AGENTICVBENCH-OMNI: EVALUATING OMNIMODAL AGENTS ON ECONOMICALLY VALUABLE WORK

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce AgenticVBench-Omni, a benchmark that tests whether omnimodal agents can turn video, audio, images, and documents into economically valuable work. Its 45 tasks span eight families and work activities linked to 10 occupations, requiring precise records, quantitative annotations, editable assets, or completed environment states. Designed around complex professional workflows, the tasks require sustained evidence gathering, cross-modal reasoning, and tool use while preserving temporal, spatial, and identity relationships. Each task specifies an output contract and uses a fully deterministic verifier to compute scores, without rubric-based grading by a model. Across five frontier agents evaluated three times per task, the strongest agent achieves a mean score of just 0.283, and 60% of attempts score below 0.10. Verifier outputs and execution traces show that agents recover partial information but struggle with precise localization, complete coverage, accurate reconstruction, and self-verification of the final deliverable.

1 INTRODUCTION

Professional work often requires integrating omnimodal information into a usable deliverable while preserving its temporal, spatial, and identity relationships. For example, a music producer reconstructs a mixing session from a narrated screen recording; an operations analyst aligns runway video with radio communications; a scientific annotator links vocalizations to individual animals. Success requires both understanding the evidence and carrying its details into a complete work product.

This motivates a central question: *how do omnimodal capabilities translate into economically valuable work at the complexity expected of human experts?* Existing benchmarks commonly test perception and reasoning through questions, summaries, or short responses (Fu et al., 2025; Zhou et al., 2025; Li et al., 2025). These formats leave unclear whether an agent can sustain accurate evidence integration through a complete professional workflow. Recent benchmarks add multimedia tools and workspace artifacts (Heo et al., 2026), but evaluation still needs to connect occupational breadth, complex deliverables, and precise completion criteria. We address this need with tasks designed to require expert-level attention to detail, sustained tool use, and verifiable outputs.

We introduce **AgenticVBench-Omni**. **Occupational breadth** spans 45 tasks across eight families, with activities linked to 10 occupations (Figure 1). These occupations account for approximately $67 billion in annualized U.S. wages, estimated from employment and mean wages in the May 2023 BLS data (U.S. Bureau of Labor Statistics, 2024); Appendix A.1 details the calculation and mapping. **Workflow realism** comes from recurring professional activities, such as reconstructing an editable production project or compiling a complete operations ledger, with enough dependencies and precision requirements to demand sustained reasoning. **Fully deterministic verification** computes scores through task-specific checks on outputs, without rubric-based grading by a model. The acceptance protocol also requires a reference solution and human checks of solvability. Section 2 presents task examples, benchmark construction, and quality control.

We evaluate five frontier agents with three runs per task. GPT-6-Astra has the highest mean score, just 0.283; 60% of all attempts score below 0.10. GPT-6-Astra often groups media processing and inspection within a single decision. Claude Opus 5.5 frequently reuses saved scripts and delegates evidence inspection; in delegating runs, its helpers perform most image views. The two Flash agents more often execute code inline. Across these strategies, agents share a gap between recovering partial evidence and delivering precise, complete results.

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



Figure 1: Task families and example deliverables in AgenticVBench-Omni. Sector angles and rim counts show the distribution of 45 tasks; each outer segment represents one task. The surrounding cards pair an input frame with a requested deliverable. Across families, 37 tasks request records or quantitative annotations, seven request reconstructed assets or editable projects, and one requests a resulting environment state.

Most zero-score executions in the diagnostic sample still produce output files, and local self-checks can confirm file structure while missing errors in the content being scored. These findings make evidence integration, complete delivery, and self-verification against the task's acceptance criteria central challenges.

2 Benchmark Construction

2.1 Task families and deliverables

The unit of evaluation is a bounded work activity with a specified deliverable. Figure 1 shows the eight families: sports analysis, game-state reconstruction, scientific observation, transport and operations, meetings and procedures, spatial measurement and tracking, music and synchronization, and asset reconstruction. Contributors design and build tasks within these predefined domains. Identity tracking, documentation, and measurement recur across families. Appendix A.1 gives the occupational mapping and full output breakdown (Tables 3 and 4). These activity-level links are informed by O*NET descriptions; the collection is not weighted by employment or economic output.

Each task connects a work request to source materials, an output contract, available tools, and a resource budget. Video, audio, images, and documents appear across the collection; individual tasks use different combinations. The contract specifies the artifact or environment state and its required fields, units, and relationships. Of the 45 tasks, 37 request structured records or quantitative annotations, seven request reconstructed assets or editable projects, and one requests a resulting environment state (Appendix A.1, Table 4). All packages include solve instructions, a configuration, and a verifier directory. Agents choose how to inspect evidence and construct the output within the task's access policy and time limit of 1 to 15 hours. These limits were set at design time from estimates of expert completion time.

2.2 TASK EXAMPLES

Figure 2 illustrates three tasks through their source evidence, deliverables, and deterministic checks. In **asset reconstruction**, agents rebuild an editable mixing session from a narrated screen recording and audio assets; the verifier checks track settings, processor chains, and routing. In **transport and operations**, agents align runway video with radio communications to link operations and aircraft callsigns; scoring jointly checks time, operation, aircraft type, and callsign. In **scientific observation**, agents combine video and audio to attribute vocal bouts to individual birds; the verifier checks identity and temporal boundaries for the annotated bouts.

