0.3120
	Nanobot	64.89	53.28	25.18	30.7	780.5	22.0	0.2479

Continued on next page

General-task results (continued)

Model	Harness	Score	$R_{0.8}$	R_1	Traj. (K)	Tokens (K)	Requests	Cost ($)
DeepSeek-V4-Flash	Claude Code	64.81	51.46	25.18	48.0	562.8	21.3	0.0329
	Codex	**68.34**	52.19	27.74	27.4	447.0	15.7	0.0202
	OpenCode	66.24	50.73	26.28	25.4	371.9	15.9	0.0211
	OpenClaw	66.78	51.46	28.47	38.8	236.5	14.6	0.0199
	Hermes	67.35	**55.47**	**29.93**	30.7	346.9	16.4	0.0232
	Goose	65.91	51.46	28.47	24.5	259.7	13.6	0.0235
	Nanobot	60.92	45.99	23.72	28.5	629.9	19.4	0.0327
Kimi-K3	Claude Code	60.01	48.91	25.18	18.1	199.3	17.6	0.3522
	Codex	65.12	52.55	26.64	17.6	307.4	15.1	0.3537
	OpenCode	**69.49**	**56.20**	**32.85**	21.3	271.9	15.0	0.5583
	OpenClaw	67.27	51.82	28.83	15.0	130.7	11.8	0.3132
	Hermes	68.16	54.01	27.74	20.4	217.9	16.6	0.4596
	Goose	66.95	52.92	29.56	17.2	153.1	11.1	0.5172
	Nanobot	65.18	52.92	31.02	19.7	297.5	13.8	0.5726
Gemma-4-31B	Claude Code	59.04	42.34	24.45	14.9	451.6	32.9	—
	Codex	58.87	**42.70**	**26.64**	13.6	302.3	16.6	—
	OpenCode	59.27	39.78	21.53	13.9	228.7	15.4	—
	OpenClaw	58.78	40.88	23.36	21.7	180.6	16.0	—
	Hermes	**59.63**	38.32	22.99	20.8	224.4	17.4	—
	Goose	52.69	39.78	24.82	11.2	217.9	17.9	—
	Nanobot	52.24	39.05	25.55	11.1	250.1	15.2	—
Qwen3.8-27B	Claude Code	61.67	46.35	28.83	29.8	359.2	15.1	—
	Codex	62.42	48.18	29.56	26.7	403.3	14.9	—
	OpenCode	**65.10**	**51.09**	**33.21**	28.4	391.2	15.8	—
	OpenClaw	58.32	45.62	26.28	27.0	211.0	14.2	—
	Hermes	61.75	47.81	25.18	35.6	271.3	19.8	—
	Goose	25.91	22.63	17.88	13.4	106.8	8.2	—
	Nanobot	34.25	28.10	17.15	19.8	317.7	12.1	—

Table 8: Multimodal results for every model–harness configuration (26 tasks per configuration).

Model	Harness	Score	$R_{0.8}$	R_1	Traj. (K)	Tokens (K)	Requests	Cost ($)
GPT-5.6 Sol	Claude Code	68.38	11.54	11.54	16.4	100.7	14.1	0.2882
	Codex	**68.81**	**23.08**	**19.23**	12.1	307.9	15.2	0.4247
	OpenCode	60.31	19.23	**19.23**	21.7	305.7	17.4	0.4897
	OpenClaw	60.73	11.54	11.54	14.3	154.3	16.3	0.3656
	Hermes	66.31	15.38	15.38	24.4	749.1	32.3	0.8219
	Goose	54.50	3.85	3.85	3.8	81.3	8.7	0.2573
	Nanobot	65.65	0.00	0.00	17.3	427.3	21.7	0.5941
GPT-5.6 Terra	Claude Code	58.00	3.85	3.85	9.9	58.4	10.3	0.0982
	Codex	**70.46**	15.38	15.38	13.6	364.2	17.4	0.2667
	OpenCode	56.65	**23.08**	**23.08**	12.2	126.3	10.3	0.1401
	OpenClaw	50.31	**23.08**	**23.08**	12.1	144.0	11.4	0.1815
	Hermes	60.73	7.69	7.69	18.2	549.2	27.7	0.5755
	Goose	57.54	19.23	19.23	3.9	90.6	10.0	0.1267
	Nanobot	67.54	19.23	19.23	10.0	232.3	15.7	0.2021

Continued on next page

Multimodal results (continued)

Model	Harness	Score	$R_{0.8}$	R_1	Traj. (K)	Tokens (K)	Requests	Cost ($)
GPT-5.6 Luna	Claude Code	58.65	23.08	19.23	11.7	70.0	11.7	0.0118
	Codex	62.88	11.54	7.69	14.8	334.4	15.8	0.0282
	OpenCode	36.54	7.69	3.85	12.5	123.8	10.0	0.0144
	OpenClaw	58.69	23.08	**23.08**	15.2	191.2	13.8	0.0222
	Hermes	58.96	7.69	7.69	26.0	798.4	37.3	0.0825
	Goose	51.19	19.23	15.38	3.2	78.3	9.8	0.0117
	Nanobot	**64.38**	**26.92**	19.23	7.0	208.4	14.4	0.0211
GLM-5.3-Flash	Claude Code	52.92	23.08	19.23	50.6	1513.7	45.2	0.1644
	Codex	71.19	26.92	23.08	29.7	1172.3	37.8	0.0884
	OpenCode	57.69	26.92	26.92	48.7	1671.2	39.2	0.1781
	OpenClaw	**76.69**	**50.00**	**50.00**	34.3	397.5	33.3	0.0311
	Hermes	64.65	42.31	42.31	69.3	1783.6	45.6	0.1841
	Goose	65.08	30.77	30.77	32.7	1703.4	28.9	0.2366
	Nanobot	57.46	30.77	26.92	33.3	4402.4	67.3	0.4843
Kimi-K3	Claude Code	60.92	26.92	23.08	37.9	866.8	41.5	1.4662
	Codex	28.08	0.00	0.00	33.0	1228.6	36.4	2.2851
	OpenCode	64.35	26.92	26.92	56.9	1917.1	51.4	2.7277
	OpenClaw	**72.73**	**30.77**	**30.77**	21.8	472.4	26.9	0.8926
	Hermes	57.92	23.08	23.08	57.3	2682.0	77.1	2.9275
	Goose	68.73	**30.77**	**30.77**	46.0	1120.0	34.0	2.9737
	Nanobot	53.04	26.92	26.92	44.9	1956.1	50.3	3.0120
Gemma-4-31B	Claude Code	**45.12**	**15.38**	**0.00**	35.7	576.3	33.3	—
	Codex	17.77	0.00	**0.00**	10.6	297.6	19.7	—
	OpenCode	28.15	3.85	**0.00**	24.1	865.0	30.6	—
	OpenClaw	29.19	7.69	**0.00**	34.8	602.0	32.0	—
	Hermes	23.69	3.85	**0.00**	37.5	977.6	42.3	—
	Goose	39.50	7.69	**0.00**	19.4	354.1	29.4	—
	Nanobot	32.31	0.00	**0.00**	17.6	490.0	28.3	—
Qwen3.8-27B	Claude Code	**69.04**	30.77	**30.77**	87.5	1400.3	48.8	—
	Codex	47.38	**34.62**	**30.77**	122.9	912.8	31.7	—
	OpenCode	45.92	23.08	23.08	110.5	1684.7	48.7	—
	OpenClaw	55.88	30.77	**30.77**	94.3	1104.0	41.7	—
	Hermes	50.23	23.08	23.08	108.9	1544.2	48.3	—
	Goose	50.88	15.38	15.38	83.3	1078.7	36.9	—
	Nanobot	31.58	3.85	3.85	104.2	3317.3	87.7	—

C Analysis Protocol and Detailed Results

The following subsections supplement the model-dependence, task-requirement, and cost analyses in Section 5. General-task comparisons use all 274 tasks for each of the ten models and seven harnesses; paired contrasts hold the task and model fixed. API costs follow Appendix B. Unless otherwise specified, we use 2,000 task-bootstrap resamples (seed 20260923), retaining all configuration outcomes for each selected task, and report percentile 95% intervals.

C.1 Harness Dependence across Models

Sensitivity across harnesses. For each model, we compute the population SD of its seven configuration-level mean scores (divisor seven). This measures variation across harnesses, not variation across tasks or repeated runs. Correlations treat models as observations. Table 9 reports the mean-score SD and the range of full-credit rates. Across models, Spearman correlations between mean oracle score and the cross-harness SD of $R_{0.8}$ and R_1 are -0.73 and $+0.16$, respectively.

The corresponding mean-score–SD correlation on the separately evaluated seven-model multimodal suite is -0.57.

Model	Mean score	Cross-harness SD	R_1 range (pp)
GPT Sol	70.67	1.93	9.12
GLM Flash	67.70	1.71	13.50
GPT Terra	67.49	1.62	8.03
GPT Luna	67.48	1.44	5.84
GLM-5.3	67.30	2.52	13.50
DS Pro	67.11	1.91	4.01
Kimi K3	66.03	2.85	7.66
DS Flash	65.76	2.23	6.20
Gemma 4	57.22	3.02	5.11
Qwen3.8	52.77	14.64	16.06

Table 9: Harness sensitivity on the general task suite. Mean score averages all seven harnesses; SD is computed across their mean scores. The R_1 range is the best-minus-worst full-credit rate across the same harnesses.

Model–harness interactions. Interaction residuals subtract the row and column means and add back the grand mean on the full general task suite. Their values depend on the set of evaluated models: excluding Qwen3.8 changes Kimi-K3–Claude Code's residual from -6.95 to -6.06 points, while GLM-5.3-Flash–Goose changes from $+4.62$ to $+1.92$ points. Residuals locate departures from the specified additive baseline; execution traces are needed to explain them.

