AGENTIC-WORKBENCH: EVALUATING MULTIMODAL AGENTS ON COMPLEX VISUAL WORKFLOWS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal work often requires agents to acquire visual evidence, use it to guide successive actions, and verify the resulting artifacts or service states. Existing work-agent evaluations leave open how reliably visual information can support this full workflow. We introduce **Agentic-WorkBench**, a benchmark of 115 tasks across eleven work domains in controlled, stateful environments. The benchmark targets *workflow-level visual intelligence*, organized into visual grounding, sensemaking, and realization, with observations that may become available only during execution. Across nine model configurations, the strongest achieves 64.0% graded completion, solves 41.7% of tasks within three trials, and succeeds in all three on only 22.6%. The lowest success rates concentrate on visual sensemaking demands, and success varies substantially across work domains and objectives. Harness effects vary across models, and greater resource use does not consistently accompany higher success. These results identify a gap between partial progress and repeatable completion, motivating agents that maintain visual state and use feedback throughout practical work.

1 INTRODUCTION

AI agents are increasingly expected to complete workflows on behalf of users rather than answer isolated questions (OpenAI, 2026c; OpenClaw, 2026). Real-world work is inherently multimodal: users express goals through text and visual references, agents encounter information in documents and interfaces, and deliverables often combine written content with visual structure. Visual information shapes both how this work should proceed and what constitutes a satisfactory outcome. An agent must translate its understanding of visual materials into decisions and actions, acquire and interpret new visual information during execution, and verify that the resulting work satisfies the task's requirements. These capabilities are interdependent: correctly interpreting a visual constraint is useful only if subsequent actions respect it, while inspecting an intermediate result may reveal the need to revise earlier decisions. This motivates evaluating whether agents can connect visual understanding, action, and verification throughout the workflow.

Existing work-agent benchmarks have expanded from packaged professional tasks with concrete deliverables (Patwardhan et al., 2025; Mazeika et al., 2025; Li et al., 2026b) to broader workspace contexts (Tang et al., 2026). Recent benchmarks further move toward interactive workplaces, where agents complete work through tool-mediated interaction with stateful services (Shepard & Salimans, 2026; Li et al., 2026a). These advances capture increasingly rich work contexts, but leave a key evaluation gap: whether agents can use visual information as part of the work itself, carrying it from task understanding through execution to the final outcome. The presence of visual inputs or deliverables alone does not resolve this question. What remains underexplored is whether visual information actually shapes what the agent seeks, how it acts, how it adapts during execution, and what it ultimately produces.

We refer to this capability as **Workflow-Level Visual Intelligence** (Fig. 1): the ability to integrate visual understanding, reasoning, interaction, and creation to complete work. We organize it into three connected stages. (1) *Visual Grounding* establishes the task context by interpreting visual materials and identifying relevant entities, relations, and constraints. (2) *Visual Sensemaking* guides execution by acquiring, inspecting, and reconciling visual information, then using it to inform decisions, tool use, and service interactions. It supports closed-loop visual agency: actions lead to new

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.

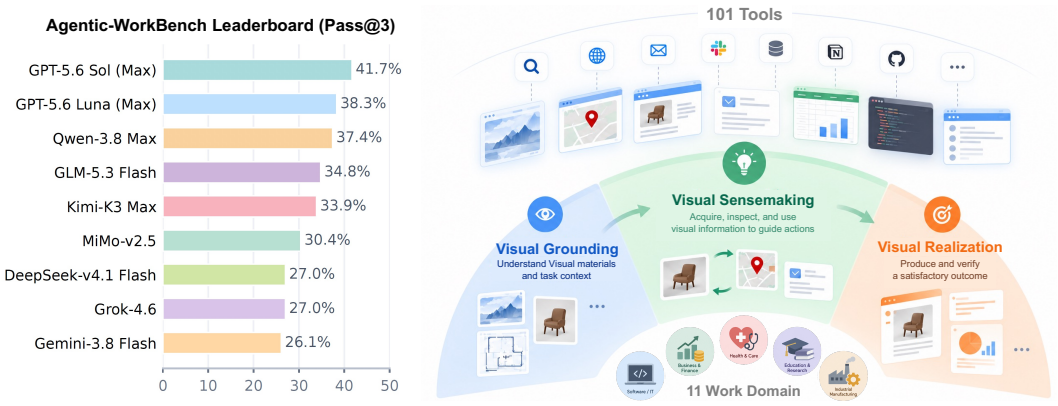

Figure 1: **Workflow-level visual intelligence.** The three connected stages, Visual Grounding, Visual Sensemaking, and Visual Realization, together with the 101 tools and 11 work domains exercised by Agentic-WorkBench.

visual observations, which can in turn change what the agent does next. (3) *Visual Realization* involves producing and verifying a final artifact or service state that satisfies the task's substantive and visual requirements. These stages are connected rather than independent: constraints identified during grounding must inform execution and remain satisfied in the final outcome, while intermediate or final inspections may prompt further interpretation and action.

To evaluate this capability, we introduce **Agentic-WorkBench**, a benchmark for workflow-level visual intelligence in complex work. Agentic-WorkBench contains 115 tasks across eleven work domains and places agents in controlled, stateful environments exposing 101 tools across services including Web, CRM, databases, Slack, email, and other business or personal applications. Unlike task-bounded evaluations in which most relevant information is packaged upfront, task-relevant information in Agentic-WorkBench may be distributed across services or become available only through execution. Agents therefore need to combine multimodal understanding with information acquisition, tool use, service interaction, intermediate decision-making, and final artifact production. Visual requirements are systematically embedded across the full agent loop—input, execution, and output, with tasks designed to require visual information to be understood, used during execution, and reflected in the final outcome.

We evaluate 9 agent models on Agentic-WorkBench and examine how visual demands, work contexts, and execution harnesses affect task outcomes. The strongest configuration reaches 64.0% graded completion but succeeds on all three trials for only 22.6% of tasks. The lowest success rates occur on several visual sensemaking demands, with information-need recognition, visual state tracking, and feedback-conditioned adaptation remaining particularly difficult. Success also varies substantially across work domains and objectives, while harness improvements depend on the model being evaluated. These findings reveal a gap between partial progress on visual work and reliably delivering the requested outcome.

Our contributions are threefold:

- We formulate **Workflow-Level Visual Intelligence** as a distinct evaluation target for work agents, covering visual grounding at input, visual sensemaking during execution, and visual realization at output, with closed-loop visual agency arising within execution.
- We introduce **Agentic-WorkBench**, a benchmark of 115 tasks across eleven work domains in stateful environments with 101 tools spanning Web, CRM, databases, Slack, email, and other services, with systematic visual requirements throughout the full agent loop.
- We provide a diagnostic evaluation of frontier multimodal agents, measuring whether visual competence can be maintained and effectively used across long-horizon, tool-mediated work workflows rather than only in isolated perception or generation tasks.

Table 1: Comparison of representative work-centric benchmarks across interaction modality and visual evaluation scope. *Full-workflow visual dependency* links visual requirements across grounding, execution, and realization in full workflow. ✓, △, and × indicate systematic, partial, and no primary coverage.