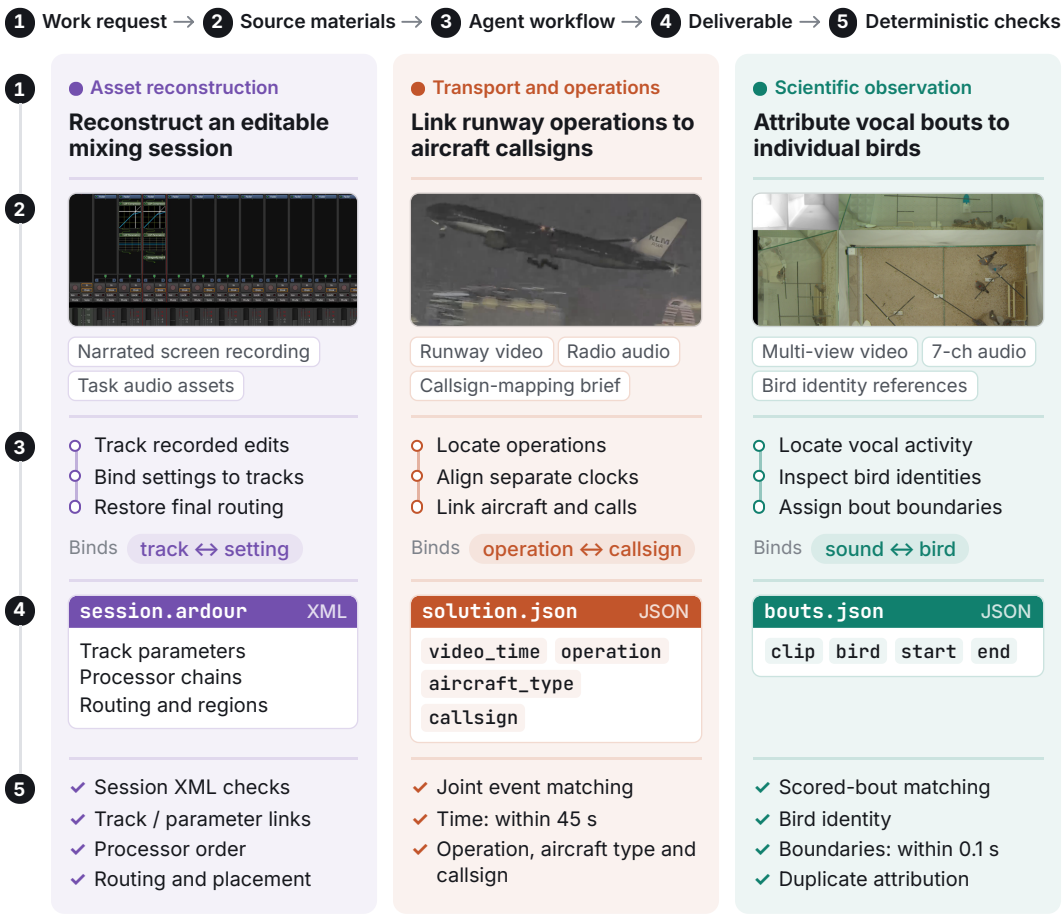


Figure 2: Three example tasks from asset reconstruction, transport and operations, and scientific observation. Images are cropped task input frames (the runway frame is brightness-adjusted); workflows and output structures are schematic. Each example connects a work request to evidence, a deliverable, and deterministic checks.

2.3 CONSTRUCTION AND ACCEPTANCE CRITERIA

Contributors propose the task, assemble evidence, define the output, and provide reference solutions and acceptance rules. The construction protocol has two review loops (Figure 3). In design review, a proposal agent and a maintainer check scope, evidence, and difficulty. In implementation review, contributors submit the specification, oracle solution, verifier, development scores, and raw trajectories. Reviewers request corrections, rerun the oracle and a strong agent, and inspect attempts for shortcuts. Affected checks are repeated before resubmission. Human checks and committee review complete the protocol. Appendix A.2 provides further implementation details.

References come from source annotations, logged states, controlled transformations, or target assets. Source-derived tasks retain inherited annotations while adding a work contract. In meetings and

procedures, examples include a cooking-step ledger and an assembly-state history. Independent reimplementations reproduced reference transformations for these workflows and for cell-lineage and bird-behavior records in scientific observation.

Fully deterministic verifiers assess correctness against reference values, completeness over scored items, consistency of identities and relationships, and functionality where a task defines a behavioral or state test. Verification covers the specified output requirements. For the mixing-session example in Section 2.2, it checks the saved track and processor settings; it does not audition the mix.

2.4 Quality control and difficulty validation

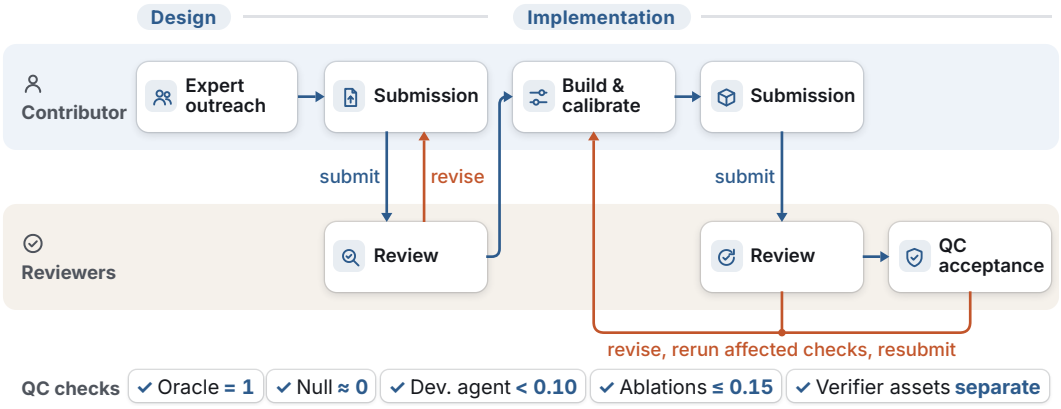


Figure 3: Construction and quality-control workflow. Design and implementation each include review and revision. The lower row gives the protocol's acceptance and calibration targets; input-restriction targets apply where a task defines the corresponding control. Section 2.4 describes the acceptance criteria.

Reference acceptance and evidence sufficiency. The protocol requires full oracle reward, near-zero reward for empty submissions, and manual review of whether the source evidence supports the reference and requested precision. The oracle provides a complete solution under the output contract. Reviewers check both the answer and the evidence needed to recover it, including event boundaries, units, identities, and relationships where applicable. Replaying saved submissions tests whether verifier changes preserve the intended scoring behavior.

Difficulty calibration and shortcut controls. The development protocol targets scores below 0.10 for strong contemporary agents, with trajectories documenting substantive attempts. Low scores require checks for missing evidence, execution failures, and invalid output contracts before they support a difficulty claim. Applicable controls remove media, restrict input to a single frame or modality, or provide a frame dump without tools; their target reward is at most 0.15. Review also checks public annotations, metadata, and visible summaries for shortcuts. Verifier references and oracle assets are kept separate from the solving workspace. Calibration targets are development criteria, not upper bounds on later evaluation scores.

Human review and adjudication. Human review connects the scored reference to the source evidence. In a sports-analysis task, a reviewer who had not seen the draft reference scanned the full video and recorded candidate routines. Author adjudication then resolved inclusion and boundary decisions against the source frames. Of 24 candidate routines, 23 were retained; one was excluded because its mount was not visible. All seven scored fields for the retained routines agree with the adjudication sheet and the oracle. The retained records separate initial annotations from refined timestamps and link them to the final reference and oracle through file hashes. Appendix B documents this process and gives examples of how source annotations become task-specific deliverables.

3 EXPERIMENTS AND ANALYSIS

We evaluate five frontier agents in their vendors' native harnesses (Table 1), abbreviating GPT-6-Astra as Astra. Results describe each model and harness together. Reasoning effort is medium for Astra, Opus 5.5, Opus 5, and Gemini 3.8 Flash; DeepSeek 4.1 Flash uses DSH's default high setting because medium is unavailable. Model, harness, task and verifier version, reasoning setting, and resource budget are held fixed across the three runs for each task and agent; runs differ in stochastic execution. Search and fetch tools are disabled by default. Appendix C documents access exceptions and configuration details.

3.1 HEADLINE PERFORMANCE

Astra and Opus 5.5 achieve mean scores of 0.283 and 0.278, compared with 0.106 for Opus 5, 0.057 for Gemini 3.8 Flash, and 0.058 for DeepSeek 4.1 Flash (Table 1).

Agent	Harness	Mean	95% CI
GPT-6-Astra	Codex 0.154.0	0.283	[0.218, 0.354]
Claude Opus 5.5	Claude Code 2.1.280	0.278	[0.200, 0.359]
Claude Opus 5	Claude Code 2.1.269	0.106	[0.060, 0.165]
Gemini 3.8 Flash	Antigravity CLI 1.2.2	0.057	[0.036, 0.081]
DeepSeek 4.1 Flash	DSH 0.1.5-rc.1	0.058	[0.035, 0.084]

Table 1: Mean task scores over 45 tasks, with three runs per task and agent (675 rollouts). Each task contributes its three-run mean with equal weight. Intervals are 95% percentile bootstrap confidence intervals over task means (20,000 resamples; seed 20260925), with paired task indices across agents. They reflect variation across tasks, holding the observed means fixed. Figure 4 uses the same procedure within each family for its error bars.