Model-specific harness selection. One policy chooses a single harness for all models; the other chooses one per model. Both use training tasks only and are evaluated on held-out tasks. Tasks that share a grading template are kept together so that the same template does not appear on both sides of the split. We use two template definitions: identical scoring rubrics, and matching module weights, requirement-role metadata, and gates. Table 10 averages 20 repetitions of grouped five-fold evaluation. Full-credit gains are positive in all 20 repetitions under both definitions.

Held-out groups	Mean score	$R_{0.8}$	R_1
Exact rubric	+0.14	+0.25	+2.73
Scoring structure	-0.14	+0.31	+2.53

Table 10: Held-out gains from selecting a harness per model rather than one globally. Each policy optimizes the reported metric on training tasks only. Rows average 20 repetitions of grouped five-fold evaluation. Rate gains are percentage points.

C.2 Task Requirements and Complementarity

Category-level delivery. Table 11 uses the current six-category taxonomy. Multimodal tasks require multimodal model capabilities during task execution and are evaluated separately.

Category	Tasks	Score	$R_{0.8}$	R_1
Business Operations	71	68.31	55.19	21.71
Content Production	16	69.55	58.04	41.07
Research and Analysis	40	64.55	53.00	25.46
Software Engineering	81	57.34	39.37	22.75
Systems Engineering	65	69.87	56.66	32.66
Web Development	1	67.36	82.86	1.43

Table 11: Task-category outcomes on the full general suite. Each task contributes 70 attempts.

Complementary full-credit outcomes. With DeepSeek Pro, Codex and OpenCode differ by only 0.57 mean-score points, yet their full-credit sets overlap on 49 tasks, with 25 completed only by Codex and 31 only by OpenCode; neither receives full credit on the remaining 169. Their union

exceeds the better fixed configuration by 9.12 percentage points, an observed upper bound for task-level selection. Among the 56 discordant tasks, only 21 retain the same winner when score differences are averaged across the other nine models. Complementary task outcomes therefore depend on the model.

Appendices D and E provide the annotation protocol and paired execution evidence; oracle scores remain artifact-based.

Delivery by execution requirement. Task specifications carry overlapping execution-demand tags, assigned independently of the observed scores. We group the 274 general tasks by these tags and retain all ten models and seven harnesses on every task in each group. Tables 12 and 13 report equally weighted task and model averages. Tags identify intended demands, not isolated interventions; an untagged task may still involve the same behavior. All tasks carry a multi-step execution tag, which is omitted because it reproduces the full-suite comparison.

Table 12: Mean oracle scores by task-design demand. N counts tasks, each evaluated under all 70 general configurations. Best harness average within each demand is bold. Groups overlap.

Demand	N	Claude Code	Codex	OpenCode	OpenClaw	Hermes	Goose	Nanobot
Evidence grounding	198	67.8	68.2	**69.2**	68.2	68.4	65.0	62.5
State management	126	71.7	72.4	**73.1**	71.7	71.8	67.9	65.1
Tool selection	114	68.7	69.0	**70.3**	67.8	69.3	63.4	60.6
Error recovery	71	80.9	81.1	**82.9**	81.0	81.3	77.4	72.0
Verification	236	65.4	65.8	**67.1**	65.5	65.8	61.8	59.4
Output contract	87	57.2	57.3	58.6	**59.1**	58.1	55.7	54.3
Safety boundary	166	69.6	69.6	**70.6**	69.9	70.2	66.7	63.5
Context management	12	77.7	**80.2**	79.8	75.9	79.1	73.8	67.6

Table 13: Full-credit attainment by task-design demand. N counts tasks, each evaluated under all 70 general configurations. Best harness average within each demand is bold. Groups overlap.

Demand	N	Claude Code	Codex	OpenCode	OpenClaw	Hermes	Goose	Nanobot
Evidence grounding	198	26.6	26.0	**29.1**	28.3	27.6	28.4	26.2
State management	126	23.0	22.8	**25.7**	24.0	23.8	24.8	22.8
Tool selection	114	24.4	23.3	26.5	26.1	**27.5**	25.3	23.7
Error recovery	71	32.5	31.8	35.2	35.1	33.4	**36.1**	32.0
Verification	236	25.7	25.0	**28.1**	27.5	27.3	26.8	25.0
Output contract	87	25.9	24.6	**28.3**	27.6	26.2	25.5	24.3
Safety boundary	166	26.2	25.4	28.4	**28.7**	27.0	28.2	25.3
Context management	12	19.2	21.7	**22.5**	**22.5**	19.2	20.0	16.7

The same demand does not select the same harness for every model. Across the 71 recovery-tagged tasks, Goose has the highest model-averaged full-credit rate (36.06%), but the paired OpenCode–Goose contrast favors OpenCode for Gemma and Qwen and Goose for the other eight models (Table 14). For Gemma, OpenCode's full-credit advantage is 8.45 percentage points on recovery-tagged tasks and reverses to -7.39 on the remaining 203 tasks. Its mean-score advantage also grows from 3.28 to 16.04 points. Harness preferences vary across both models and requirement-defined task groups.

Checking work-category composition. For the Gemma contrast, we also compare the two task groups within each work category, then average these differences using category task counts as weights. Five categories contain both groups: Business Operations (34 recovery / 37 remaining), Content Production (4/12), Research and Analysis (9/31), Software Engineering (5/76), and Systems Engineering (18/47). Across these 273 tasks, the category-standardized difference between the two groups remains 10.80 score points and 24.87 percentage points in R_1. Thus, differing work-category proportions alone do not account for the observed contrast. Demand tags remain overlapping in this analysis.

Execution mechanisms recur across distinct tasks. Evidence-and-requirement, state-and-artifact, and verification errors each span more than 100 tasks and all ten models and seven harnesses

Table 14: Paired OpenCode minus Goose differences within each model. Recovery: 71 recovery-tagged tasks; remaining: 203 other tasks. Score differences are points; R_1 differences are percentage points. Positive values favor OpenCode.

	Mean score difference		R_1 difference	
Model	Recovery	Remaining	Recovery	Remaining
GPT Sol	+1.06	-1.03	-4.23	-1.48
GPT Terra	-0.90	-3.18	-2.82	-2.96
GPT Luna	-2.48	-1.01	-4.23	-1.48
GLM-5.3	+0.56	+3.02	-2.82	+5.91
GLM Flash	-3.75	+0.80	-9.86	-1.97
DS Pro	+2.28	+3.16	-1.41	+4.93
DS Flash	-1.10	+0.84	-4.23	-1.48
Kimi K3	-0.65	+3.66	-4.23	+5.91
Gemma 4	+16.04	+3.28	+8.45	-7.39
Qwen3.8	+43.45	+37.69	+16.90	+14.78

(Table 15). These annotations describe observed behavior, complementing the task-design demand tags.

Table 15: Coverage of observed mechanisms in the unchanged base samples. Runs and tasks are distinct within each category; categories overlap. Models/harnesses gives the number of represented backends and harnesses. Full counts labeled runs that nevertheless receive full credit.

Suite	Mechanism	Runs	Tasks	Models/harnesses	Full
General	Evidence and requirements	596	110	10/7	68
	Tool invocation	694	193	10/7	197
	Recovery	262	122	10/7	67
	Execution stagnation	55	36	9/7	0
	State and artifacts	399	143	10/7	75
	Verification	948	160	10/7	24
Multimodal	Evidence and requirements	66	13	7/7	5
	Tool invocation	34	14	7/7	11
	Recovery	27	14	7/5	3
	Execution stagnation	9	5	3/7	0
	State and artifacts	21	15	6/7	0
	Verification	217	25	7/7	2

Harness-level annotation profiles. Table 16 aggregates the base samples with equal model coverage: 800 general runs and 84 multimodal runs per harness. The profiles distinguish tool-invocation, recovery, and state-and-artifact problems across harnesses. Frequencies describe the score-stratified samples rather than benchmark-wide incidence; configuration-level breakdowns remain in Appendix D.1.

C.3 Delivery Quality and Execution Cost

Quality–cost contrasts. Table 17 reports paired quality and cost contrasts for DeepSeek Flash and GPT Sol on all 274 general tasks. We use 3,000 task-paired bootstrap resamples (seed 20260925) to obtain percentile 95% intervals. The DeepSeek Flash–Nanobot cache fraction is total reported cached input divided by total input across the same suite.

Figure 5(b) uses the same full suite for GPT Luna. Empirical frontier membership is computed separately for mean score, $R_{0.8}$, and R_1, with identical task weights and cost estimates in every panel.

Requests and delivery. Let T_i be the recorded number of provider transactions, Q_i the final oracle score, N the number of runs, and b a request count on the horizontal axis. We plot $S(b) = N^{-1}\sum_i \mathbf{1}[T_i > b]$ and $A_q(b) = N^{-1}\sum_i \mathbf{1}[T_i \le b, Q_i \ge q]$, for $q = 0.80$ and $q = 1$. Provider transactions can include auxiliary, delegated, or failed requests; one transaction is not nec-

Table 16: Runs with each mechanism label, by harness and base sample. E: evidence and requirements; T: tool invocation; R: recovery; P: execution stagnation; S: state and artifacts; V: verification. Labels overlap.

Suite	Harness	n	E	T	R	P	S	V
General	Claude Code	800	70	95	14	14	34	127
	Codex	800	71	79	20	5	34	120
	OpenCode	800	89	73	51	5	35	143
	OpenClaw	800	80	90	19	3	32	161
	Hermes	800	100	81	38	8	41	130
	Goose	800	78	171	14	6	26	131
	Nanobot	800	108	105	106	14	197	136
Multimodal	Claude Code	84	7	8	3	1	1	36
	Codex	84	5	3	0	1	3	31
	OpenCode	84	13	6	1	2	3	28
	OpenClaw	84	11	5	0	1	1	33
	Hermes	84	10	5	9	2	3	25
	Goose	84	10	2	1	1	2	30
	Nanobot	84	10	5	13	1	8	34

Comparison (first minus second)	Score difference	R_1 difference	Cost ratio
DS Flash: Codex / Nano	+7.41 [+4.06, +10.84]	+4.01 [-0.36, +8.39]	0.62 [0.56, 0.68]
GPT Sol: Nano / OCl	-6.59 [-10.18, -3.03]	-6.20 [-10.95, -1.09]	2.97 [2.61, 3.38]

Table 17: Paired quality and cost contrasts on all 274 general tasks, with 95% intervals. Rate differences are percentage points; costs are first/second ratios.

essarily one tool operation. Curves use final scores at observed termination, not counterfactual scores under a smaller budget. Provider request counts are distinct from interaction-turn limits.