Benchmark	Setting family	Interaction	Work domains	Tasks	Multimodal I/O	Full-workflow visual dependency
GDPval (Patwardhan et al., 2025)	Task package	Files	9	1,320	✓	×
RLI (Mazeika et al., 2025)	Task package	Files	–	240	△	×
JobBench (Li et al., 2026b)	Task package	Files	10	130	△	×
OneMillion-Bench (Yang et al., 2026)	Task package	Files	5	400	×	×
SWE-Lancer (Miserendino et al., 2025)	Task package	Files	1	1,488	×	×
APEX-Agents (Vidgen et al., 2026)	Persistent workspace	Tool/API	3	480	△	×
Workspace-Bench (Tang et al., 2026)	Persistent workspace	Tool/API	5	388	△	×
VisualWebArena (Koh et al., 2024)	GUI environment	GUI	–	910	✓	×
OSWorld 2.0 (Yuan et al., 2026)	GUI environment	GUI	–	108	✓	×
TheAgentCompany (Xu et al., 2024)	Simulated services	Tool/API	–	175	×	×
WorkBench (Styles et al., 2024)	Simulated services	Tool/API	–	690	×	×
AutomationBench (Shepard & Salimans, 2026)	Simulated services	Tool/API	6	600+	×	×
ClawsBench (Li et al., 2026a)	Simulated services	Tool/API	–	44	×	×
Claw-Eval (Ye et al., 2026)	Simulated services	Tool/API	–	300	✓	×
WildClawBench (Ding et al., 2026)	Native runtime	Tool/API	–	60	✓	×
Agentic-WorkBench	Simulated services	Tool/API	11	115	✓	✓

2 Related Work

2.1 Multimodal Agents

Large language models have evolved from text generation systems into agents that can reason over context, call tools, and execute multi-step actions (OpenAI, 2022; 2023; Anthropic, 2023). In parallel, large vision-language models extend these capabilities to visual information, enabling models to interpret images, documents, interfaces, and other visual media (Liu et al., 2023; Bai et al., 2023; OpenAI et al., 2024). This progress has motivated benchmarks that evaluate multimodal agents through visual understanding, planning, and interaction across embodied, GUI, and other visual environments (Liu et al., 2024; Zhang et al., 2024; Ashraf et al., 2025). Related work has also examined multimodal agents in GUI and computer-use settings (Xie et al., 2024; Koh et al., 2024), where agents perceive rendered interface states and act over web or desktop applications. In these benchmarks, vision primarily serves as part of the interaction interface. By contrast, Agentic-WorkBench treats visual information as task-relevant content within workflows where relevant visual evidence may be absent initially, acquired through service interaction, and subsequently used to inform downstream visual or non-visual actions and outcomes.

2.2 Work-centric Benchmarks

Work-centric benchmarks have expanded from task-bounded professional work to persistent workspaces and interactive workplace environments (Tab. 1). Task-bounded evaluations provide instructions, supporting materials, and expected deliverables, and assess professional work through domain-specific criteria and work-product quality (Patwardhan et al., 2025; Mazeika et al., 2025; Li et al., 2026b; Yang et al., 2026; Hu et al., 2026; Miserendino et al., 2025). Workspace-oriented benchmarks introduce persistent context that agents must search and reconcile across files and dependencies (Vidgen et al., 2026; Tang et al., 2026). More recent environment-oriented benchmarks place agents in stateful services and require multi-step coordination across workplace applications (Xu et al., 2024; Styles et al., 2024; Shepard & Salimans, 2026; Li et al., 2026a), but remain primarily text- and API-centric in how task information is represented and accessed.

Recent benchmarks also introduce multimodal tasks into long-horizon agent evaluation. Claw-Eval (Ye et al., 2026) combines multimodal tasks with evaluation of task completion, safety, robustness, and execution trajectories, while WildClawBench (Ding et al., 2026) emphasizes multimodal, long-horizon tasks executed in native agent runtimes with real tools. Their evaluation focus is not specifically on how visual information is carried across different stages of a work workflow. Agentic-WorkBench instead makes this cross-stage visual dependency the central evaluation target, examining how visual information grounds the task, guides execution, and constrains the final outcome.

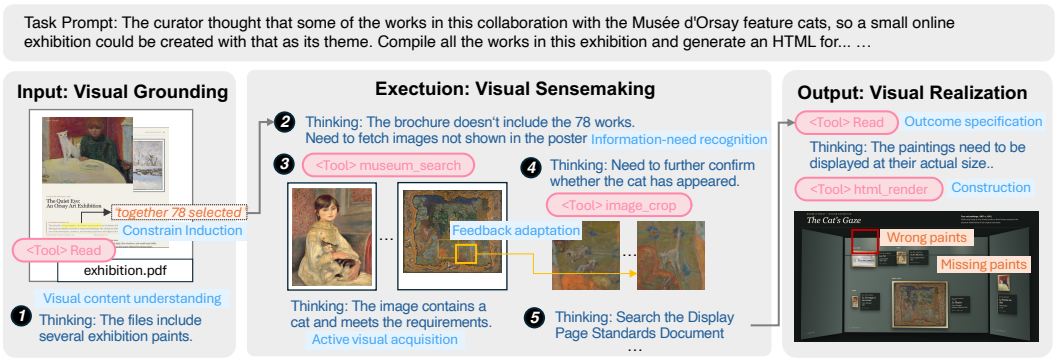

Figure 2: Example workflow spanning the three stages of workflow-level visual intelligence.

3 AGENTIC-WORKBENCH

Agentic-WorkBench evaluates visual intelligence through the work an agent completes. Each task places the agent in a stateful environment where visual information informs decisions, actions reveal or alter observations, and success is determined by the resulting artifact or service state. This design makes the connection between visual understanding and consequential action the central object of evaluation.

3.1 WORKFLOW-LEVEL VISUAL INTELLIGENCE

We define *workflow-level visual intelligence* as the ability to acquire, interpret, and act on visual information throughout a task, while preserving its implications in the final outcome. The relevant information may be present in the initial materials, distributed across services, or revealed by an intermediate action. Such workflows may require the agent to decide both what to do with an observation and which observation to obtain next.

Visual Grounding. The agent establishes the visual task state by identifying relevant entities, interpreting their attributes and relationships, and deriving constraints from the user's request and visual materials. Grounding connects a work goal to the objects and properties on which subsequent actions depend.

Visual Sensemaking. The agent uses visual evidence to guide execution: it acquires missing views, reconciles sources, tracks changes, and revises actions in response to feedback. This stage closes the observation–action loop, allowing newly available visual information to change the course of the workflow.

Visual Realization. The agent produces and verifies an artifact or service state that satisfies the task's semantic and visual requirements. Depending on the goal, this may involve selecting an item, modifying an interface, or constructing a visual deliverable. The final result must preserve the constraints established during grounding and refined through execution.

Figure 2 illustrates how visual grounding, sensemaking, and realization interact within a single workflow, with incomplete visual evidence during execution propagating to missing or incorrect content in the final artifact. These stages recur and interact: inspecting an intermediate artifact can reveal a violated constraint and trigger further grounding or revision. We characterize their demands through eleven visual functions, defined in App. C, and evaluate success at the workflow level.