3.2 PERFORMANCE AND FAILURES BY TASK FAMILY

Astra leads sports analysis, game-state reconstruction, and spatial measurement and tracking; Opus 5.5 leads the other five families (Figure 4). Margins range from 0.038 to 0.200 across families of three to eight tasks. Spatial measurement and tracking has the lowest mean for Astra, Opus 5.5, and Opus 5. The two Flash agents average approximately 0.011 in this family.

Failure patterns distinguish incomplete delivery from imprecise reconstruction. Opus 5 and the Flash agents receive delivery or coverage labels in 18 to 23 sampled executions each, compared with seven for Astra and 12 for Opus 5.5. The leaders instead lose credit most often on localization. Meetings and procedures, and spatial measurement and tracking, commonly fail on temporal or spatial tolerances; asset reconstruction more often fails on parameter values. Appendix C defines the classification rules and reports their sensitivity.

On six tasks from five families in the diagnostic sample where every agent produced a deliverable, Astra and Opus 5.5 recognize 91% and 88% of reference items, but localize only 35% and 30% within tolerance. Recognition and localization rates are 76% and 17% for Opus 5, 50% and 11% for Gemini 3.8 Flash, and 57% and 7% for DeepSeek 4.1 Flash. These selected tasks illustrate the gap between finding an item and measuring it precisely.

Every execution in the diagnostic sample processes media programmatically. Astra's full-credit game-state reconstruction solution uses change detectors on cropped display regions at a resolution finer than the 750 ms tolerance. By contrast, contact sheets, used in 43 of Astra's 45 rollouts, locate events only to their sampling interval. Agents process audio in 54 of the 55 rollouts whose scores depend on it. Media processing alone therefore does not explain successful completion.

3.3 PERFORMANCE AND COST FRONTIER

Astra averages 27.8 minutes per task, compared with 75.3 for Opus 5.5, and is faster on 40 of 45 tasks. Their mean model-cost estimates are $36.84 and $33.59. Opus 5 takes 91.5 minutes and costs $35.43 while scoring substantially lower (Figure 5). The Flash agents take about 50 minutes

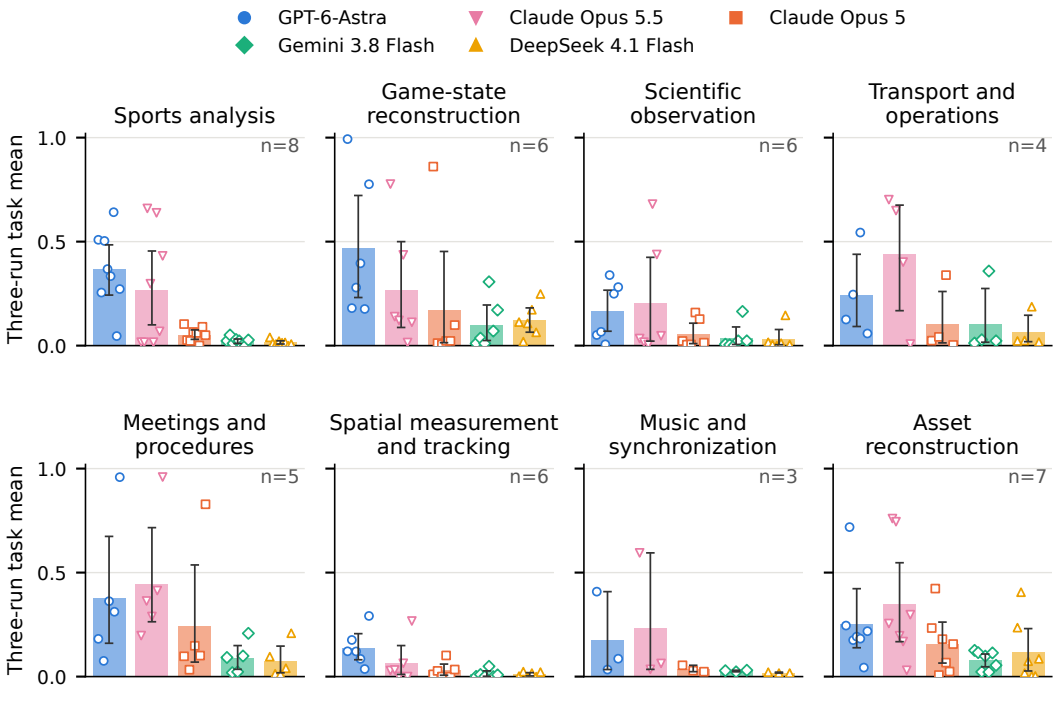


Figure 4: Three-run scores across the eight task families. Bars average the per-task means; open markers show individual task means. Table 1 describes the error bars; n is the number of tasks. Each task has three runs per agent. Interrupted rollouts score zero.

and cost $5.59 and $1.45 per task. These comparisons use matched scores and resources from the diagnostic sample; more time or spending does not consistently produce a higher score.

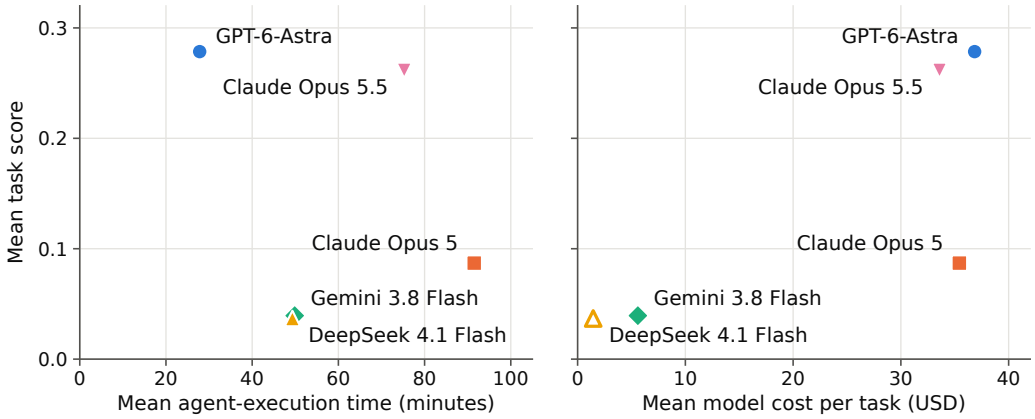


Figure 5: Mean score against execution time and model cost in the matched diagnostic sample, with 45 executions per agent. Filled markers use estimated model costs; DeepSeek uses recorded spend over 44 executions (hollow marker). Costs exclude virtual machines; times exclude setup and verification. Appendix C details the measurements.

3.4 Agent Behavior

Working styles. Opus 5 and the Flash agents execute most code inline, whereas Astra and Opus 5.5 more often reuse saved scripts (Figure 7a). Saved scripts account for a median 40% of Opus

5.5's shell calls, compared with 2.5% for Gemini 3.8 Flash. Opus 5 has the lowest action rate in all eight families (Appendix D.1). This rate counts actions by the main agent and sub-agents per minute, including time spent generating responses, running tools, and waiting. On the 26 tasks where neither Opus agent delegates, Opus 5.5 views a median 2.02 images per minute, compared with 0.85 for Opus 5, and has the higher rate on 23 tasks.