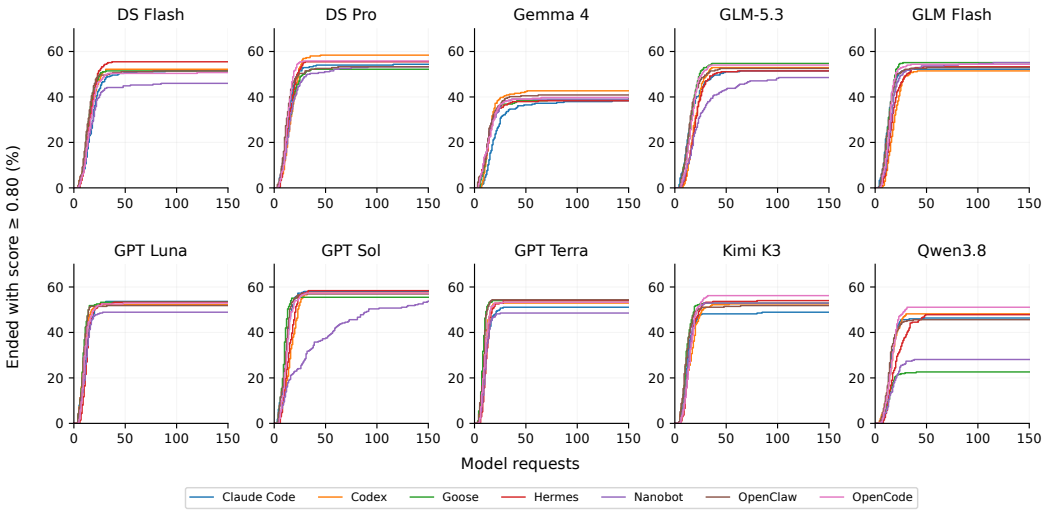


Figure 7: Observed termination with score at least 0.80 for all ten models and seven harnesses, on the full general task suite.

Table 18 reports paired quality and request-count differences on all 274 general tasks.

Request distributions. Table 19 complements the per-configuration means and quality metrics in Appendix B.3 with median requests and counts of runs exceeding 40 requests.

Comparison	N	Score difference	R_1 difference	Request difference
Sol	274	+6.59 [+3.20, +10.23]	+6.20 [+1.46, +10.95]	-27.73 [-31.83, -23.65]
GLM-5.3	274	+8.53 [+4.61, +12.39]	+7.30 [+2.92, +12.04]	-12.62 [-15.92, -9.63]

Table 18: Sol compares OpenClaw minus Nanobot; GLM-5.3 compares OpenCode minus Nanobot. Paired differences with percentile 95% task-bootstrap intervals. R_1 differences are percentage points. Both comparisons use all 274 general tasks.

Table 19: Request distributions on 274 general tasks per configuration. Each cell gives the median request count / number of runs with more than 40 requests. Requests include auxiliary and delegated calls.

Model	Claude Code	Codex	OpenCode	OpenClaw	Hermes	Goose	Nanobot
GPT Sol	11 / 0	15 / 0	12 / 0	11 / 0	14 / 0	9 / 0	26 / 85
GPT Terra	10 / 0	9 / 0	10 / 1	9 / 0	11 / 0	8 / 0	10 / 0
GPT Luna	11 / 0	9 / 0	10 / 1	9 / 0	12 / 0	9 / 0	11 / 0
GLM-5.3	15 / 16	18 / 5	16 / 0	15 / 7	19 / 8	15 / 2	20 / 54
GLM Flash	12 / 2	17 / 9	14 / 0	13 / 4	15 / 3	12 / 0	15 / 18
DS Pro	14 / 13	17 / 3	12 / 0	11 / 7	13 / 2	14 / 1	15 / 43
DS Flash	17 / 23	15 / 1	15 / 3	13 / 4	15 / 1	13 / 1	16 / 20
Kimi K3	12 / 22	14 / 4	15 / 0	10 / 3	14 / 8	10 / 0	13 / 1
Gemma 4	19 / 44	14 / 8	14 / 7	14 / 9	14 / 11	12 / 19	13 / 9
Qwen3.8	14 / 5	15 / 0	15 / 1	13 / 3	18 / 12	5 / 1	10 / 3

D TRAJECTORY ANNOTATION

Sampling. The corpus contains 6,864 unique runs covering all 300 tasks. The general base sample contains 80 runs per model–harness configuration: 20 full-credit and 60 non-full-credit runs. We initially allocate 20 places to each band: zero, $(0, 0.80)$, and $[0.80, 1)$. An undersized band is exhausted, and remaining places go to the non-full-credit band with the most remaining capacity. Within each band, task-type strata receive at least two runs when available; remaining places are allocated proportionally to residual stratum sizes using largest remainders. Selection strata and inclusion probabilities are retained; task coverage is summarized with the current taxonomy.

Multimodal sampling selects 12 runs per configuration. Each nonempty score band receives up to three initial places; remaining places are allocated proportionally to available capacity. Within a band, each represented task subcategory receives at least one run when available, with the remainder allocated proportionally. Fixed-seed hash rankings select runs within strata. The resulting general score-band counts are 1,321 zero, 1,485 positive but below 0.80, 1,394 high-partial, and 1,400 full-credit; the multimodal counts are 123, 312, 21, and 132, respectively.

Matched successful runs add same-task, same-model contrasts under a different harness; diagnostic records ensure task and stratum coverage (Table 20). Each record retains its role, stratum, and inclusion probability. The main distribution figure and configuration-level error percentages use only the base samples. Matched controls and diagnostic additions support case analysis and are not pooled into those denominators.

Sample component	Selected	Structured records
General probability sample	5,600	5,600
Multimodal probability sample	588	588
Matched successful runs	620	620
Task and stratum coverage additions	56	56
Total	6,864	6,864

Table 20: Trajectory annotation coverage by sampling role. Structured records pass format and evidence-citation checks.

Evidence-linked annotation. Agent-assisted inspection of execution trajectories identifies six recurring error categories. GPT-5.6 Luna then classifies sampled trajectories using evidence packets

containing task requirements, execution events, tool feedback, and artifact excerpts. Each record states the task requirement and cites at least two specific observations with event locators and verbatim evidence. Positive labels additionally describe the behavior, recovery outcome, relation to delivery, confidence, and a possible harness design response. Multiple labels are allowed when distinct behaviors are supported. Automated checks validate record structure and cited text; a separate targeted semantic review checks whether the cited evidence supports the interpretation. The six categories in Table 21 describe execution mechanisms rather than score levels.

Human validation and adjudication. Two human annotators independently reviewed the same 100 trajectories randomly sampled from the base samples. They saw only the execution trajectories, artifacts, and final scores; model-generated error labels were withheld. Before adjudication, complete error-label-set agreement was 97% between the human annotators and 98% and 97% between each annotator and GPT-5.6 Luna, respectively. All three agreed on 96 trajectories. On one trajectory, both humans agreed on a label set different from Luna's; on the remaining three, the humans disagreed and Luna matched one of them. This agreement supports Luna-assisted classification under the annotation protocol. Disagreements were adjudicated and labels corrected. All uncertain model classifications receive manual review and annotation. Other supported errors fall outside the six-category codebook; records without sufficient evidence for a supported classification remain unclassified.

Mechanism	Evidence required	Design direction
Evidence and requirements	An applicable rule or acquired fact conflicts with the subsequent decision.	Preserve constraints and decision-relevant facts.
Tool invocation	A call uses an incompatible tool, argument, or target, with observable feedback.	Check interfaces and target types before dispatch.
Recovery	Actionable feedback is followed by an unchanged invalid operation or a repair that introduces another defect.	Carry the required correction into the next action.
Execution stagnation	Repeated settled decisions or investigation produce no corresponding action or new evidence.	Materialize a checkable intermediate result.
State and artifacts	Recorded operations and final artifacts reveal missing persistence, inconsistent outputs, or duplicate effects.	Track updates and reconcile final outputs.
Verification	A specific check or completion claim conflicts with visible evidence or omits an explicit delivery requirement.	Tie verification to the delivery contract.

Table 21: Execution mechanisms, the evidence needed to identify them, and the harness design questions they motivate. Categories can overlap within a run.

Recovery and delivery. For each observed error, recovery is recorded as complete, partial, absent, or unknown. Its relation to delivery is recorded separately as direct, contributing, not established, or unknown. An invalid call followed by a corrected call can therefore be a recovered tool error without being a recovery failure. Conversely, recognizing a mistake in reasoning does not establish recovery unless subsequent actions address it. A high final score can coexist with earlier corrected errors. The absence of a step from a bounded excerpt does not establish that the step never occurred, and a hidden-oracle failure alone does not establish that the agent ignored visible feedback.

Final annotation corpus. The completed corpus contains 6,864 accepted records: 6,235 general and 629 multimodal. Of these, 3,079 carry at least one confirmed error label, 2,318 contain no confirmed error, and 1,467 remain unclassified because the inspected evidence does not support a determination. Counts are reported separately by sampling role and modality. Unconfirmed errors may be absent or unobservable in the inspected evidence; mechanisms outside the codebook are tracked separately.