3.2 BENCHMARK CONSTRUCTION

We construct Agentic-WorkBench through four stages: work-domain definition, workflow design and annotation, environment and grader implementation, and quality control.

Work-domain definition. We organize tasks by the work context in which the user goal arises, using the International Standard Industrial Classification of All Economic Activities (ISIC Rev. 5)

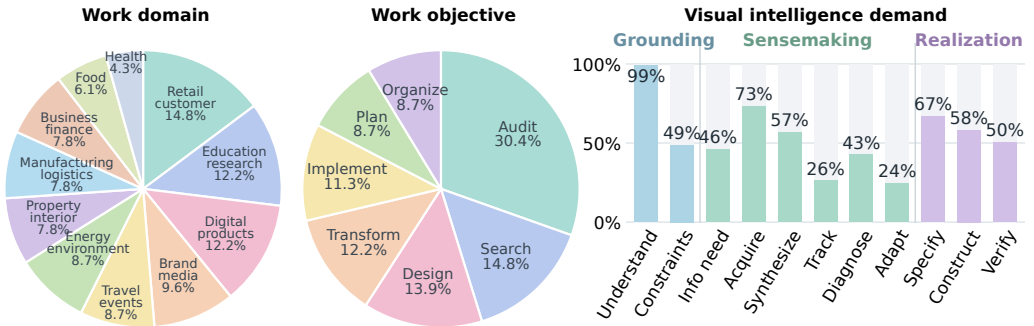

Figure 3: Coverage of the 115 tasks by work domain, work objective, and visual function. Each task has one primary domain and objective; visual demands can overlap. Colored bars show the share of tasks for which each function is central, against a pale 100% reference. Category definitions and annotation details are provided in App. A and App. C.

(United Nations, 2024) as an external reference and consolidating related economic activities into eleven benchmark-specific work domains. The consolidation preserves distinctions that are meaningful for agentic work while avoiding overly sparse categories at the benchmark scale. Each task is assigned one primary domain according to its work context. Domain experts additionally annotate one of seven primary work objectives according to the intended outcome of the user request, such as selecting, reconstructing, or repairing. Definitions and annotation criteria are provided in App. A.

Workflow design and visual annotation. For each task, we specify the user request, constraints, initial materials, available tools, environment state, expected deliverables, and grading criteria. Every workflow is designed to exercise all three stages of workflow-level visual intelligence: visual grounding, visual sensemaking, and visual realization. Relevant visual information may be available initially, distributed across services, or revealed through interaction with the environment, requiring the agent to maintain and use visual information across multiple steps of execution. After construction, domain experts annotate each task with the eleven visual functions defined in App. C; multiple functions may apply to a single task.

Environment and grader implementation. We instantiate each workflow in a controlled, stateful environment containing the required files, webpages, service records, visual materials, and editable artifacts. Agent actions can retrieve information, update service state, or modify intermediate artifacts, allowing the environment to reflect the consequences of execution. Task-specific graders evaluate completion through observable evidence including backend service state, structured values, artifact properties, and visual quality.

Quality control. We initially construct 245 candidate tasks and iteratively review them for work realism, executability, three-stage visual coverage, and outcome verifiability. We further check that required information is accessible through the provided tools and that grading criteria correspond to the requested outcome rather than incidental execution details. Visual materials are additionally reviewed for usability and provenance, while duplicate, unstable, or underspecified tasks are revised or removed. This process yields the final benchmark of 115 tasks.

3.3 Benchmark Coverage

Agentic-WorkBench contains 115 tasks spanning eleven work domains and seven work objectives (Fig. 3). Tasks combine visual materials with documents, structured records, webpages, service states, and editable artifacts, requiring agents to carry information and constraints across multiple representations. Each task exposes 5.22 tools on average, with up to 14, reflecting the multi-step and service-mediated nature of the workflows.

The benchmark places particular emphasis on visual demands that arise during execution. Among the 115 tasks, 84 centrally require active visual acquisition and 65 require cross-source visual synthesis. Visual state tracking is central to 30 tasks, while 28 require agents to adapt subsequent actions based on visual feedback. Because these functions can co-occur within a task, the benchmark covers workflows in which agents must not only interpret available visual materials, but also seek new evidence, integrate it across sources, track changes, and respond to the consequences of their actions.

We further characterize workload variation using expert annotations of constraint density, evidence and service dependencies, output integration, and conditional workflow structure. These annotations define a fixed Easy subset of 38 tasks and Hard subset of 77 tasks prior to model evaluation, enabling stratified analysis without using model performance to determine task difficulty.

4 EXPERIMENTS

We evaluate agents on the proposed Agentic-WorkBench to examine whether current agents can complete visual work reliably and which aspects of their execution remain limiting. We first compare model performance, then examine the effect of the execution harness before analyzing visual demands, work contexts, and resource use.

4.1 EXPERIMENT SETUP

Models. We evaluate nine multimodal model configurations: GPT-5.6 Sol and Luna (OpenAI, 2026b;a), Kimi-K3 (Team et al., 2026b), Qwen-3.8 Max (QwenTeam, 2026), GLM-5.3 Flash (Z.ai, 2026), MiMo-v2.5 (Team et al., 2026a), DeepSeek-v4.1 Flash (DeepSeek-AI et al., 2026), Grok-4.6 (xAI, 2026), and Gemini-3.8 Flash (Google, 2026). Each model is evaluated in its canonical execution configuration for the main comparison. For DeepSeek, GLM 5.3, and Kimi, we additionally compare OpenClaw (OpenClaw, 2026), OpenCode (Anomaly, 2026), and Hermes (Nous Research, 2026) against the corresponding canonical Base configuration. These harnesses provide alternative tool-routing, session, and execution protocols while the task pool, model, and trial budget remain fixed.

Metrics. Each configuration is run for three trials per task, with a 100-turn limit and a 7,200-second timeout per trial. All comparisons use the same task pool and trial budget; App. E provides the aggregation details. We measure task success using the binary outcomes returned by the task-specific graders. **Pass@1** is success on the first budgeted trial (one sample per task), **Pass@3** is the fraction of tasks passed at least once in three trials, and **Pass^3** is the fraction passed in all three trials. These metrics characterize single-trial success, success with retries, and repeatability, respectively. Alongside these binary measures, we report **Completion (Comp.)**, a continuous measure of how much of the requested work the agent completes under the task-specific rubric. We average it over the three trials for each task and then equally across tasks. Missing or errored trials count as failures and contribute zero completion; all metrics retain the full evaluation budget.

4.2 TASK COMPLETION AND RELIABILITY

Current agents struggle with workflow-level visual intelligence. All nine configurations find the 115-task pool challenging. GPT-5.6 Sol is the strongest configuration, yet reaches only 41.7% Pass@3; its Pass@3 drops from 57.9% on Easy tasks to 33.8% on Hard tasks. The decline shows that longer visual workflows amplify difficulty even for the strongest current agent.