Operation grouping. Astra groups multiple operations into 67% of decisions, compared with 41% for Opus 5.5; only 2.8% of Astra's multi-operation decisions launch calls concurrently. Most bundles connect successive steps, such as running a command and inspecting its output. In one trajectory, Astra records event times from earlier views, generates contact sheets for the next time windows, and opens them within one decision. This connects evidence bookkeeping to the next inspection step. Figure 6 shows the operation breakdown; decision units differ across harnesses.

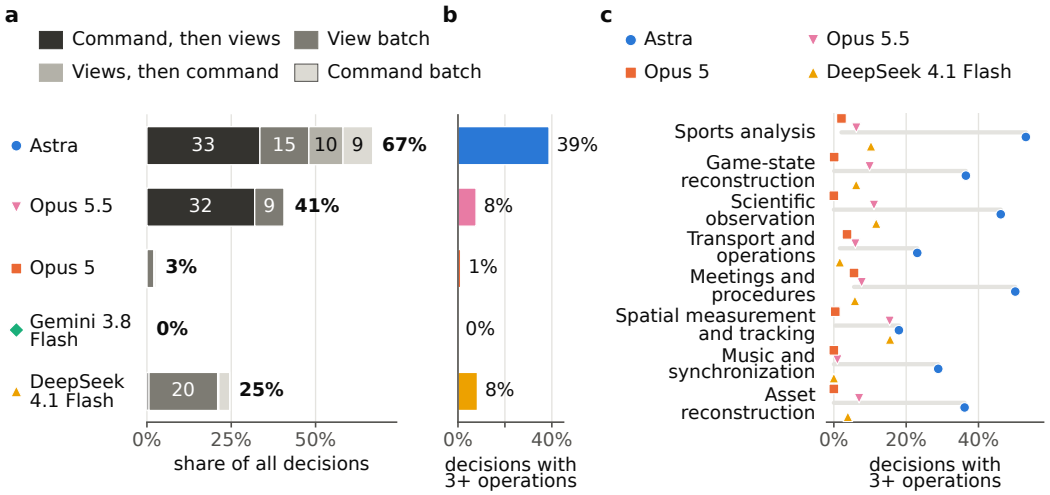


Figure 6: Operations per model decision in the diagnostic sample. (a) Multi-operation decisions, grouped by operation order. (b) Decisions containing three or more operations. (c) The same share by family. Decision units differ across harnesses; Antigravity records one call per step, so Gemini cannot show multi-operation decisions. Appendix D.1 gives the counting rules.

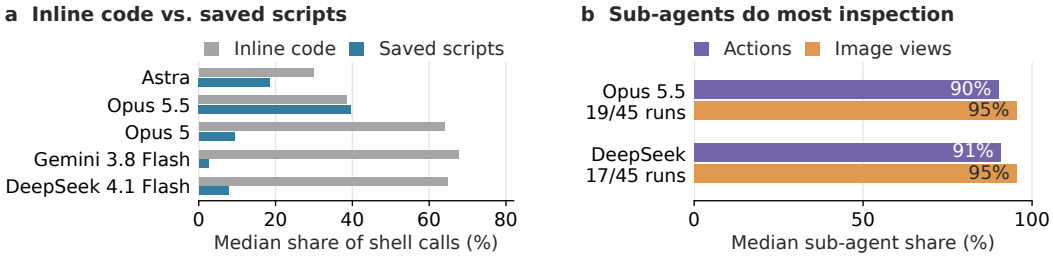


Figure 7: Code execution and delegation in the diagnostic sample. (a) Median shares of shell calls. (b) Median sub-agent shares among delegating runs; counts show delegation frequency.

Delegation. Opus 5.5 and DeepSeek 4.1 Flash delegate in 19 and 17 of 45 rollouts. In their median delegating rollout, sub-agents perform about 90% of actions and 95% of image views (Figure 7b). Opus 5.5's delegating and other rollouts average 0.256 and 0.265, but cover different tasks. Opus 5 and Gemini have delegation tools but do not use them; Astra's system prompt prohibits unrequested delegation. With most media inspection assigned to helpers, the main session must combine their findings into a single deliverable. Identities, time references, and output fields must remain consistent across separately inspected portions of the source.

Self-verification. Agents sometimes check that a project loads, that joint geometry is consistent, or that an output matches their own schema, while missing the task's actual requirements (Table 2). Five Gemini 3.8 Flash submissions use the wrong outer container, and four pass the agent's own check. Adding only the required outer key improves verifier reward in two cases but leaves three at zero.

Task family and deliverable	Agent's local check	Deterministic verifier result
Asset reconstruction: editable mixing session (Astra)	Reloads the session and checks that processor chains survive.	Chains and faders: 1.00; equalizer: 0.364; compressor: 0.40. Loading preserves inaccurate settings.
Spatial measurement: 3-D hand poses (Astra)	Validates 720 joints; maximum bone-length error is 0.00001744 m.	Position score: 0.0271. Consistent bone lengths do not determine correct joint positions.
Sports analysis: table-tennis rally records (Gemini 3.8 Flash)	Accepts a list of 89 rallies and passes its own schema check.	Score: 0. The task requires an object containing the list, so the verifier reads no rallies.

Table 2: Self-verification can check the wrong property. Three cases contrast an agent's local check with the task requirement. Scores measure different task-specific properties; the cases illustrate failure mechanisms without estimating their frequency.

These cases reveal a mismatch between what the agent checks and what the task accepts. A final check should connect the output to the source evidence, test the required structure, and verify scored quantities at the requested precision.

3.5 Failure modes

Files without credit. Of 81 zero-score executions in the diagnostic sample, 63 produce an output file (Figure 8a). Creating a file therefore does not establish that the task was completed. Early writes often produce placeholders: 37 of 38 detected writes in the first 5% of a rollout create empty skeletons or templates.

Missing deliverables. Across all agents, 17 of 18 empty output directories follow a timeout, waiting termination, or process error (Figure 8b). Opus 5 times out nine times, leaving no graded output file in six; Astra never times out. The two outcome groups call for different remedies: preserving partial work before an interruption and correcting an existing artifact that earns no credit. The latter accounts for most zero-score cases in this sample.

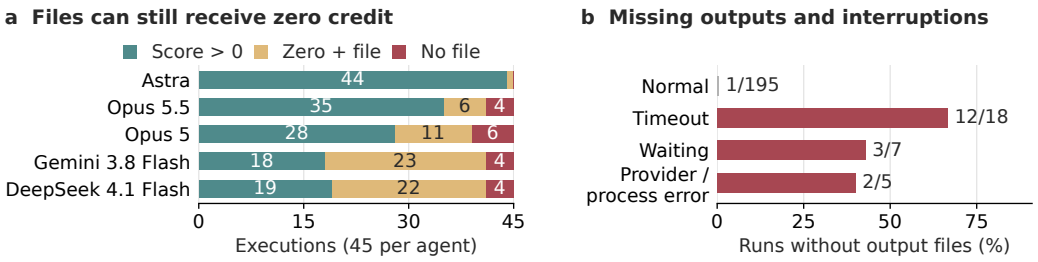


Figure 8: Delivery outcomes in 225 diagnostic executions. (a) Scores and output-file presence. (b) Rates of missing output by termination type; labels give counts and denominators.

Copying prompt examples. Gemini 3.8 Flash copies 12 of 50 fabricated example rows from the instructions into its outputs across six tasks. The other agents copy none.