	General: 5,600 runs			Multimodal: 588 runs		
Mechanism	Runs	Labels	Recovered	Runs	Labels	Recovered
Evidence / requirements	596	614	94	66	69	6
Tool invocation	694	729	655	34	37	29
Recovery	262	263	160	27	27	8
Execution stagnation	55	55	2	9	9	0
State / artifacts	399	401	26	21	21	0
Verification	948	962	7	217	220	0

Table 22: Final descriptive counts for the two probability samples. *Runs* reports distinct annotated runs per category; *Labels* reports observations; *Recovered* reports labels followed by complete recovery. Categories can overlap. Matched successful controls and diagnostic additions are excluded. The score-stratified sample counts are not weighted estimates of benchmark-wide prevalence.

D.1 Error Distributions for Every Model–Harness Configuration

Tables 23 and 24 report all 119 configurations. N_{all} counts every annotated run for the configuration, including matched controls and diagnostics; n is its base-sample size and the denominator for all percentages. E: evidence and requirements; T: tool invocation; R: recovery; P: execution stagnation; S: state and artifacts; V: verification; O: other confirmed categories outside the six-category codebook. Each category counts distinct runs, so categories can overlap.

None means no error was confirmed in the inspected evidence; *Uncls.* means the inspected evidence is insufficient for a supported classification. Neither is inferred from the oracle score. *Full* is the percentage of base-sample runs receiving full oracle credit, which can include recovered errors. The percentage with any confirmed error is $100-$*None*$-$*Uncls.*, up to rounding. All percentages describe the score-stratified samples, not weighted population prevalence. Total annotation coverage is 6,235 general and 629 multimodal runs; their base samples contain 5,600 and 588 runs.

Table 23: General-task annotation profiles. Counts: N_{all}, n; all remaining columns: percentages of n.

Model	Harness	N_{all}	n	E	T	R	P	S	V	O	None	Uncls.	Full
DS Flash	Claude Code	85	80	2.5	7.5	1.2	0.0	6.2	16.2	5.0	40.0	25.0	25.0
	Codex	86	80	6.2	13.8	2.5	0.0	6.2	11.2	7.5	31.2	27.5	25.0
	OpenCode	90	80	10.0	8.8	5.0	1.2	3.8	20.0	6.2	32.5	21.2	25.0
	OpenClaw	92	80	8.8	7.5	1.2	0.0	3.8	22.5	5.0	40.0	20.0	25.0
	Hermes	95	80	7.5	2.5	2.5	1.2	5.0	17.5	6.2	35.0	26.2	25.0
	Goose	90	80	10.0	28.8	3.8	0.0	3.8	20.0	2.5	31.2	15.0	25.0
	Nanobot	90	80	5.0	11.2	18.8	1.2	28.8	18.8	2.5	21.2	16.2	25.0
DS Pro	Claude Code	88	80	12.5	7.5	0.0	0.0	1.2	21.2	3.8	35.0	25.0	25.0
	Codex	93	80	11.2	7.5	1.2	0.0	2.5	17.5	0.0	42.5	21.2	25.0
	OpenCode	97	80	8.8	6.2	3.8	1.2	3.8	17.5	0.0	41.2	22.5	25.0
	OpenClaw	90	80	8.8	5.0	0.0	0.0	2.5	20.0	0.0	35.0	31.2	25.0
	Hermes	92	80	8.8	5.0	2.5	1.2	2.5	13.8	1.2	40.0	26.2	25.0
	Goose	88	80	11.2	41.2	2.5	1.2	0.0	17.5	1.2	20.0	16.2	25.0
	Nanobot	89	80	7.5	8.8	10.0	3.8	28.8	18.8	1.2	23.8	17.5	25.0
Gemma 4	Claude Code	93	80	11.2	15.0	10.0	5.0	6.2	20.0	3.8	28.8	15.0	25.0
	Codex	91	80	15.0	8.8	10.0	1.2	6.2	22.5	1.2	32.5	18.8	25.0
	OpenCode	89	80	17.5	26.2	13.8	1.2	6.2	16.2	1.2	18.8	12.5	25.0
	OpenClaw	92	80	11.2	21.2	11.2	2.5	3.8	23.8	3.8	25.0	13.8	25.0
	Hermes	88	80	23.8	22.5	17.5	1.2	10.0	20.0	2.5	18.8	10.0	25.0
	Goose	92	80	13.8	11.2	6.2	1.2	8.8	17.5	1.2	32.5	17.5	25.0
	Nanobot	93	80	17.5	17.5	11.2	1.2	25.0	15.0	2.5	17.5	21.2	25.0

Continued on next page

General-task annotation profiles (continued)

Model	Harness	N_{all}	n	E	T	R	P	S	V	O	None	Uncls.	Full
GLM-5.3	Claude Code	87	80	6.2	11.2	1.2	0.0	2.5	25.0	1.2	31.2	28.8	25.0
	Codex	86	80	6.2	25.0	1.2	1.2	1.2	13.8	2.5	32.5	22.5	25.0
	OpenCode	88	80	12.5	1.2	11.2	0.0	2.5	25.0	0.0	32.5	23.8	25.0
	OpenClaw	88	80	13.8	13.8	1.2	0.0	2.5	23.8	3.8	31.2	22.5	25.0
	Hermes	90	80	12.5	11.2	3.8	1.2	2.5	23.8	0.0	30.0	23.8	25.0
	Goose	91	80	7.5	15.0	0.0	0.0	1.2	15.0	5.0	35.0	28.8	25.0
	Nanobot	87	80	18.8	12.5	16.2	1.2	17.5	25.0	0.0	16.2	16.2	25.0
GLM Flash	Claude Code	90	80	10.0	7.5	0.0	2.5	6.2	15.0	6.2	36.2	22.5	25.0
	Codex	86	80	10.0	11.2	3.8	0.0	10.0	21.2	7.5	30.0	18.8	25.0
	OpenCode	93	80	13.8	7.5	6.2	0.0	10.0	25.0	3.8	36.2	10.0	25.0
	OpenClaw	90	80	12.5	10.0	0.0	0.0	3.8	26.2	2.5	35.0	20.0	25.0
	Hermes	90	80	8.8	22.5	2.5	1.2	3.8	20.0	3.8	23.8	23.8	25.0
	Goose	92	80	7.5	20.0	1.2	1.2	0.0	17.5	3.8	33.8	20.0	25.0
	Nanobot	88	80	16.2	10.0	15.0	0.0	17.5	20.0	1.2	20.0	12.5	25.0
GPT Luna	Claude Code	85	80	10.0	17.5	0.0	0.0	3.8	10.0	2.5	33.8	27.5	25.0
	Codex	85	80	8.8	0.0	0.0	0.0	1.2	10.0	0.0	42.5	40.0	25.0
	OpenCode	86	80	12.5	15.0	1.2	0.0	5.0	13.8	2.5	30.0	27.5	25.0
	OpenClaw	88	80	8.8	18.8	2.5	0.0	7.5	17.5	5.0	28.8	25.0	25.0
	Hermes	86	80	16.2	10.0	3.8	1.2	10.0	10.0	2.5	33.8	22.5	25.0
	Goose	83	80	12.5	16.2	1.2	0.0	6.2	15.0	2.5	32.5	23.8	25.0
	Nanobot	87	80	13.8	23.8	3.8	1.2	23.8	17.5	0.0	15.0	21.2	25.0
GPT Sol	Claude Code	87	80	6.2	16.2	1.2	0.0	1.2	15.0	1.2	33.8	28.8	25.0
	Codex	84	80	6.2	8.8	1.2	0.0	1.2	16.2	0.0	37.5	32.5	25.0
	OpenCode	93	80	7.5	5.0	3.8	0.0	1.2	12.5	0.0	37.5	36.2	25.0
	OpenClaw	91	80	10.0	10.0	1.2	0.0	2.5	17.5	2.5	27.5	31.2	25.0
	Hermes	95	80	8.8	7.5	2.5	0.0	1.2	13.8	0.0	35.0	35.0	25.0
	Goose	89	80	5.0	16.2	1.2	0.0	0.0	17.5	5.0	37.5	23.8	25.0
	Nanobot	87	80	10.0	12.5	16.2	0.0	26.2	11.2	0.0	23.8	25.0	25.0
GPT Terra	Claude Code	85	80	7.5	20.0	2.5	0.0	6.2	15.0	8.8	27.5	23.8	25.0
	Codex	84	80	5.0	0.0	0.0	0.0	2.5	12.5	0.0	41.2	40.0	25.0
	OpenCode	89	80	7.5	11.2	5.0	0.0	2.5	13.8	0.0	35.0	30.0	25.0
	OpenClaw	91	80	11.2	13.8	0.0	0.0	2.5	15.0	5.0	33.8	25.0	25.0
	Hermes	89	80	15.0	6.2	2.5	1.2	6.2	10.0	1.2	38.8	25.0	25.0
	Goose	90	80	8.8	7.5	1.2	0.0	6.2	15.0	1.2	35.0	26.2	25.0
	Nanobot	85	80	15.0	18.8	7.5	1.2	21.2	10.0	2.5	23.8	16.2	25.0
Kimi K3	Claude Code	83	80	8.8	6.2	0.0	8.8	7.5	8.8	3.8	35.0	23.8	25.0
	Codex	88	80	8.8	15.0	1.2	2.5	3.8	15.0	1.2	28.8	30.0	25.0
	OpenCode	90	80	10.0	5.0	7.5	0.0	5.0	18.8	1.2	32.5	26.2	25.0
	OpenClaw	91	80	11.2	5.0	0.0	0.0	3.8	20.0	2.5	35.0	25.0	25.0
	Hermes	89	80	13.8	7.5	6.2	1.2	5.0	22.5	0.0	33.8	23.8	25.0
	Goose	87	80	13.8	7.5	0.0	0.0	2.5	17.5	3.8	41.2	22.5	25.0
	Nanobot	90	80	17.5	7.5	15.0	3.8	28.8	15.0	0.0	21.2	15.0	25.0
Qwen3.8	Claude Code	95	80	12.5	10.0	1.2	1.2	1.2	12.5	6.2	32.5	27.5	25.0
	Codex	93	80	11.2	8.8	3.8	1.2	7.5	10.0	2.5	32.5	32.5	25.0
	OpenCode	94	80	11.2	5.0	6.2	2.5	3.8	16.2	3.8	33.8	23.8	25.0
	OpenClaw	90	80	3.8	7.5	6.2	1.2	7.5	15.0	1.2	28.8	32.5	25.0
	Hermes	92	80	10.0	6.2	3.8	0.0	5.0	11.2	2.5	36.2	28.8	25.0
	Goose	84	80	7.5	50.0	0.0	3.8	3.8	11.2	2.5	18.8	11.2	25.0
	Nanobot	81	80	13.8	8.8	18.8	3.8	28.8	18.8	1.2	20.0	17.5	25.0

Table 24: Multimodal-task annotation profiles. Counts: N_{all}, n; all remaining columns: percentages of n.