Retries improve coverage without stable correctness. Qwen achieves higher Pass@3 than Kimi (37.4% versus 33.9%) but lower Pass^3 (12.2% versus 18.3%). Pass@3 records whether a task is solved at least once, whereas Pass^3 requires success in all three trials. The broad gap between these metrics shows that current models can produce a successful trajectory without reproducing it reliably: resampling increases coverage, but does not necessarily improve the reliability of the underlying policy.

Agents need more efficient tool use within the interaction budget. DeepSeek and Gemini often exhaust the 100-turn budget before completing a task; their traces commonly contain repeated tool calls or failed tool calls. Expanding the budget to 300 turns raises Pass@3 from 27.0% to 33.9%

Table 2: Performance on 115 tasks by expert-annotated workload level. All metrics are percentages; bars share a 0–100 scale. Bold marks the best values.

	GPT-5.6 Sol (Max)	Kimi-K3 Max	Qwen-3.8 Max	GLM-5.3 Flash	GPT-5.6 Luna (Max)	MiMo-v2.5	DeepSeek-v4.1 Flash	Grok-4.6	Gemini-3.8 Flash
Easy (38)									
Comp.	**75.3**	70.2	68.5	68.3	63.8	57.4	55.4	50.6	46.3
Pass@1	44.7	**47.4**	39.5	36.8	42.1	34.2	23.7	23.7	28.9
Pass@3	**57.9**	52.6	**57.9**	50.0	55.3	39.5	44.7	44.7	47.4
Pass^3	**39.5**	31.6	23.7	26.3	21.1	13.2	15.8	21.1	2.6
Hard (77)									
Comp.	**58.4**	53.3	51.5	51.0	50.3	49.8	42.3	41.2	31.7
Pass@1	**24.7**	18.2	13.0	18.2	16.9	18.2	13.0	9.1	10.4
Pass@3	**33.8**	24.7	27.3	27.3	29.9	26.0	18.2	18.2	15.6
Pass^3	**14.3**	11.7	6.5	9.1	10.4	7.8	6.5	3.9	3.9
Overall (115)									
Comp.	**64.0**	58.9	57.1	56.7	54.8	52.3	46.6	44.3	36.5
Pass@1	**31.3**	27.8	21.7	24.3	25.2	23.5	16.5	13.9	16.5
Pass@3	**41.7**	33.9	37.4	34.8	38.3	30.4	27.0	27.0	26.1
Pass^3	**22.6**	18.3	12.2	14.8	13.9	9.6	9.6	9.6	3.5

for DeepSeek and from 26.1% to 31.9% for Gemini, but both remain below Sol's canonical 41.7%. App. F details these reruns. The additional budget recovers some unfinished trajectories, while the remaining failures indicate limitations beyond max-turn termination in the models' ability to interpret visual evidence and complete the workflow.

4.3 Sensitivity to the Execution Harness

Fig. 4 compares the alternative harnesses with each model's Base.

Different harness designs lead to different execution effects. Among the evaluated harnesses, OpenCode provides the most consistent gains, improving Pass@3 across all three model families, including a +13.9 percentage-point gain for DeepSeek. Its design provides more explicit tool routing, guided visual-media inspection, and structured session organization. In representative traces, these mechanisms reduce redundant exploration and move the agent earlier toward artifact-producing actions, suggesting that harness design can improve how effectively visual evidence is translated into workflow progress.

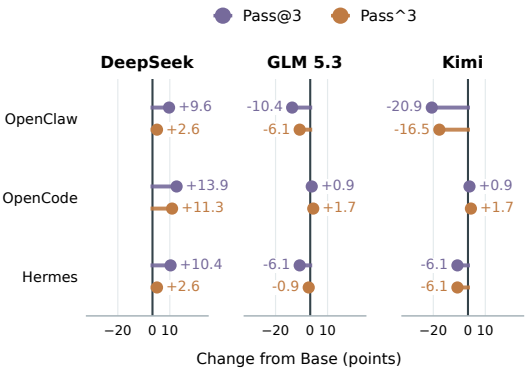

Figure 4: Points show changes from Base in Pass@3 (purple) and Pass^3 (orange), measured in percentage points.

Harness effects reflect model–protocol compatibility. OpenClaw and Hermes improve DeepSeek but reduce both metrics for GLM and Kimi. Representative traces show DeepSeek benefiting from the additional execution structure, whereas GLM and Kimi more often enter long exploratory loops, repeated tool calls, or failed calls before producing a final artifact. The same harness can therefore change the model's behavior in opposite directions, making the model–harness pair the relevant unit of comparison.

4.4 Performance Across Visual Demands

Fig. 5 groups tasks by the eleven visual functions introduced in Sec. 3.1. Each group contains tasks where the function is central, according to the annotations in App. C. Each group reports whole-task success under a central visual demand, and aggregate values below average over the nine canonical configurations.

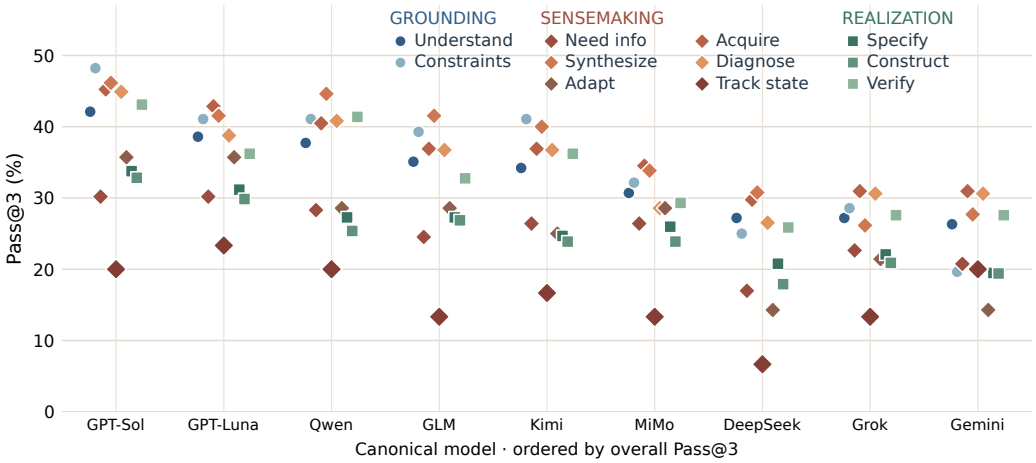

Figure 5: Markers show Pass@3 by central visual demand; shapes denote Grounding (circles),

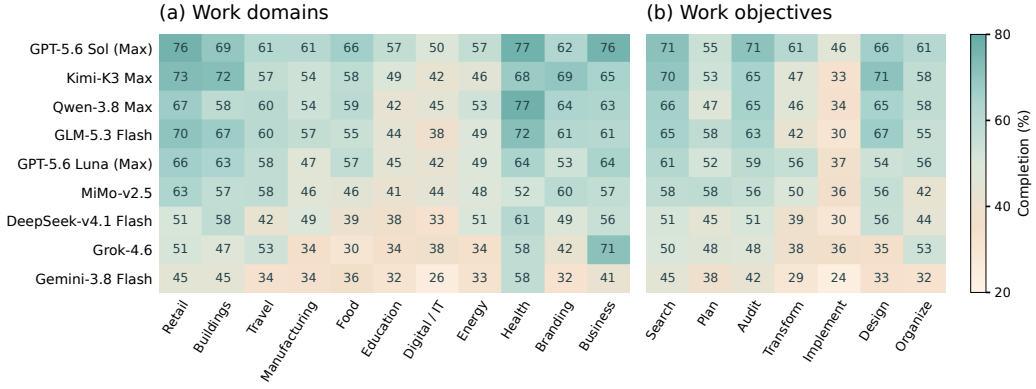

Figure 6: Completion is compared by work domain (left) and objective (right).