Misjudging elapsed time. In 23 coded DeepSeek rollouts, 62 of 63 elapsed-time statements made without checking a clock overestimate duration, by a median factor of 3.3. Appendix D.2 gives detection rules, exclusions, and denominators.

4 Related Work

Agents for professional work. Benchmarks evaluate complete workflows in software, web, desktop, and terminal environments (Jimenez et al., 2024; Zhou et al., 2024; Xie et al., 2024; Merrill et al., 2026). SWE-Lancer, TheAgentCompany, GDPval, the Remote Labor Index, and Agents' Last Exam extend this direction to paid projects and occupational deliverables (Miserendino et al., 2025; Xu et al., 2025; Patwardhan et al., 2026; Mazeika et al., 2025; Sun et al., 2026). AgenticVBench applies workflow-oriented evaluation to video post-production (Cao et al., 2026). AgenticVBench-Omni focuses on work products that require integrating evidence across modalities and preserving temporal, spatial, and identity relationships. Its occupational framing connects these requirements to recurring activities, while the output contracts specify the records, measurements, or editable assets an agent must deliver.

Multimodal understanding and tool use. Image, video, and audio benchmarks measure perception and reasoning through question answering and structured responses (Yue et al., 2024; Fu et al., 2025; Zhou et al., 2025; Sakshi et al., 2025; Li et al., 2025). OmniGAIA adds cross-modal reasoning and multi-turn tool use (Li et al., 2026); Omni-DeepSearch and VideoDR connect audio or video clues to web retrieval (Yu et al., 2026; Liu et al., 2026), while M3-Agent uses memory of visual and auditory observations (Long et al., 2026). Other benchmarks connect media understanding to actions: VideoWebArena uses video tutorials to guide web tasks (Jang et al., 2025), and MMTB evaluates terminal workflows over audio and video files (Heo et al., 2026). CutVerse studies GUI-based media editing, while BVB asks agents to reconstruct videos as animated Blender scenes (Hu et al., 2026; Tang et al., 2026). AgenticVBench-Omni brings evidence integration, sustained tool use, and complete deliverables together across eight occupational task families.

Verifiability. RLI evaluates professional deliverables through expert comparisons (Mazeika et al., 2025), and AgenticVBench uses expert-scored rubrics for subjective Repurpose criteria (Cao et al., 2026). Agents' Last Exam uses mostly deterministic checks but retains model judges for some visual and semantic criteria (Sun et al., 2026). AgenticVBench-Omni combines complex professional work and omnimodal evidence integration with fully verifiable rewards across all 45 tasks. Scores are computed by deterministic code using reference values, explicit tolerances, coverage requirements, and output structure. No human or model interprets a rubric at evaluation time. For a fixed task version, the same submitted artifact receives the same score, and failures can be traced to concrete requirements. Component scores also let researchers test specific repairs to an artifact and identify which acceptance requirements still prevent full credit.

5 Discussion and Conclusion

Limitations and future work. Building each task requires more than 40 hours of development, including preparing the evidence and output contract, implementing the deterministic verifier, and checking the reference solution. This cost limits the current benchmark to 45 tasks and constrains coverage within each family. We plan to expand the collection with more tasks while preserving its requirements for workflow complexity, verifiability, and quality control. Within each family, new tasks should vary the source materials, modality combinations, and output requirements to test whether an agent's performance generalizes across instances of the same professional activity. A larger collection would also reduce the influence of individual task designs on aggregate results. Reusable tools for media preparation, reference annotation, and verifier testing could reduce development time while preserving task-specific checks of evidence, output requirements, and precision.

Conclusions. AgenticVBench-Omni connects omnimodal evaluation to complex professional work through 45 tasks across eight families linked to 10 occupations. It combines realistic work requests, explicit output contracts, fully deterministic verifiers, and a construction and quality-control protocol. Across five frontier agents evaluated three times per task, the strongest mean is just 0.283, and 60% of attempts score below 0.10. Verifier outputs and execution traces reveal persistent gaps in precise measurement, complete delivery, and self-verification of scored requirements. The benchmark provides concrete tasks and diagnostic evidence for improving how agents turn omnimodal information into usable professional work products.

AI USE STATEMENT

We used Anthropic's Claude models and OpenAI's GPT-6-Astra to assist with manuscript revision and with literature retrieval, discovery, and citation checking. The benchmark construction protocol includes a proposal agent that assists with reviewing task designs (Appendix A.2). The authors retain responsibility for reviewing AI-assisted text, claims, citations, code, and figures and for all final research and editorial decisions.

ETHICS STATEMENT

The tasks draw on recorded or simulated work activities, including media depicting people. Source-media access and redistribution require task-specific provenance and permissions. Occupational mappings describe bounded activities and do not establish labor savings, professional competence, or readiness for deployment.

REPRODUCIBILITY STATEMENT

Section 3 specifies the evaluated agents, harness versions, and score aggregation. Section 2 describes resource limits. Appendix C records scoring rules, network-access exceptions, verifier changes, failure-label rules, and repeated runs. Appendices B and D document construction checks and behavioral analysis.

REFERENCES

Zongheng Cao, Yi Zheng, Rui Song, and Xinyu Hu. AgenticVBench: Can AI agents complete real-world post-production tasks? *arXiv preprint arXiv:2605.27705*, 2026. URL https://arxiv.org/abs/2605.27705.

Alex Hoi Hang Chan, Neha Singhal, Onur Kocahan, Andrea Meltzer, Saverio Lubrano, Miyako H. Warrington, Michel Griesser, Fumihiro Kano, and Hemal Naik. CHIRP dataset: towards long-term, individual-level, behavioral monitoring of bird populations in the wild. *arXiv preprint arXiv:2603.25524*, 2026. URL https://arxiv.org/abs/2603.25524.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, et al. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118, 2025. URL https://arxiv.org/abs/2405.21075.

Chiyeong Heo, Jaechang Kim, Junhyuk Kwon, Hoyoung Kim, Dongmin Park, Jonghyun Lee, and Jungseul Ok. MMTB: Evaluating terminal agents on multimedia-file tasks. *arXiv preprint arXiv:2605.10966*, 2026. URL https://arxiv.org/abs/2605.10966.

Haobo Hu, Xiangwu Guo, Zhiheng Chen, Difei Gao, Haotian Liu, Libiao Jin, and Qi Mao. CutVerse: A compositional GUI agents benchmark for media post-production editing. *arXiv preprint arXiv:2605.19484*, 2026. URL https://arxiv.org/abs/2605.19484.

Lawrence Jang, Yinheng Li, Dan Zhao, Charles Ding, Justin Lin, Paul Pu Liang, Rogerio Bonatti, and Kazuhito Koishida. VideoWebArena: Evaluating long context multimodal agents with video understanding web tasks. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2410.19100.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can language models resolve real-world GitHub issues? In *International Conference on Learning Representations (ICLR)*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/edac78c3e300629acfe6cbe9ca88fb84-Abstract-Conference.html.

Dai Fei Elmer Ker, Sungeun Eom, Sho Sanami, et al. Phase contrast time-lapse microscopy datasets with automated and manual cell tracking annotations. *Scientific Data*, 5:180237, 2018. doi: 10.1038/sdata.2018.237. URL https://www.nature.com/articles/sdata2018237.