Model	Harness	N_{all}	n	E	T	R	P	S	V	O	None	Uncls.	Full
Gemma 4	Claude Code	12	12	0.0	33.3	25.0	0.0	0.0	83.3	16.7	0.0	8.3	0.0
	Codex	12	12	8.3	8.3	0.0	0.0	16.7	25.0	8.3	8.3	33.3	0.0
	OpenCode	12	12	25.0	25.0	0.0	0.0	8.3	50.0	0.0	0.0	16.7	0.0
	OpenClaw	12	12	25.0	8.3	0.0	0.0	8.3	58.3	16.7	8.3	16.7	0.0
	Hermes	12	12	8.3	16.7	8.3	0.0	16.7	41.7	8.3	8.3	16.7	0.0
	Goose	12	12	33.3	8.3	0.0	0.0	16.7	58.3	0.0	0.0	16.7	0.0
	Nanobot	12	12	8.3	16.7	8.3	0.0	8.3	75.0	8.3	0.0	16.7	0.0
GLM Flash	Claude Code	14	12	16.7	0.0	0.0	0.0	0.0	33.3	0.0	16.7	33.3	25.0
	Codex	12	12	8.3	16.7	0.0	0.0	0.0	16.7	16.7	41.7	16.7	33.3
	OpenCode	13	12	16.7	8.3	8.3	0.0	0.0	25.0	0.0	25.0	16.7	33.3
	OpenClaw	16	12	0.0	0.0	0.0	0.0	0.0	33.3	16.7	33.3	16.7	50.0
	Hermes	14	12	8.3	0.0	0.0	8.3	0.0	16.7	0.0	25.0	41.7	41.7
	Goose	15	12	8.3	0.0	0.0	8.3	0.0	41.7	8.3	16.7	16.7	33.3
	Nanobot	13	12	25.0	8.3	25.0	0.0	0.0	8.3	8.3	25.0	16.7	33.3
GPT Luna	Claude Code	14	12	16.7	0.0	0.0	0.0	0.0	50.0	0.0	25.0	8.3	25.0
	Codex	12	12	0.0	0.0	0.0	0.0	0.0	50.0	0.0	25.0	25.0	16.7
	OpenCode	12	12	25.0	0.0	0.0	0.0	8.3	41.7	0.0	8.3	33.3	8.3
	OpenClaw	13	12	16.7	0.0	0.0	0.0	0.0	25.0	0.0	41.7	16.7	33.3
	Hermes	12	12	16.7	8.3	25.0	0.0	0.0	16.7	0.0	16.7	25.0	16.7
	Goose	13	12	8.3	0.0	8.3	0.0	0.0	25.0	0.0	16.7	41.7	25.0
	Nanobot	12	12	0.0	8.3	0.0	0.0	0.0	41.7	0.0	41.7	8.3	25.0
GPT Sol	Claude Code	12	12	0.0	8.3	0.0	0.0	0.0	50.0	8.3	16.7	25.0	25.0
	Codex	14	12	0.0	0.0	0.0	0.0	8.3	50.0	0.0	16.7	33.3	25.0
	OpenCode	12	12	8.3	0.0	0.0	0.0	0.0	33.3	16.7	16.7	25.0	25.0
	OpenClaw	12	12	8.3	8.3	0.0	0.0	0.0	58.3	0.0	16.7	8.3	25.0
	Hermes	13	12	8.3	0.0	8.3	0.0	0.0	25.0	0.0	16.7	41.7	25.0
	Goose	12	12	0.0	0.0	0.0	0.0	0.0	16.7	0.0	16.7	66.7	8.3
	Nanobot	12	12	16.7	0.0	8.3	0.0	8.3	33.3	0.0	0.0	33.3	0.0
GPT Terra	Claude Code	12	12	8.3	0.0	0.0	0.0	0.0	33.3	0.0	16.7	50.0	8.3
	Codex	12	12	0.0	0.0	0.0	0.0	0.0	75.0	0.0	25.0	0.0	25.0
	OpenCode	13	12	16.7	0.0	0.0	0.0	0.0	33.3	8.3	41.7	8.3	33.3
	OpenClaw	14	12	8.3	8.3	0.0	0.0	0.0	25.0	8.3	50.0	0.0	33.3
	Hermes	12	12	8.3	0.0	25.0	0.0	0.0	33.3	8.3	8.3	41.7	16.7
	Goose	14	12	0.0	0.0	0.0	0.0	0.0	33.3	8.3	25.0	41.7	25.0
	Nanobot	13	12	25.0	0.0	8.3	0.0	8.3	25.0	0.0	25.0	16.7	33.3
Kimi K3	Claude Code	13	12	16.7	16.7	0.0	0.0	8.3	16.7	0.0	41.7	8.3	33.3
	Codex	12	12	16.7	0.0	0.0	0.0	0.0	8.3	8.3	8.3	58.3	0.0
	OpenCode	13	12	8.3	0.0	0.0	8.3	0.0	33.3	0.0	25.0	25.0	33.3
	OpenClaw	12	12	25.0	8.3	0.0	0.0	0.0	50.0	0.0	25.0	0.0	25.0
	Hermes	13	12	25.0	8.3	8.3	0.0	8.3	41.7	0.0	25.0	8.3	33.3
	Goose	14	12	25.0	8.3	0.0	0.0	0.0	41.7	0.0	25.0	8.3	33.3
	Nanobot	14	12	8.3	0.0	33.3	8.3	16.7	33.3	0.0	16.7	8.3	33.3
Qwen3.8	Claude Code	13	12	0.0	8.3	0.0	8.3	0.0	33.3	8.3	16.7	33.3	33.3
	Codex	14	12	8.3	0.0	0.0	8.3	0.0	33.3	0.0	25.0	25.0	33.3
	OpenCode	12	12	8.3	16.7	0.0	8.3	8.3	16.7	0.0	16.7	33.3	33.3
	OpenClaw	14	12	8.3	8.3	0.0	8.3	0.0	25.0	0.0	25.0	25.0	33.3
	Hermes	14	12	8.3	8.3	0.0	8.3	0.0	33.3	0.0	25.0	16.7	33.3
	Goose	13	12	8.3	0.0	0.0	0.0	0.0	33.3	8.3	33.3	16.7	25.0
	Nanobot	12	12	0.0	8.3	25.0	0.0	25.0	66.7	0.0	8.3	8.3	8.3

E Execution Case Studies

The cases below follow three execution transitions: feedback into corrected actions, retained evidence into continued work, and task requirements into final artifacts. Same-task, same-model contrasts are complemented by a successful recovery and cross-model tool-use profiles. Request indices refer to provider traces; OpenCode may include a separate title-generation request. Case scores use the original 0–1 scale.

E.1 Tool Feedback and Recovery

Model-specific tool use. Table 25 separates incompatible tool selection from its eventual recovery. The same Goose tool surface is associated with different observed behaviors across models. For example, DeepSeek Pro uses an image reader on the text contract in task 081, then recovers through a shell read. DeepSeek Flash attempts unavailable text-reading calls in task 222, receives the available-tool list, and switches to shell reads. Qwen also recovers image-reader errors in task 041 by using shell reads. These positive corrections coexist with the persistent erroneous calls illustrated in Table 26. A separate Qwen interface audit finds 273 image-reader calls targeting common text-file extensions across 153 tasks. Interface friction and failure to recover are distinct observations; neither is inferred solely from the final score.

	Goose			OpenCode
Model	Wrong type	All tool errors	Recovered	Wrong type
DS Flash	21	25	24	0
DS Pro	28	33	31	0
Gemma 4	0	9	8	0
GLM-5.3	5	13	12	0
GLM Flash	13	16	15	0
GPT Luna	0	13	12	2
GPT Sol	1	13	11	0
GPT Terra	0	6	6	1
Kimi K3	2	6	5	0
Qwen3.8	35	41	40	0

Table 25: Tool-use labels in the general probability sample, with 80 runs per model–harness configuration. Wrong type is the codebook's wrong-tool-type subtype; Recovered counts recovery among all tool-error labels. The sampled tasks and score strata differ across configurations, so these are descriptive profiles rather than paired estimates of population incidence.

From reasoning to execution. Table 26 compares Qwen executions on the same tasks. The cases were selected from attempts where Goose received zero credit and OpenCode received at least 0.80, followed by inspection of the task instructions, model responses, tool feedback, and final artifacts.

Table 26: Paired semantic and action-level evidence. The observations identify concrete differences in the recorded executions; scores in the first column are the official results. Full request and artifact evidence is preserved in the analysis logs.