The lowest success rates concentrate on Sensemaking-related demands. Visual state tracking is particularly difficult, with Pass3 of 16.3%, while feedback-conditioned adaptation also remains low at 25.8%. In contrast, tasks centered on acquiring or synthesizing visual evidence are solved more often. This pattern suggests that the harder part of Sensemaking is not simply obtaining visual information, but maintaining it across actions and revising subsequent decisions as the visual state evolves.

Model rankings vary by visual demand. GPT-5.6 Sol leads or ties on ten of the eleven visual functions, but Luna has the highest Pass@3 on visual state tracking (23.3%, compared with 20.0% for Sol and Qwen). The close middle group also has distinct profiles: Qwen reaches 40.5% on active acquisition and 44.6% on cross-source synthesis, whereas GLM reaches only 13.3% on state tracking. Overall rankings therefore do not determine which visual demands a model handles reliably.

4.5 Performance Across Work Domains and Objectives

Fig. 6 examines the same configurations by work domain and objective. These two views distinguish the requested operation from the context in which it must be performed.

Work-domain differences identify concentrated difficulty. Pass@3 ranges from 60.8% on Retail and Customer Service to 10.3% on Digital Products and IT and 11.1% on Manufacturing and Logistics. Seven of the nine manufacturing tasks are never passed by any canonical configuration, and their mean Pass@3 is only 11.1%, compared with 54.7% for the other 26 audit tasks. The domain view therefore identifies where current agents encounter concentrated failure cases and complements the objective analysis below.

Work objective exposes a shared end-to-end bottleneck. Implementation and repair tasks have the lowest mean Comp. (34.0%) and Pass@3 (6.8%) among the seven objectives; ten of the 13 tasks

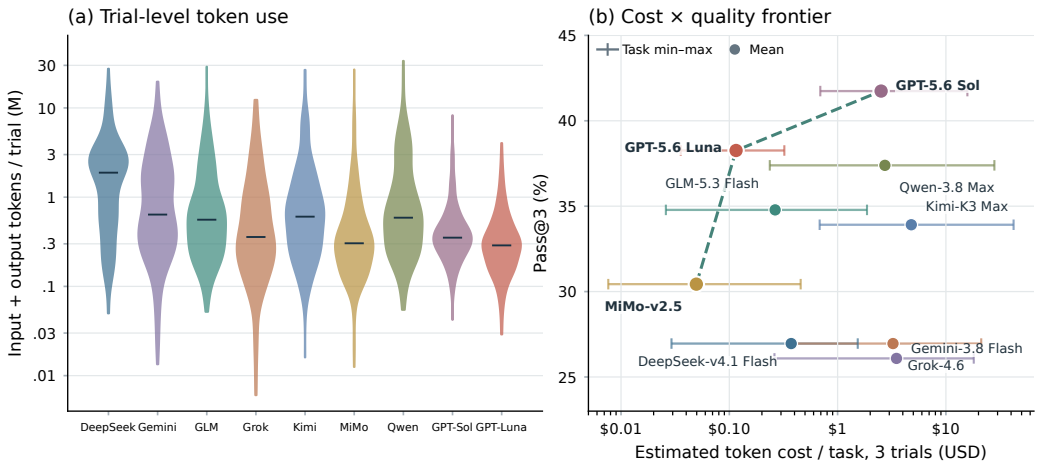

Figure 7: (a) Processed tokens per trial. (b) Pass@3 versus estimated token cost for three trials; the dashed line marks the observed frontier.

are never passed, and no configuration passes any of them in all three trials. These tasks couple visual feedback, state-changing actions, artifact modification, and final verification. The objective analysis therefore points to a broader failure in completing observation–action–verification loops, rather than to a weakness confined to one interface type.

4.6 Resource Use and Task Success

Fig. 7 compares processed tokens per trial with the estimated token cost of three attempts at a task. Costs use public token rates and tasks with complete resource records, while Pass@3 retains all 115 tasks; App. B gives pricing and coverage.

Across models, greater token use does not track higher success. Token use varies widely within each configuration, with upper tails extending to millions of processed tokens per trial. Across the same task pool, however, DeepSeek has the highest typical consumption, whereas Sol and Luna process substantially fewer tokens and achieve the two highest Pass@3 values. The resource burden of a workflow thus depends strongly on the configuration executing it. High token consumption can accompany limited task success, making the quality of execution a separate consideration from its length.

Lower-cost configurations can outperform more expensive alternatives. Luna achieves higher Pass@3 than Qwen, Kimi, and GLM at lower mean estimated cost, while MiMo outperforms the more expensive DeepSeek, Gemini, and Grok. Sol attains the highest success rate but at substantially greater cost than Luna. The resulting frontier presents a meaningful trade-off between expense and success, rather than a uniform benefit from selecting a more costly configuration. Alongside the reliability results, these comparisons motivate assessing visual work agents by whether they complete a task, how consistently they do so, and the resources required.

5 Conclusion

We introduced Agentic-WorkBench, a 115-task benchmark for workflow-level visual intelligence spanning grounding, sensemaking, and realization. Its stateful workflows require agents to acquire visual evidence, use it to guide actions, and produce verifiable outcomes. Across nine model configurations, substantial partial completion coexists with limited repeatability: the strongest passes all three trials on only 22.6% of tasks. Several of the lowest success rates arise on visual sensemaking demands, particularly information-need recognition, visual state tracking, and feedback adaptation. Success also varies sharply across work domains and objectives, while harness effects vary by model and greater resource use does not consistently accompany higher success. These findings motivate agents that maintain visual state and use feedback throughout execution. Agentic-WorkBench provides a testbed for progress toward this sustained use of visual information in practical work.

AI USE STATEMENT

We used AI tools for language polishing and for generating statistical plots. The author team checked the data, analyses, figures, and final manuscript and retains responsibility for all claims and content.

ETHICS STATEMENT

Agentic-WorkBench evaluates agents in controlled environments with simulated services and task-specific visual materials. Its outcomes measure performance within these environments and do not establish readiness for unsupervised use in clinical, financial, safety-critical, or other consequential settings. Broader deployment would require validation of the relevant services and data, along with appropriate human oversight. The benchmark's work-domain coverage also does not represent all occupations, populations, or cultural contexts.

REPRODUCIBILITY STATEMENT

Section 3.2 describes task construction and verification. Appendices A, C, and D define the work categories, visual-demand annotations, and human feasibility check. Appendix E specifies trial accounting, metric aggregation, and uncertainty estimation, and Appendix B documents the resource measurements and pricing assumptions. The main comparisons use a common 115-task pool and a fixed three-trial budget; additional turn-budget results are reported separately in Appendix F.