Xiaoxi Li, Wenxiang Jiao, Jiarui Jin, Haoxuan Li, Hao Wang, Shijian Wang, et al. OmniGAIA: Towards native omni-modal AI agents. *arXiv preprint arXiv:2602.22897*, 2026. URL https://arxiv.org/abs/2602.22897.

Yizhi Li, Ge Zhang, Yinghao Ma, Ruibin Yuan, Kang Zhu, Hangyu Guo, et al. OmniBench: Towards the future of universal omni-language models. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 38, 2025. URL https://proceedings.nips.cc/paper_files/paper/2025/hash/2c7c4a12a9dcf7dace9896d08154c705-Abstract-Datasets_and_Benchmarks_Track.html.

Chengwen Liu, Xiaomin Yu, Zhuoyue Chang, Zhe Huang, Shuo Zhang, Heng Lian, et al. Watching, reasoning, and searching: A video deep research benchmark on open web for agentic video reasoning. *arXiv preprint arXiv:2601.06943*, 2026. URL https://arxiv.org/abs/2601.06943.

Lin Long, Yichen He, Wentao Ye, Yiyuan Pan, Yuan Lin, Hang Li, Junbo Zhao, and Wei Li. Seeing, listening, remembering, and reasoning: A multimodal agent with long-term memory. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2508.09736.

Mantas Mazeika, Alice Gatti, Cristina Menghini, Udari Madhushani Sehwag, Shivam Singhal, Yury Orlovskiy, et al. Remote labor index: Measuring AI automation of remote work. *arXiv preprint arXiv:2510.26787*, 2025. URL https://arxiv.org/abs/2510.26787.

Mike A. Merrill, Alexander G. Shaw, Nicholas Carlini, Boxuan Li, Harsh Raj, Ivan Bercovich, et al. Terminal-Bench: Benchmarking agents on hard, realistic tasks in command line interfaces. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/444a3737adaee10d86ad2ef5f74468e6-Abstract-Conference.html.

Samuel Miserendino, Michele Wang, Tejal Patwardhan, and Johannes Heidecke. SWE-Lancer: Can frontier LLMs earn $1 million from real-world freelance software engineering? In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 44412–44450, 2025. URL https://proceedings.mlr.press/v267/miserendino25a.html.

O*NET OnLine. Airfield operations specialists (53-2022.00), 2026a. URL https://www.onetonline.org/link/summary/53-2022.00. Occupational profile; accessed September 17, 2026.

O*NET OnLine. Special effects artists and animators (27-1014.00), 2026b. URL https://www.onetonline.org/link/summary/27-1014.00. Occupational profile; accessed September 13, 2026.

O*NET OnLine. Biological technicians (19-4021.00), 2026c. URL https://www.onetonline.org/link/summary/19-4021.00. Occupational profile; accessed September 13, 2026.

O*NET OnLine. Court, municipal, and license clerks (43-4031.00), 2026d. URL https://www.onetonline.org/link/summary/43-4031.00. Occupational profile; accessed September 13, 2026.

O*NET OnLine. Coaches and scouts (27-2022.00), 2026e. URL https://www.onetonline.org/link/summary/27-2022.00. Occupational profile; accessed September 13, 2026.

O*NET OnLine. Architectural and civil drafters (17-3011.00), 2026f. URL https://www.onetonline.org/link/summary/17-3011.00. Occupational profile; accessed September 13, 2026.

O*NET OnLine. Industrial engineering technologists and technicians (17-3026.00), 2026g. URL https://www.onetonline.org/link/summary/17-3026.00. Occupational profile; accessed September 13, 2026.

O*NET OnLine. Software quality assurance analysts and testers (15-1253.00), 2026h. URL https://www.onetonline.org/link/summary/15-1253.00. Occupational profile; accessed September 13, 2026.

O*NET OnLine. Sound engineering technicians (27-4014.00), 2026i. URL https://www.onetonline.org/link/summary/27-4014.00. Occupational profile; accessed September 13, 2026.

O*NET OnLine. Zoologists and wildlife biologists (19-1023.00), 2026j. URL https://www.onetonline.org/link/summary/19-1023.00. Occupational profile; accessed September 13, 2026.

Tejal Patwardhan, Rachel Dias, Elizabeth Proehl, Grace Kim, Michele Wang, Olivia Watkins, et al. GDPval: Evaluating AI model performance on real-world economically valuable tasks. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/290c2430f91912204f30bbcc990fff1d-Abstract-Conference.html.

Rohith Peddi, Shivvrat Arya, Bharath Challa, Likhitha Pallapothula, Akshay Vyas, Bhavya Gouripeddi, Qifan Zhang, Jikai Wang, Vasundhara Komaragiri, Eric Ragan, Nicholas Ruozzi, Yu Xiang, and Vibhav Gogate. CaptainCook4D: A dataset for understanding errors in procedural activities. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f4a04396c2ed1342a5d8d05e94cb6101-Abstract-Datasets_and_Benchmarks_Track.html.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, et al. MMAU: A massive multi-task audio understanding and reasoning benchmark. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2410.19168.

Tim J. Schoonbeek, Tim Houben, Hans Onvlee, Peter H. N. de With, and Fons van der Sommen. IndustReal: A dataset for procedure step recognition handling execution errors in egocentric videos in an industrial-like setting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4353–4362, 2024. doi: 10.1109/WACV57701.2024.00431. URL https://arxiv.org/abs/2310.17323.

Yiyou Sun, Xinyang Han, Weichen Zhang, et al. Agents' last exam. *arXiv preprint arXiv:2606.05405*, 2026. URL https://arxiv.org/abs/2606.05405.

Yolo Y. Tang, Daiki Shimada, Jiayue Meng, Jing Bi, Pinxin Liu, Yicheng Wang, et al. BVB: Benchmarking agentic video understanding via programmatic reconstruction in Blender. *arXiv preprint arXiv:2609.15478*, 2026. URL https://arxiv.org/abs/2609.15478.

U.S. Bureau of Labor Statistics. May 2023 national occupational employment and wage estimates: United states, 2024. URL https://www.bls.gov/oes/2023/may/oes_nat.htm. Occupational Employment and Wage Statistics; accessed September 26, 2026.

U.S. Bureau of Labor Statistics. Occupational employment and wage statistics: Frequently asked questions, 2026. URL https://www.bls.gov/oes/oes_ques.htm. Survey scope and wage definitions; accessed September 26, 2026.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, et al. OSWorld: Benchmarking multimodal agents for open-ended tasks in real computer environments. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, volume 37, 2024. doi: 10.52202/079017-1650. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/5d413e48f84dc61244b6be550f1cd8f5-Abstract-Datasets_and_Benchmarks_Track.html.

Frank F. Xu, Yufan Song, Boxuan Li, Yuxuan Tang, Kritanjali Jain, Mengxue Bao, et al. TheAgentCompany: Benchmarking LLM agents on consequential real world tasks. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, volume 38, 2025. doi: 10.52202/085713-0315. URL https://papers.nips.cc/paper_files/paper/2025/hash/0d744742f6fac4d1134c019b7cef3c8a-Abstract-Datasets_and_Benchmarks_Track.html.