Task and scores	Goose trajectory	OpenCode trajectory	Observed distinction
137: Calibration-certificate recovery Goose: 0.00 OpenCode: 0.92	Request 10 identifies the wrong image tool and incomplete filename, but repeats both in its next calls. The run contains 17 explicit tool-error results. It eventually reads the controller and queries the database, but never writes the plan; six new artifacts are absent.	Reads the controller at request 3 and writes the plan at request 8. Preview, apply, verify, staging, and replay follow. The primary outcome checks pass.	A correction expressed in reasoning does not necessarily change the emitted operation.
105: Data-management package staging Goose: 0.00 OpenCode: 0.94	Identifies the main consent restrictions and repository routes, then repeatedly discusses reference syntax, repository alternatives, and future handoff details. No plan is written; six deliverables are absent.	Writes the plan at request 6, builds at request 7, writes the handoff at request 9, and runs validation, staging, and repeated verification at request 10. The primary outcome checks pass.	Settled business decisions need to become intermediate artifacts that execution can check.
125: Product-recall recovery Goose: 0.00 OpenCode: 0.92	Correctly identifies much of the scope and event logic, but continues reasoning about whether replay remains valid after state mutation. No plan is written; four new deliverables are absent.	Writes the plan at request 6, corrects evidence references at request 7, previews and applies it, then reorders its lists and applies it again at request 10. The primary outcome checks pass.	Actual replay provides evidence about stateful behavior that reasoning alone cannot supply.
206: Engineering-change propagation Goose: 0.00 OpenCode: 1.00	Begins a detailed recovery analysis before querying the decisive database records. After receiving them, it derives the plan again. A write call starts but is incomplete; six new artifacts are absent.	Queries the database at request 5, derives and writes the plan, corrects revision-field types, and completes preview, apply, verify, staging, and replay.	Acquiring decision-relevant state before expanding a plan reduces avoidable re-derivation.

A completed recovery. Task 051 requires an offline normalization pipeline for customer CSV, order JSON, and note TXT inputs. The deliverables are normalized records, an index, a rejection ledger, and a batch summary. The task also requires deterministic ordering, globally unique accepted IDs, and an idempotent rerun. In the Gemma–Claude Code execution, the agent inspects its outputs, corrects missing note coverage, and obtains full credit. Table 27 follows the observable sequence. The evidence packet abbreviates the implementation writes, so the case identifies the coverage defect and its correction without assigning an unseen parser-level cause.

Unlike the repeated invalid calls in Table 26, the observed discrepancy here leads to new source inspection, a changed implementation, and a check of the revised outputs. Figure 8 juxtaposes these distinct responses to feedback.

Requests	Observed execution	What changes
4–5	Reads generated files and ledgers. The summary reports 6 accepted records, 19 rejections, and 0 notes.	Output inspection exposes incomplete input coverage.
5–7	Reads the note files, including their Case, Owner, and Status fields.	Acquires the concrete inputs needed for revision.
7–9	Rewrites and executes the pipeline, then checks the outputs again: 9 accepted records, 9 rejections, and 3 notes.	The revision changes the delivered records and ledgers.
9–10	Reruns the transformation and compares hashes of the output files; the command completes with no reported differences.	Checks repeatability of the revised outputs.

Table 27: A recovered coverage defect in a full-credit execution. Request indices follow the provider trace; the final oracle independently evaluates the artifacts.

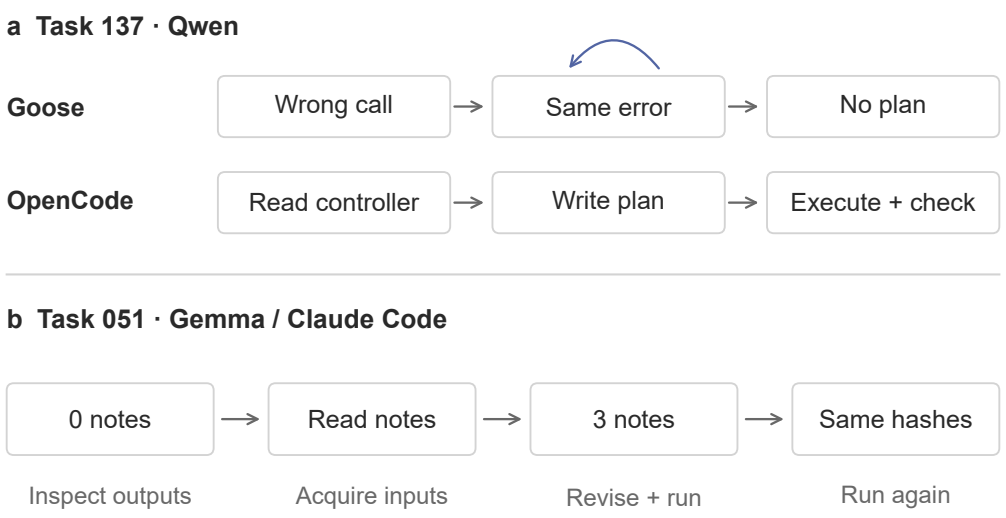


Figure 8: Feedback and recovery in two distinct tasks. (a) Qwen–Goose repeats an invalid operation after identifying the mistake in task 137, while OpenCode writes and executes a plan. Goose later reads the controller but leaves the plan unwritten. (b) Gemma–Claude Code detects missing notes in task 051, revises the pipeline, and verifies the rerun.

E.2 Evidence Retention and Context Recovery

Retaining evidence needed for implementation. In task 142 (trace-context propagation repair), the DeepSeek Flash–Nanobot run first writes implementation code at provider request 46, after repeated small-range reads. At one observed context transition, retained assistant reasoning grows from 65,760 to 142,765 characters while tool-result content decreases from 17,185 to 3,792. The inspected compaction implementation can replace recent tool results under context pressure. These counts describe retained text, not tokenizer-based token usage. The corresponding task uses 67 provider requests, compared with Codex's 12 and Goose's 14.

Resuming after a context-budget rejection. The print-preview task (multimodal 026) starts without Node.js installed; preparing the required build dependencies is part of task execution. Paired Qwen–Nanobot and Qwen–Goose runs both encounter a context-budget rejection. Figure 9 compares their recovery paths. Nanobot's two rejected requests retain 115 messages and the same output

reservation, changing image placeholder text without materially reducing the request. Goose instead makes a separate summarization call and resumes execution with a summary and the original task; the recorded input decreases from 64,007 tokens in the preceding successful request to 8,398 in the resumed request. Earlier, Nanobot recovers compacted tool evidence through smaller reads. Neither run establishes a usable Node toolchain; Goose resumes execution but writes the distribution artifact by hand without a standard build.

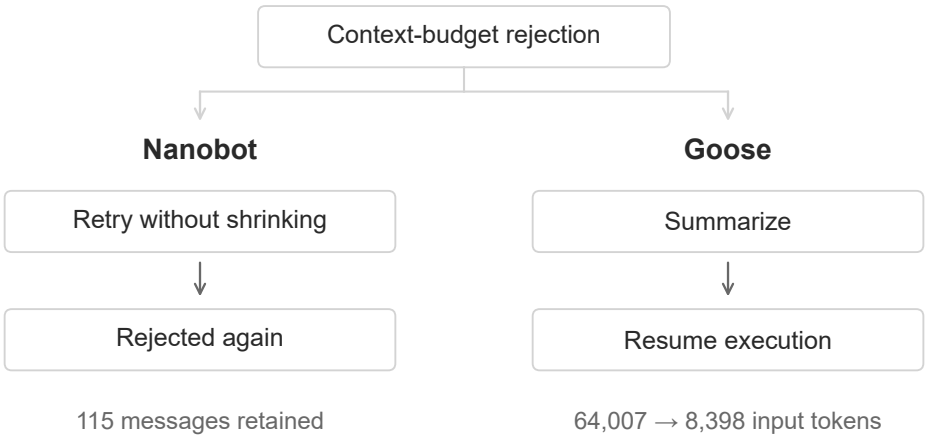


Figure 9: Qwen executions on multimodal task 026 encounter context-budget rejection. Nanobot retries without materially reducing the request, whereas Goose summarizes and resumes. Input-token counts compare Goose's last successful request before rejection with its resumed request; the summary is a separate call.

E.3 Requirements and Delivered Artifacts

Policy decisions and supporting clauses. Task 022 explicitly requires every materially applicable secondary clause, alongside the controlling clause. In case T104, the missing itemized receipt triggers exception EX-1 as well as the lodging rule. We inspect the final JSON artifacts of all 70 configurations and condition the comparison on passing the primary-rulings module. Of the 64 qualifying runs, 21 omit EX-1 from both the controlling and secondary references. Table 28 gives the model-specific counts. The count tracks the explicitly required EX-1 reference in the final artifacts.

Model	Primary-rulings pass	EX-1 retained among these
DS Flash	7 / 7	1 / 7
DS Pro	7 / 7	3 / 7
Gemma 4	5 / 7	0 / 5
GLM-5.3	7 / 7	7 / 7
GLM Flash	7 / 7	4 / 7
GPT Luna	7 / 7	7 / 7
GPT Sol	7 / 7	7 / 7
GPT Terra	5 / 7	4 / 5
Kimi K3	7 / 7	7 / 7
Qwen3.8	5 / 7	3 / 5

Table 28: The same explicit exception requirement across models and harnesses. Counts refer to the recorded final artifacts; all seven harnesses are evaluated per model.

For DeepSeek Pro, both Codex and OpenCode pass the primary rulings, blocking-information, and output-package modules. OpenCode fails the cited-reasoning module on secondary-clause references. The omitted EX-1 reference in T104 is one concrete difference between their supporting reference sets. Codex's request 7 marks its plan complete after schema checks, but requests 8–11

revisit T106 and T109 and reconcile the JSON and CSV. The final outputs agree with OpenCode on all 11 case-level decisions, controlling clauses, and reason codes.