REFERENCES

Anomaly. Opencode, 2026. URL https://github.com/anomalyco/opencode.

Anthropic. Claude 2, 2023. URL https://www.anthropic.com/news/claude-2.

Tajamul Ashraf, Amal Saqib, Hana' Khalief Ghani, Muhra AlMahri, Yuhao Li, Noor Ahsan, Umair Nawaz, Jean Lahoud, Hisham Cholakkal, Mubarak Shah, Philip H. S. Torr, Fahad Shahbaz Khan, Rao Muhammad Anwer, and Salman H. Khan. Agent-x: Evaluating deep multimodal reasoning in vision-centric agentic tasks. *ArXiv*, abs/2505.24876, 2025. URL https://api.semanticscholar.org/CorpusID:279070382.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *ArXiv*, abs/2308.12966, 2023. URL https://api.semanticscholar.org/CorpusID:263875678.

DeepSeek-AI, Anyi Xu, B. Li, Bangcai Lin, Bing Xue, BingCheng Xian, Bingzheng Xu, Bochao Wu, Bowei Zhang, Boyi Deng, C. C. Yu, Chao Jin, Chaofan Lin, Chen Dong, Chenbing Wang, Chenfan Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chengyuan Zhang, Chenhao Xu, Chenqi Zhao, Chenze Shao, Chuhao Wang, Chuqi Zhang, Damai Dai, Dejian Yang, Deli Chen, Di Huang, Di Wu, Donghao Li, Erhang Li, Eric Fu, F. Zhou, Fangwei Zhou, Fangyun Lin, Fangzhou Yuan, Feiyu Xia, Fucong Dai, Guangbo Hao, Guanglin Li, Guanting Chen, Guoai Cao, Guofan Fan, Guolai Meng, Guowei Li, Haichuan Zhang, Haiyang Ma, Haiyang Shen, Han Li, Han Yu, Han Zhang, Hangyuan Deng, Hanwei Xu, Hanxiang Xu, Hanxun Zhong, Hao Guo, Hao Jiang, Hao Li, Hao Qin, Haodong Wen, Haofen Liang, Haofeng Huang, Haohua Liu, Haoling Zhang, Haoming Luo, Haoran Yang, Haotian Xu, Haotian Yuan, Haoting Huang, Haowen Luo, Haoyang Cai, Haoyu Chen, Haozhe Ji, Hengran Zhang, Hengrui Wang, Hengxu Wu, Honghui Ding, Hongxuan Tang, Huadong Wang, Huanqi Cao, Huazuo Gao, Hui Qu, Hui Zeng, J. Yang, J. H. Jin, J. H. Zhang, J. X. Zou, Jia Yu, Jiahui Zhou, Jiajun Chen, Jialiang Huang, Jialin Zhao, Jiamin Tang, Jian Zhou, Jianan Tong, Jianwen Li, Jiaqi Zhu, Jiarui Wang, Jiasheng Ye, Jiashi Li, Jiaxin Xu, Jiaying Ding, Jibai Lu, Jiewen Hu, Jin Yan, Jincheng Zhai, Jingchang Chen, Jingcheng Hu, Jingli Zhou, Jingsheng Xu, Jingting Xiang, Jingyan Yun, Jingyang Yuan, Jingyuan Cheng, Jinhua Zhu, Jinpeng Wang, Jinyi Chen, Jinyi Hu, Jiping Yu, Jueliang Guo, Junbo Pei, Junbo Sun, Junguang Jiang, Junjie Qiu, Junkang Zhou, Junqi Liu, Junren Li, Junxian Li, Junxiao Song, Junyi Guo, Kai Dong, Kaifeng Chen, Kaige Gao, Kang Guan, Kangdong Yuan, Ke Hong, Ke Xu, Kefan Zhao, Kexin Ji, Kexin Zhang, Kexing Zhou, Kuai Yu, Lan Zhang, Lean Wang, Lecong Zhang, Lei Wang, Letian Gao, Liang Zhao, Liansheng Xu, Lihua Guo, Lingxiao Luo, Lingyue