Tao Yu, Yiming Ding, Shenghua Chai, Minghui Zhang, Zhongtian Luo, Xinming Wang, et al. Omni-DeepSearch: A benchmark for audio-driven omni-modal deep search. *arXiv preprint arXiv:2605.08762*, 2026. URL https://arxiv.org/abs/2605.08762.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, et al. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9556–9567, 2024. URL https://arxiv.org/abs/2311.16502.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, et al. MLVU: Benchmarking multi-task long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13691–13701, 2025. URL https://arxiv.org/abs/2406.04264.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, et al. WebArena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/4410c0711e9154a7a2d26f9b3816d1ef-Abstract-Conference.html.

A Occupational coverage and construction protocol

A.1 Detailed occupational mapping and output coverage

The main-text overview appears in Figure 1. Contributors design and build tasks within the eight predefined domains. Table 3 lists the eight families and their proposed occupational links. Table 4 summarizes the primary deliverables. Identity tracking, measurement, and documentation recur across families.

Table 3: Task families and task counts. Related roles are occupations mapped to at least one task in the family; a parenthetical marks a role that covers only some of the family's tasks.

Task family	Tasks	Related roles
Sports analysis	8	Coaches and scouts (O*NET OnLine, 2026e)
Game-state reconstruction	6	Coaches and scouts (esports tasks); software QA analysts (gameplay proxies) (O*NET OnLine, 2026e;h)
Scientific observation	6	Biological technicians; wildlife biologists (O*NET OnLine, 2026c;j)
Transport and operations	4	Airfield operations specialists (runway task) (O*NET OnLine, 2026a)
Meetings and procedures	5	Municipal clerks; industrial engineering technicians (assembly task) (O*NET OnLine, 2026d;g)
Spatial measurement and tracking	6	Architectural and civil drafters (layout task) (O*NET OnLine, 2026f)
Music and synchronization	3	Sound engineering technicians (synchronization task) (O*NET OnLine, 2026i)
Asset reconstruction	7	3D artists and animators (3D tasks); sound engineering technicians (audio tasks) (O*NET OnLine, 2026b;i)
Total	45	

Table 4: Primary deliverables in the 45-task inventory. Each task is counted once, by its requested output.

Primary deliverable	Tasks	Examples
Structured records and quantitative annotations	37	Event ledgers, lineages, trajectories, spatial records
Reconstructed assets and editable projects	7	Meshes, synthesizer patches, Blender projects, a mixing session, a slide deck
Resulting environment state	1	A recovered robot simulation state
Total	45	

Within asset reconstruction, Slide Deck requires rebuilding an editable deck. Robot Recovery is the only task scored on an environment state and belongs to meetings and procedures because it completes an interrupted checklist. Occupational roles are an activity-level mapping informed by O*NET descriptions. A gameplay history, for example, exercises documentation relevant to software testing. The collection is not weighted by employment or economic output.

Economic context. The mapping links activities to 10 distinct occupations. Using May 2023 national OEWS estimates, we calculate their annualized wage scale as $\sum_o E_o W_o$, where E_o is employment and W_o is the annual mean wage. The total is $66.99 billion across 941,020 jobs (Table 5). Each occupation is counted once, including roles linked to multiple families. Municipal clerks use the broader court, municipal, and license clerks category.

This calculation describes the economic scale of the linked occupations. The tasks cover selected activities within them, so the total does not estimate the value of benchmark tasks, potential automa-

tion revenue, or a share of GDP. OEWS excludes self-employed workers and employer benefit costs; annual wages generally assume 2,080 working hours, so the product is an annualized measure rather than observed payroll (U.S. Bureau of Labor Statistics, 2026).

Table 5: Economic context for the 10 linked occupations. Employment and annual mean wages are May 2023 U.S. OEWS estimates (U.S. Bureau of Labor Statistics, 2024). The last column is their product, expressed in billions of dollars; the total uses unrounded products and counts each occupation once.

Occupation	SOC code	Jobs	Mean annual wage ($)	Annualized wages ($B)
Coaches and Scouts	27-2022	238,980	58,700	14.03
Software Quality Assurance Analysts and Testers	15-1253	203,040	108,460	22.02
Biological Technicians	19-4021	76,990	56,540	4.35
Zoologists and Wildlife Biologists	19-1023	17,100	75,740	1.30
Airfield Operations Specialists	53-2022	18,320	62,140	1.14
Court, Municipal, and License Clerks	43-4031	157,960	48,760	7.70
Industrial Engineering Technologists and Technicians	17-3026	73,020	66,640	4.87
Architectural and Civil Drafters	17-3011	111,070	65,000	7.22
Sound Engineering Technicians	27-4014	14,600	74,100	1.08
Special Effects Artists and Animators	27-1014	29,940	109,630	3.28
Total		941,020		66.99

A.2 Construction and quality-control protocol

Contributors submit the task specification, source materials, output contract, oracle solution, deterministic verifier, development scores, and raw trajectories. Design review checks the proposed work activity and evidence; implementation review checks the executable package and its acceptance rules (Section 2.4). Revisions repeat the affected checks before acceptance.

References may derive from source annotations, logged states, controlled transformations, or target assets. Task and media digests associate each reference, control, and review decision with its version. Oracle acceptance checks that a reference submission receives full reward; manual review checks that the source evidence supports the requested answer and precision. Appendix B gives source-to-deliverable examples and a documented human review.

B Supporting records for benchmark construction

Inherited evidence and added work requirements. Table 6 gives four source-derived examples. Existing annotations supply reference information; the benchmark specifies a bounded work request, transforms or selects the evidence, and implements joint acceptance over the requested output. The source audit reproduced all four frozen reference transformations from the raw source records without importing the task builders. Reusing those labels does not provide a second judgement of the source media.

Human annotation and adjudication. The Balance Beam task documents a blind second pass by a human reviewer who had not created or inspected the draft reference. The reviewer scanned the 139.3-minute source video and recorded 24 candidate routines. Author adjudication retained 23 and excluded one whose mount was not visible. The records distinguish the reviewer's initial timestamps from subsequent source-frame refinement of starts, ends, and dismount takeoffs. The final reference also records school, gymnast, score, and score-availability time. The annotation sheets, adjudication decisions, and status record retain dates and file hashes. All 23 included adjudication rows agree with the seven scored fields in the reference and oracle files.

Table 6: Source-to-deliverable examples. Counts refer to the frozen task, not the full source dataset.

Task / source	Inherited evidence	Requested output and acceptance
Cooking / CaptainCook4D (Peddi et al., 2024)	Procedure recordings, step boundaries, error labels	A 314-step ledger jointly scoring sequence, boundaries, and errors
Assembly / IndustReal v2 (Schoonbeek et al., 2024)	Procedure recordings, transitions, state vectors	47 checkpoints with full successor states, incorrect transitions, and timing
Cell lineage / Ker et al. (Ker et al., 2018)	Microscopy and curated tracks	A transformed sequence; 257 divisions with ancestry and census requirements
Jay behavior / CHIRP (Chan et al., 2026)	Bird recordings, identities, peck annotations	104 individual bouts with boundaries, identity, and per-peck timing

C EVALUATION DETAILS

The score evaluation covers three runs for every task and agent. Behavioral findings were checked across all three runs. Detailed counts, verifier components, and resource measurements refer to a matched diagnostic sample of 225 executions, one per task and agent. Score summaries and variability appear at the end of this appendix.