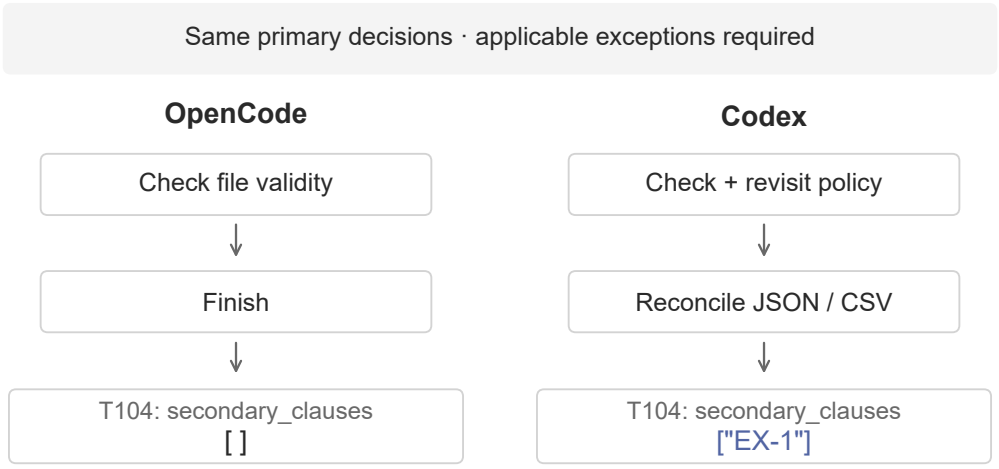


Figure 10: DeepSeek Pro executions on task 022 agree on all 11 primary decisions but differ in supporting references. Codex continues policy review after structural checks; the final-artifact comparison shows EX-1 retained for T104 in Codex and omitted in OpenCode.

An attachment omission propagates into a reminder. Figure 4 visualizes task 274 with GPT Luna. The task requires selecting the latest attachment per person and material type, preserving its bytes, filling a register, and reminding only participants with missing materials. Both runs receive the same relevant metadata at request 5: m004 contains u003's resume a004 and declaration a005; m005 contains resume a006, explicitly replacing a004. The declaration is neither replaced nor late.

Hermes downloads a006 but omits a005, reminds u003 at request 6, and registers the declaration as missing at request 7. Archive selection, register, and reminder agree on this incorrect absence. It later inspects and hashes the seven downloaded files and reports successful verification, without reconciling the selection against all eligible attachments. Goose downloads both a005 and a006, registers both as submitted, and does not remind u003. It archives eight attachments overall and receives 1.00; Hermes receives 0.00.

Hermes separately recovers terminal-inspection errors by switching to the filesystem interface, but does not recover the attachment. Checking a person-by-material evidence table before sending reminders would make each missing-material decision traceable to the eligible attachments.

F Task Construction and Representative Examples

F.1 From Workflow to Verified Task

Task 274 (Business Operations / Coordination) illustrates the mapping from a practical work objective to environment state, operational constraints, and acceptance checks. The objective is to collect two materials from each participant, retain the latest valid submissions, update a register, and remind only those with missing materials. The environment represents the participants and submissions; the constraints determine which submissions count; the oracle checks that the archive, register, and reminders implement those decisions.

Environment construction. The workflow entities become concrete resources: a six-person roster represents participants, while ten timestamped messages and eleven attachments represent submissions. Collection rules and a reminder template specify how these resources should be used. A local collaboration service provides tools for reading messages, downloading files, updating a spreadsheet, and sending reminders. This reconstruction preserves the dependencies between evidence and actions while fixing the starting state for every configuration, without contacting a live platform.

Workflow refinement. The fixtures instantiate version changes, deadline boundaries, and out-of-roster content, making the collection rules consequential. A later resume replaces an earlier resume for the same person, but does not invalidate a declaration attached to the earlier message. An administrator's sample is outside the roster, and a late submission falls after the cutoff. Instructions make the downstream obligations explicit: preserve selected attachment bytes, archive them by person and material type, maintain a fixed-order register, and send exactly one reminder to each person still missing materials. These requirements connect evidence grounding to output consistency: an omission during selection can propagate into both the register and reminders even when those outputs agree.

Oracle and reference solution. Each module checks a work product against the collection rules. Selected versions must appear faithfully in the archive and register (0.40); missing-material decisions must produce the correct reminder recipients and content (0.40); complete delivery earns the remaining 0.20. The oracle checks final artifacts and persisted service records, awarding a module's weight only when all its checks pass. A separate gate checks protected inputs and service resources. The reference solution downloads eight files, writes six register rows, sends three reminders, and receives full credit. Human review checks environment executability and agreement between the instructions and scoring requirements. Figure 4 shows how the version-selection dependency leads to different outcomes under two harnesses.

F.2 Representative Tasks

These eight examples illustrate all six work categories and eight subcategories through six general and two multimodal tasks. They were selected for contrasting execution requirements. Instructions are condensed for presentation; module weights follow the task specifications. Each module earns its weight only when all its checks pass; the separate gate checks prohibited actions and protected-resource integrity.

T004: Reconcile a sales upload with a management memo

Prepare a Q3 regional sales close package. Apply the recognition policy to orders and adjustment events, then deliver a reconciled CSV upload and an editable Word memo.

Input workspace

Transaction sources

`orders_q3.csv`
24 orders

`returns_and_adjustments.csv`
14 refund events

Selected adjustment facts

Record	Evidence
ADJ-N4	Replaces ADJ-N3
ADJ-S4	Reverses ADJ-S3
ADJ-W3	Settles 4 h before cutoff
ADJ-S2	Settles on 10-06

Policy and template

`sales_recognition_policy.md`
- Recognition rules and cutoff

`regional_sales_memo_template.docx`
- Preserve header and footer

Keep source files unchanged.

Decision points

Resolve versions

- ADJ-N4 replaces ADJ-N3: count once.
- ADJ-S4 reverses ADJ-S3: remove the refund.

Apply the cutoff

- ADJ-W3 settles before cutoff: deduct.
- ADJ-S2 settles on 10-06: report as an exception.

Check eligibility

- ADJ-N2, ADJ-W2 and ADJ-E2 are unsettled or tied to a pending order.

- Disclose these items; do not deduct.

Required outputs

`regional_rollup.csv`
Required rollup content

Content	Scope
Order counts	By region
Recognized amounts	By region

`regional_sales_memo.docx`
Required memo structure

Results
- Native Word table matching the CSV.

Interpretation
- Company reconciliation and exceptions.
- Leading region and unbooked exposure.

Template
- Keep the original header and footer.

Use the same figures in both artifacts.

Oracle modules

* Core module

Regional rollup*	Adjustment / cutoff*	Usable Word memo*	Management interpretation	Professional handoff
30%	25%	25%	10%	10%

T274: Close a materials collection without duplicate reminders

Collect two required submissions from each of six people in a local channel. Archive the latest eligible files and send one templated reminder only to people with missing materials.

Input workspace

Collection rules and roster

`inputs/material_requirements.md`
Required materials and cutoff

`inputs/person_roster.csv`
Six listed people; fixed roster order

`inputs/reminder_template.txt`
Prescribed reminder wording

`hb_tools/` Collaboration CLIs

Selected channel records

Item	Sender	Submission
a004	u003	Resume, earlier
a006	u003	Resume, latest
a011	u002	Declaration
a010	admin001	Off-roster file
m008	u006	No attachment

Preserve inputs and protected service state.

Decision points

Select the eligible revision

- Use the latest eligible file per person and type.
- a006 replaces a004 for u003: archive a006.

Apply the cutoff and scope

- Cutoff: 07-20 10:00 (+08:00).
- a011 arrives at 10:30: exclude as late.
- a010 is off-roster; ignore example files.

Remind missing submitters

- m008 has no attachment: u006 is incomplete.
- Verify one send per recipient.

- No reminders to complete members.

Required outputs

`archive/<id>/<type>/`
Byte-exact attachment archive
- Store the original bytes of eligible files.

`outputs/material_register.csv`
One consistent material register

Content	Requirement
People	Roster order
Materials	Completeness

`reminders via collab_cli`
Exactly one reminder per person

`u002` `u004` `u006`

Use the supplied template wording.

Verify the archive, register and reminders.

Oracle modules

* Core module

Archive fidelity and register*	Reminder recipients and content*	Complete delivery and verification
40%	40%	20%

T142: Repair trace propagation across trust boundaries

Repair a service-mesh adapter that mishandles malformed trace state and leaks sensitive baggage to external services. Implement the contract generally, then run and inspect the outputs.

Input workspace

Contract and implementation

`docs/trace_contract.md`
Trace propagation rules

`policy.json`
Vendor key and sensitive keys

`src/ tests/ bin/`
Adapter, tests, and runner

Public request contexts

`requests.json`

q1 · internal
Valid trace; participates
`tracestate: acme=old`

q2 · external
All-zero trace ID
`baggage: auth=secret`

q3 · external
Valid trace; pass-through
`tracestate: bad==value`
`baggage: user.id=bob`

Keep protected fixtures intact.

Decision points

Choose trace identity

- Forward, participate, or restart each trace.
- Use the correct flags for each disposition.

Validate optional state

- Drop malformed tracestate or baggage.
- Retain a valid trace when optional state fails.

Enforce the trust boundary

- Strip sensitive baggage at external boundaries.
- Do not hard-code request IDs or policy keys.

Input validation
Malformed input must fail predictably.
Replay fixtures to reproduce decisions and metrics.

Required outputs

Adapter and execution artifacts

`src/trace_context.py`
Repaired adapter

`out/decisions.json`
One decision per request

`out/metrics.json`
Disposition and drop counts

`TRACE_NOTES.md`
Identity, state, privacy, verification

Required public-case behavior

Hidden variants differ.

Req.	Required treatment
q1	Participated; acme=v2 first.
q2	Restarted; auth removed.
q3	Forwarded; invalid tracestate dropped; user.id removed.

Offline only: no network, live telemetry export, or service mutation.

Oracle modules

* Core module

Identity / participation*	State / privacy*	Fixture replay*	Validation / offline*	Trace handoff
30%	25%	10%	15%	20%

T007: Recover the right database history, then prove it

Build an isolated recovery candidate for a synthetic PostgreSQL incident. Select a valid backup on the correct timeline and an unambiguous restore target; leave production untouched.