Fu, Litao Deng, Litong Wang, Liyue Zhang, Longhao Chen, Lu Chen, Luotian Huang, Luyao Ma, Luyao Wang, M. S. Di, Max Mei, Menghao Ye, Miao Cui, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingjing Zhang, Mingqi Wei, Mingshu Chen, Mingxing Liu, Mingxu Zhou, Mingyu Xu, Mingyu Yang, Mingze Wang, Muyang Chen, Ni Shentu, Ning Wang, Niufang Ning, Panpan Huang, Peixin Cong, Peiyi Wang, Peiyuan Xin, Pengfei Ren, Pengfei Yan, Pengle Zhang, Qi Kang, Qi Tang, Qiancheng Wang, Qiang Li, Qihao Zhu, Qingyang Li, Qinyu Chen, Qiushi Du, Qizhou Guo, Rongxian Xu, Rui Ding, Rui Hu, Rui Tian, Rui Yu, Ruidong Zhu, Ruifan Xu, Ruihan Yang, Ruihang Xia, Ruijie Lu, Ruilin Geng, Ruipeng Hong, Ruiqi Ge, Ruisong Zhang, Ruize Sun, Ruizhe Pan, Runji Wang, Runqian Chen, Runxin Xu, Ruohong Tian, Ruomeng Shen, Ruoyu Zhang, Ryan X., S. H. Liu, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaofei Cai, Shaoheng Nie, Shaoyuan Chen, Shengding Hu, Shengkai Lin, Shengwen Ran, Shengyu Liu, Shengyuan Jia, Shi Bai, Shi Feng, Shicheng Xu, Shichun Liu, Shiqiang Hu, Shirong Ma, Shiyu Wang, Shiyuan Feng, Shufan Gong, Shuhan Lin, Shuiping Yu, Shunfeng Zhou, Shuo Yang, Shuomeng Wang, Shuting Guo, Shuting Pan, Shuying Yu, Sinuo Cao, Siyi Lin, Sizhe Chen, Songyang Chen, Songyang Zhou, Tao Ni, Tao Yun, Tian Jin, Tian Pei, Tian Ye, Tianle Lin, Tianran Ji, Tianyi Cui, Tianyuan Yue, Tingting Yu, Tongrui Xiong, Wangding Zeng, Wei Liu, Wei Zhang, Weibin Xu, Weihao Zeng, Weilin Zhao, Wen Liu, Wenfeng Liang, Wenjie Pang, Wenjing Luo, Wenjing Yao, Wenjun Gao, Wenkai Shao, Wenkai Yang, Wenli Zhang, Wenlu Wang, Wenlve Huang, Wenqian Yan, Wentao Zhang, Xi Gao, Xiang He, Xiang Li, Xiangli Li, Xiangwen Wang, Xiangying Zhang, Xiankui Wei, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojian Qu, Xiaokang Chen, Xiaokang Zhang, Xiaotao Nie, Xiaoyao Zou, Xiaoyuan Li, Xicheng Guo, Xieting Chu, Xin Cheng, Xin Liu, Xin Xie, Xinbo Xu, Xingchao Liu, Xingchen Liu, Xingkai Yu, Xingyou Li, Xintong Yao, Xinyang Chen, Xinyong Jiang, Xinyu Yang, Xinyu Yang, Xu Chen, Xuanyu Wang, Xubei Zhong, Xuecheng Su, Xuejie Liu, Xuheng Lin, Xujie Fan, Xuncheng Zhao, Xuwei Fu, Y. C. Yan, Y. H. Jiang, Y. T. Wu, Y. W. M., Y. Z. Wang, Yafei Gao, Yang Yang, Yang Zhang, Yanru Ma, Yanwen Huang, Yao Li, Yao Li, Yao Meng, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yaoyang Ye, Yehang Yin, Yexinrui Wu, Yi Qian, Yi Tao, Yi Yu, Yichao Zhang, Yichen Jiang, Yicheng Wang, Yifan Ding, Yifan Shi, Yifeng Peng, Yifeng Zhai, Yijia Wu, Yiliang Xiong, Yilun Wang, Ying He, Ying Zhou, Yingjia Luo, Yinmin Zhong, Yiping Wang, Yisong Wang, Yixiang Zhang, Yixiao Chen, Yixuan Tan, Yixuan Wei, Yiyang Ma, Yiyao Yang, Yiyuan Liu, Yizai Cai, Yizhen Wei, Yizhi Wang, Yonglun Yang, Yongqi Zhuo, Yongqiang Guo, Yongtong Wu, Yu Wu, Yu Zhang, Yuan Bian, Yuan Cheng, Yuan Ou, Yuan Sun, Yuanfan Xu, Yuanhang Sun, Yuanhao Li, Yuchen Liu, Yuchen Yao, Yudong Han, Yuduan Wang, Yuhan Wu, Yuhao Meng, Yuheng Zou, YuKun Li, Yunchuan Wang, Yunfan Xiao, Yunfan Xiong, Yupeng Chen, Yuqian Cao, Yuqian Wang, Yuqing Chen, Yushun Zhang, Yutong Lin, Yuwei Xiao, Yuxian Gu, Yuxiang Chen, Yuxiang Huang, Yuxiang Luo, Yuxiang You, Yuxin Chen, Yuxin Xiang, Yuxuan Liu, Yuxuan Zhou, Yuyang Zhou, Yuzhe Guo, Yuzhen Huang, Yuzhuo Bai, Z. Y. Z., Zanlin Ni, Zehao Wang, Zehua Zhao, Zehui Ren, Zejun Zhao, Zhangli Sha, Zhanying Wang, Zhaochen Zhang, Zhaoshuai Du, Zhe Fu, Zhean Xu, Zhenda Xie, Zheng Liu, Zhengyan Zhang, Zhenhua Dong, Zhewen Hao, Zhibang Wang, Zhibin Gou, Zhicheng Ma, Zhihao Li, Zhihong Shao, Zhihuan Huang, Zhijie Li, Zhirui Lu, Zhixian Huang, Zhixuan Chen, Zhixuan Chen, Zhixuan Pan, Zhiyu Wu, Zhizhou Ren, Zhu He, Zhuoshu Li, Zhuping Zhang, Zian Xu, Zihao Wang, Zihui Gu, Zijia Zhu, Zili Zhang, Zilin Li, Zilong Hou, Zilong Lyu, Ziqiao Wang, Ziwei Xie, Ziya Zhang, Ziyi Gao, Zizheng Pan, Zonglin Li, Zongqing Yao, Zui Chen, Zuofan Wu, Chenchen Ling, Chengyu Hou, Chong Chen, D. Li, Di Qi, Dongjie Ji, Fang Wei, Fanyi Xia, Fei Xie, Feiyi Tan, Hailong Guo, Haiyan Zhai, Hui Zhou, Huihui Tan, Huijie Li, Jia Luo, Jia Song, Jialu Cai, Jian Liang, Jiangting Zhou, Jiaqi Gao, Jiayi Shao, Jie Chen, Jieyu Yang, Jin Chen, Jingde Zhang, Jingzi Zhou, Jinqian Wang, Jinyang Liu, JinZhao Sun, Junhua Ling, Junmin Zheng, Kaicheng Yang, Ke Xu, Le Su, Leyi Xia, Liangfeng Ding, Lin Zhuo, Linwang Ma, Linyan Zhu, Liyu Cai, Luqi Yao, M. K. Zhang, Meng Li, Miao Lin, Miaojun Wang, Min Zhang, Mingming Li, Mingming Wang, Mingze Yin, Minmin Han, Nan Cao, Ning Wang, Ningxin Ma, Panpan Wang, Peihan Lin, Peng Sun, Peng Zhang, Qian Ying, Qiang Xiang, Qiao Wang, Qingmiao Mao, Qiwei Jiang, Rongli Jin, Ruyi Chen, Sha Tao, Shangmian Sun, Shaoqing Wu, Shichao Zou, Si Lei, Tianyang Zhang, Tianyu Sun, Tingting Yin, W. L. Xiao, Wei An, Wei Li, Wei Wang, Weiwei Lin, Wenqing Hou, X. Lin, Xiangfei Meng, Xianzhu Huang, Xiao Peng, Xiaoqian Li, Xiaoting Zhang, Xiaowen Sun, Xiaoxiang Wang, Xiaoyu Ye, Xinrou Zhang, Xinyu Zhang, Xue Cao, Xueyin Chen, Yanan Zhou, Yanhong Xu, Yao Xia, Yao Xu, Yi Shao, Yihong Zhang, Yiling Ma, Ying Tang, Yining Lou, Yiru Chen, Yishi Piao, Yixuan Chen, Yong Xiong, Yuchen Xuan, Yuehan Yang, Yuer Xu, Yukun Zha, Yunxian Ma, Yuping Lin, Yuting Yan, Yutong Xie, Yuwen Sheng, Yuxuan Zhu, Zekai Zhang, Zhe Ju, Zhenzhen

Lin, Zheren Gao, Zheyang Sun, Zhigang Yan, Zhongyu Wu, Zi Wang, Zihua Qu, Ziling Yan, and Ziyi Wan. Deepseek-v4.1-flash: Pushing the limits of kv cache compression, 2026. URL https://arxiv.org/abs/2609.19969.

Shuangrui Ding, Xuan jun Dai, Long Xing, Sheng-Yuan Ding, Ziyu Liu, Jingyi Yang, Penghui Yang, Zhixiong Zhang, Xilin Wei, Xinyu Fang, Yubo Ma, Haodong Duan, Jing Shao, Jiaqi Wang, Dahua Lin, Kai Chen, and Yuhang Zang. Wildclawbench: A benchmark for real-world, long-horizon agent evaluation. *ArXiv*, abs/2605.10912, 2026. URL https://api.semanticscholar.org/CorpusID:288257725.

Google. Gemini 3.8 flash, 2026. URL https://deepmind.google/models/gemini/flash/.