Web access. Search and fetch tools are disabled by default through harness settings or a denying hook. On 43 tasks, network access is restricted to the model provider's API host. Dota 2 and IndustReal allow internet access; seven of their ten sampled executions installed general-purpose libraries. Astra, Opus 5, and Gemini 3.8 Flash on GSW ran without the web-tool restrictions. Gemini made 46 searches returning content before timing out with score zero; Astra and Opus 5 made no web calls. The eight executions with library installs or returned search content retain their scores. Zeroing the five nonzero scores among them leaves the diagnostic sample's agent ordering unchanged.

Execution and scoring. All reported executions ran on Modal cloud VMs. Native logs record medium reasoning effort for Astra, Opus 5.5, Opus 5, and Gemini 3.8 Flash; DeepSeek uses DSH's default high setting. Opus 5.5 uses Claude Code 2.1.280 and Opus 5 uses 2.1.269. Infrastructure failures were rerun; the reported Gemini Blender Advanced rerun also ended in provider errors and scores zero despite a raw verifier reward of 0.101. Agent timeouts, provider errors, and abnormal exits score zero; successful solve steps retain their verifier rewards. Collection failures after scoring do not change scores. Slide Deck's verifier was corrected during evaluation: Opus 5.5 uses the corrected version and the other sampled agents use the original. Rescoring their saved decks raises each reward by 0.001 to 0.020 without changing any agent's mean at three decimals.

Failure-mode classification. Each rollout receives at most one label. The one rollout ended by a provider fault is excluded. A rollout has no valid deliverable (N) if the agent's own timeout or exit interrupted it, or if the verifier found the deliverable missing, unreadable, schema-invalid, or empty. Otherwise a task-specific mapping turns verifier fields into losses, each a share of the task's reference items or of reward weight: low coverage (C, reference items never found at the loosest level the verifier reports), wrong identity (I, an item found but credited to the wrong individual), imprecise localization (L, time or position outside tolerance), and wrong values (V, wrong label, count, parameter, or state); O covers the one task scored on an environment state. On the lineage census, whose gates zero every rollout's reward, the whole loss counts as coverage when division F1 is below 0.5 and as identity otherwise. Elsewhere the label is the largest loss; the one full score gets none. Four rules settle ambiguous cases: items never listed count as coverage loss; when the loosest match joins fields from several categories, the unmatched share is split equally among them; when only per-field marginals are reported, a category's loss is its largest single-field loss; and exact ties resolve C, then I, L, V, O. Ten rollouts change label under an alternative split or tie order. Recognized and localized shares (Section 3.2) are means over tasks whose verifier, or a relaxed replay, reports both, using valid deliverables only.

Audio processing. A rollout processed audio if any command, code, or written file of its main session or sub-agents uses an audio library to transcribe, load, or analyze sound, runs ffmpeg to

decode or analyze an audio stream, calls sox, or opens an audio file with a native read tool; metadata probes alone do not count. Of 19 tasks with an audio track, 11 score something observable only in the audio. The one audio-scored rollout that never processed it is Astra's on a synthesizer-session reconstruction; the other nine unprocessed tracks carry broadcast commentary on two sports tasks.

Outer-container intervention. For the five Gemini 3.8 Flash submissions whose top-level container violated the task's schema, we added only the task-specified outer object key, kept every prediction, and rescored with the same task verifier (Table 7). This offline diagnostic leaves aggregate scores unchanged; the table reports verifier rewards on the saved files.

Task	Original	Wrapped	Residual result
Badminton	0	0	No recognized checkpoints
OpenTTGames	0	0.000272	Joint timing and semantic errors
Piano Duo	0	0	Zero true positives
Siberian Jay	0	0	Zero true positives
LoL	0	0.1532	17 matched events

Table 7: Outer-container intervention on five Gemini 3.8 Flash failures. Predictions are unchanged; only the task-required outer object key is added. Both conditions use the same task verifier; values are verifier rewards, and the LoL rollout reached its time limit, so it scores zero in both conditions under the main scoring rule.

Three-run score aggregation. Every one of the 45 tasks has three scores for each of the five agents. Records are joined by task identifier and agent, using final scores under the interruption policy. Each task contributes its three-run mean equally to the benchmark and family averages. Model, harness, task and verifier version, reasoning setting, and resource budget are fixed for each task and agent across runs; only the run or seed differs.

Agent	Run 1	Run 2	Run 3	Mean	Mean within-task SD
GPT-6-Astra	0.279	0.281	0.288	0.283	0.025
Claude Opus 5.5	0.261	0.282	0.289	0.278	0.025
Claude Opus 5	0.087	0.109	0.122	0.106	0.025
Gemini 3.8 Flash	0.039	0.060	0.071	0.057	0.020
DeepSeek 4.1 Flash	0.037	0.061	0.075	0.058	0.020

Table 8: Aggregate scores for all three runs over 45 tasks. The final column averages the per-task sample standard deviation across three runs, using denominator $n - 1 = 2$.

Resource comparison. Figure 5 pairs scores and resource measurements from the same diagnostic sample. DeepSeek's aggregate cost covers 44 sampled executions, excluding Robot Recovery; the other cost values are framework estimates. Costs exclude virtual machines; execution times exclude setup and verification.

D BEHAVIOR AND DELIVERY MEASUREMENT

Behavioral findings were checked across all three runs. Numerical trace summaries use the matched diagnostic sample defined in Appendix C.

D.1 OPERATION COUNTING AND CODE EXECUTION

Decision units. A decision is one Codex exec call, Claude Code assistant message, Antigravity step, or DSH step. Antigravity records one call per step. Astra explicitly launches calls concurrently in 93 of 3,325 multi-operation decisions. Claude Code and DSH may execute one message's tool calls concurrently. Figure 6 includes sub-agents and groups operations by their order: commands before views, views only, views before commands, or commands and file operations only. Other mixes contribute under 0.5% and count toward totals. Command-text patterns identify order, without establishing that a command produced the images viewed afterward.

Code execution and action rates. Inline code is detected through shell heredocs and `-c` or `-e` arguments; saved-script execution through commands that run a script file. These heuristics can overlap and need not cover every shell call. Figure 7a reports medians over 45 executions per agent. Delegation shares in Figure 7b are conditional on delegation. Action rates include main-session and sub-agent actions divided by elapsed execution time. Opus 5 has the lowest median rate in each of the eight families.

D.2 DELIVERY AND LOCAL VERIFICATION

Output detection. File presence is measured from the collected output directories joined to all 225 score records. First-write detection instead matches output-path references in tool-call text. It can miss writes within separately executed scripts and includes two commands that only create directories. Detected times therefore need not be actual first-write times. Explicit echo or printf placeholders are excluded. Writes are detected in 43, 37, 34, 32, and 33 executions for Astra, Opus 5.5, Opus 5, Gemini, and DeepSeek, respectively. Among 38 detections in the first 5% of an execution, 37 create empty skeletons or templates.

Copied examples. A prompt example counts as copied when all numeric and non-placeholder text fields appear unchanged in a submitted object and the row is absent from the reference. Unparseable submissions are omitted. Two tasks are excluded because their examples match reference events: one within 1 s and another with four reference rows. Gemini reproduces 12 of 50 eligible rows across six tasks; other agents reproduce none of 28 to 52 eligible rows each.

Elapsed-time statements. The analysis codes 44 DeepSeek executions. Of these, 23 contain 63 elapsed-time statements made without checking a clock; 62 overestimate duration, with a median ratio of 3.3 to actual elapsed time. Other agents were not coded for this measure. DSH provides no elapsed-time signal to the agent.