Input workspace

Backup candidates

3 backups · production timeline 2

Candidate	Assessment
`base-t2-0840` Timeline 2	Reject Bad checksum
`base-t2-0920` Timeline 2	Select Verified backup
`base-t3-1005` Timeline 3	Reject Abandoned drill

Evidence and controller

`evidence/*.json`
Backups, WAL, incident evidence

`policy/recovery_policy.json`
Recovery rules and constraints

`pitrctl.py · pitr_lib.py`
Local recovery controller

`state/recovery_state.json`
Controller-managed state

Keep evidence and controller intact.

Decision points

Match the incident timeline

- A valid checksum is not enough: match the backup to production timeline 2.
- Select the verified base-t2-0920 backup.

Choose an unambiguous target

Target candidate	Assessment
`10:15:42Z` RP-880 and TX-881 share it	Reject Ambiguous
`before_batch_cleanup` Restore point · 0/C120000	Select Named target

Use a named target for shared timestamps.

- Verify continuous WAL segments.

Prove the isolated recovery

- Verify the bad transaction is excluded.
- Run two identical probes.
- Replay idempotently; never promote.
- Stop at manual review.

Required outputs

Recovery plan and probes

`work/recovery_plan.json`
Selected backup and named target

`work/probe_report.json`
Two identical probes

Receipt and handoff

`work/recovery_receipt.json`
Gate receipt and idempotent replay

`work/recovery_handoff.md`
Timeline, rollback, promotion boundary

Controller state

`state/recovery_state.json`
Update only through pitrctl.py.

State	Required property
Candidate	Isolated
Production	Unchanged

Production remains untouched.

Oracle modules

* Core module

Backup / timeline*	WAL / target plan*	Restore / probes*	Receipt / replay*	Recovery handoff
20%	20%	25%	15%	20%

T088: Repair a revenue join without multiplying orders

Repair a failed monthly close whose draft SQL multiplies orders by joining raw dimensions. Build a reusable many-to-one pipeline, then rerun the protected close.

Input workspace

Orders, dimensions and runner

`orders.csv` 12 orders

`customers.csv, rates.csv`
Dimension rows may duplicate keys.

`policy.md` Join contract

`pipeline/revenue_join.sql`
Failing draft; modify this SQL only.

`bin/run_close.py`
Protected close runner

Customer dimension evidence

Order / key	Customer rows
o2 / c2	2 rows; 1 current
o5 / c3	2 current rows
o9 / c4	2 rows; none current
o4 / c9	No row
o3 / null	Null keys never match

Evidence: 7 of 12 orders; held-out keys differ.
Inputs, policy and runner stay read-only.

Decision points

Resolve duplicate keys first

- Resolve a duplicate only when exactly one row is current.
- o2 matches; o5 and o9 remain ambiguous.

Preserve one row per order

- Keep all 12 orders exactly once.
- Flag o3 as null-key and o4 as unmatched.
- Null keys must never match.

Count valid refunds

- Count valid refunds only if both joins match.

Rate and refund evidence

Order / key	Treatment
o6 / GBP 07-01	2 current: ambiguous
o12 / c5, -10	Both match: count refund

Required outputs

`pipeline/revenue_join.sql`
Reusable many-to-one query
- Resolve keys before joining orders.

`close/order_join.csv`
Unique-order close ledger

Scope	Invariant
All 12 orders	One row each

`close/join_audit.json`
Reconcile the executed close

Check	Coverage
Join audit	Counts and reasons
Close totals	Revenue, cardinality

`close_handoff.md`
Maintainer handoff
- Repair and safety on new data.

No external financial-system update claims.

Oracle modules

* Core module

Visible-data close ledger*	Held-out keys / duplicates*	Executable final close*	Reconciled audit	Maintainer handoff
30%	30%	20%	10%	10%

T125: Resume a recall without applying the transaction twice

Recover a product-recall transaction after a worker crash. Reconcile a newer signed scope, an interrupted journal, and an out-of-order feed with a duplicate and an unknown shipment.

Input workspace

Recovery inputs

`recall.db`
Lots, shipments, and journal

`signed_scope_revisions.json`
Three scope revisions

`shipment_events.jsonl`
7 events across 8 feed lines

`recall_checkpoint.json`
Interrupted state at EV-504

`recall_policy.md`
`recall_seed.sql`

Choose the trusted scope

Revision	Trust / treatment
SCOPE-01	Signed; 02 supersedes.
SCOPE-02	Signed 08-31; descendants; controls.
SCOPE-03-DRAFT	Draft signature; ignore.

Change state only via recallctl.py; never replace the database.

Decision points

Trust the signed scope

- Use the latest trusted signed revision; drafts never control.
- Follow descendant lots in SCOPE-02.

Reconcile shipment events

- Per known shipment, apply only the newest sequence above stored state.
- Quarantine unknown shipments; apply duplicate events only once.

Resume once, then replay

- Resume prepared EV-504 once from the interrupted checkpoint.
- Replaying the transaction must be a no-op.

Inventory, lot, and channel closure

Keep quarantine conservative. Verification and staging receipts must bind the final state.

Required outputs

Recovered state and evidence

`recall.db`
Updated local state

`recall_plan.json`
Evidence-linked recovery plan

`recall_verification.json`
Verification report

`recall_stage_receipt.json`
Stable receipt bound to final state

`recall_handoff.md`
Complete, reviewable recall handoff

Required event handling

Selected evidence · 4 of 7 events

Event	Evidence / treatment
EV-500	SH-A: seq. 2 after seq. 3; stale.
EV-502	Exact duplicate; apply once.
EV-505	Unknown shipment; quarantine.
EV-504	Prepared at crash; resume once.

No forged receipts or real notices, refunds, or stop-sale actions.

Oracle modules

* Core module

Signed scope / genealogy*	Events / shipment state*	Inventory / lot / channel closure*
22%	18%	22%

Journal recovery / replay*	Verification / local staging	Recall handoff
18%	12%	8%

M024: Turn visual differences into an interactive annotator

Build a local React/TypeScript app over a protected before/after image. Read marker position, side, and severity from pixels; records provide IDs and labels only.

Protected visual evidence



Protected input · before-after.png

Read from pixels

4 visible differences

ID	Side	Severity
DIFF-NAV	Before	Review
DIFF-BTN	After	Review
DIFF-PRICE	After	Watch
DIFF-GAP	After	Info

Pixel-derived attributes

Crosshair
Marker center
Before / after window
Side
Badge color
Severity

Input workspace

Protected assets

`web/public/before-after.png`
Before / after image with crosshairs

`web/src/data/localRecords.json`
Four IDs and labels; no geometry

Local app starter

`web/` Vite · React · TypeScript
Pinned local dependencies

Preserve the image and records.

Decision points

Anchor markers to pixels

- Center each marker on its crosshair.
- Infer side from the window and severity from badge color.

Keep annotation state in sync

- Select a marker to show its details.
- Side filters and marking controls act on the same records.

Mark / unmark updates all linked views.

Required outputs

Local app artifacts

`web/src/` Annotator interface
`web/dist/` Built application

Synchronized views

Open diffs | Diff count
After-only | Overlay

Responsive, accessible surface

Desktop / mobile · focus · reduced motion

Oracle modules

* Core module

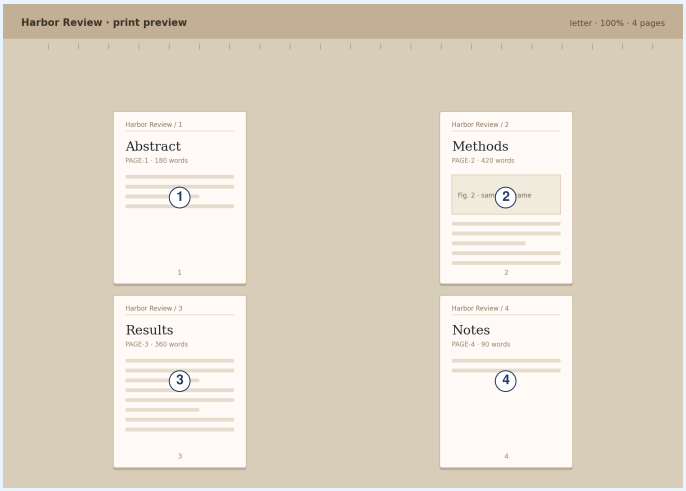
Build / runtime contract*	Protected records / image*	Markers, selection, summaries*	Responsive accessible surface
22%	22%	36%	20%

Boundary: preserve protected inputs; use pinned local dependencies, with no package locks or remote services.

M026: Connect a page image to a working print preview

Build a local paginated preview over a protected article-page image. Read page centers, page numbers, and word counts from pixels; records provide IDs and titles only.

Protected visual evidence



Read from the image

Page	Section	Words
1	Abstract	180
2	Methods	420
3	Results	360
4	Notes	90

Records are not geometry

IDs and titles only; no page centers or word counts.

Illustrative pinned state

Pages 1 + 3 pinned

2 pages · 540 words

Numbered overlays identify the page centers.

Input workspace

Protected page image
`web/public/print-pages.png`
Four page layouts and printed counts.

Local records
`web/src/data/localRecords.json`
Four IDs and titles; no geometry.

Starter app
`web/`
Vite + React project.

Keep the image and records unchanged.

Decision points

Locate all four pages
Read page centers, printed page numbers, and word counts directly from pixels.

Connect the interactions
Page selection, paper-size controls, and pin/clear actions must work together.

Keep summaries synchronized
Update Pinned pages, Pinned count, Word total, and the running header together.

Required outputs

Print-preview interface
`web/src/`
Page markers, selection, paper size, and pin/clear controls.

Built local app
`web/dist/`

Responsive and accessible
Desktop and mobile layouts, visible keyboard focus, and reduced-motion support.

Oracle modules

* Core module

Build / runtime contract*	Protected records and image*	Markers, selection, summaries*	Responsive accessible surface
22%	22%	36%	20%

Boundary: preserve protected inputs; use pinned local dependencies, with no package locks or remote services.