Xiaomeng Hu, Ying-Er Zhang, Fei Huang, Jianhong Tu, Yang Su, Liang-Hao Deng, Yuxuan Liu, Yan-Tao Liu, Dayiheng Liu, and Tsung-Yi Ho. Occubench: Evaluating ai agents on real-world professional tasks via language environment simulation. *ArXiv*, abs/2604.10866, 2026. URL https://api.semanticscholar.org/CorpusID:287432902.

Jing Yu Koh, Robert Lo, Lawrence Keunho Jang, Vikram Duvvur, Ming Chong Lim, Po-Yu Huang, Graham Neubig, Shuyan Zhou, Ruslan Salakhutdinov, and Daniel Fried. Visualwebarena: Evaluating multimodal agents on realistic visual web tasks. *ArXiv*, abs/2401.13649, 2024. URL https://api.semanticscholar.org/CorpusID:267199749.

Xiangyi Li, Kyoung Whan Choe, Yimin Liu, Xiaokun Chen, Chujun Tao, Bin You, Wen-Bo Chen, Zong-Lin Di, Jiankai Sun, Sheng-Chun Zheng, Jiajun Bao, Yuanli Wang, Weixiang Yan, Yiyuan Li, and Han Lee. Clawsbench: Evaluating capability and safety of llm productivity agents in simulated workspaces. *ArXiv*, abs/2604.05172, 2026a. URL https://api.semanticscholar.org/CorpusID:287208605.

Yuetai Li, Yichen Feng, Zhangchen Xu, Zixian Ma, Kaiyuan Zheng, Fengqing Jiang, Xinghua Sun, Rulin Shao, Zichen Chen, Yue Huang, Xinyan Han, Brian Lee, Kayla Y. Xu, Shenglai Zeng, Hang Hua, Xiangliang Zhang, Basel Alomair, Ranjay Krishna, Luke S. Zettlemoyer, Pang Wei Koh, Bhaskar Ramasubramanian, Luyao Niu, Xiang Yue, and Radha Poovendran. Jobbench: Aligning agent work with human will. *ArXiv*, abs/2605.26329, 2026b. URL https://api.semanticscholar.org/CorpusID:288671779.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2304.08485.

Xiao Liu, Tianjie Zhang, Yu Gu, Iat Long Iong, Yi-Fan Xu, Xixuan Song, Shudan Zhang, Hanyu Lai, Xinyi Liu, Hanlin Zhao, Jiadai Sun, Xinyue Yang, Yu Yang, Ze-Han Qi, Shuntian Yao, Xueqiao Sun, Siyi Cheng, Qi Zheng, Hao Yu, Hanchen Zhang, Wenyi Hong, Ming Ding, Li-Hang Pan, Xiao-Tao Gu, Ao-Han Zeng, Zheng-Xiao Du, Chan Hee Song, Yu Su, Yu-Xiao Dong, and Jie Tang. Visualagentbench: Towards large multimodal models as visual foundation agents. *ArXiv*, abs/2408.06327, 2024. URL https://api.semanticscholar.org/CorpusID:271854812.

Mantas Mazeika, Alice Gatti, Cristina Menghini, Udari Madhushani Sehwag, Shivam Singhal, Yury Orlovskiy, Steven Basart, Manasi Sharma, Denis Peskoff, Elaine Lau, Jaehyuk Lim, Lachlan Carroll, Alice Blair, V. Sivakumar, Sumana Basu, Brad Kenstler, Yuntao Ma, Julian Michael, Xiaoke Li, Oliver Ingebretsen, Aditya Mehta, Jean Mottola, John Teichmann, Kevin Yu, Zaina Shaik, Adam Khoja, Richard Ren, Jason Hausenloy, Long Phan, Ye Htet, Ankit Aich, Tahseen Rabbani, Vivswan Shah, A. Novykov, Felix C. Binder, K. V. Chugunov, Luis F. Ramírez, Matias Geralnik, Hern'an Mesura, Dean Lee, Ed-Yeremai Hernandez Cardona, A. S. Diamond, Summer Yue, Alexandr Wang, Bing Liu, Ernesto Hernandez, and Dan Hendrycks. Remote labor index: Measuring ai automation of remote work. *ArXiv*, abs/2510.26787, 2025. URL https://api.semanticscholar.org/CorpusID:282592087.

Samuel Miserendino, Michele Wang, Tejal Patwardhan, and Johannes Heidecke. Swe-lancer: Can frontier llms earn $1 million from real-world freelance software engineering? *ArXiv*, abs/2502.12115, 2025. URL https://api.semanticscholar.org/CorpusID:276421817.

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

OpenAI. Introducing chatgpt, 2022. URL https://openai.com/index/chatgpt/.

OpenAI. Gpt-4, 2023. URL https://openai.com/index/gpt-4-research/.

OpenAI. Gpt-5.6 luna, 2026a. URL https://developers.openai.com/api/docs/models/gpt-5.6-luna.

OpenAI. Gpt-5.6 sol, 2026b. URL https://developers.openai.com/api/docs/models/gpt-5.6-sol.

OpenAI. Workspace agents. https://openai.com/academy/workspace-agents/, 2026c.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail

Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

OpenClaw. Openclaw, 2026. URL https://github.com/openclaw/openclaw.

Tejal Patwardhan, Rachel Dias, Elizabeth Proehl, Grace Kim, Michele Wang, Olivia Watkins, Sim'on Posada Fishman, Marwan Aljubeh, Phoebe Thacker, Laurance Fauconnet, Natalie S. Kim, Patrick Chao, Samuel Miserendino, Gildas Chabot, David Li, Michael Sharman, Alexandra Barr, Amelia Glaese, and Jerry Tworek. Gdpval: Evaluating ai model performance on real-world economically valuable tasks. *ArXiv*, abs/2510.04374, 2025. URL https://api.semanticscholar.org/CorpusID:281843768.

QwenTeam. Qwen3.8-max: A new bar for coding and cowork, 2026. URL https://qwen.ai/blog?id=qwen3.8.

Dan Shepard and Robin Salimans. Automationbench. *ArXiv*, abs/2604.18934, 2026. URL https://api.semanticscholar.org/CorpusID:287636396.

Olly Styles, Sam Miller, Patricio Cerda-Mardini, Tanaya Guha, Victor Sanchez, and Bertie Vidgen. Workbench: a benchmark dataset for agents in a realistic workplace setting. *ArXiv*, abs/2405.00823, 2024. URL https://api.semanticscholar.org/CorpusID:269502506.

Zi-Rui Tang, Xuanhe Zhou, Yu-Mou Liu, Linchun Li, Yu-Kai Wu, Wei-Zheng Wang, Hong-Yi Huang, Wei Zhou, Jun Zhou, Jiacheng Song, Shao-Lin Yu, Jinqi Wang, Zihan Zhou, Hongyi Zhou, Yu-Ting Lv, Jin-Yang Li, Jia-Shuo Liu, Ruoyu Chen, Chun-Wei Liu, Guoliang Li, Jihua Kang, and Fan Wu. Workspace-bench 1.0: Benchmarking ai agents on workspace tasks with large-scale file dependencies. *ArXiv*, abs/2605.03596, 2026. URL https://api.semanticscholar.org/CorpusID:287962808.

Core Team, Bangjun Xiao, Bingquan Xia, Bo Yang, Bofei Gao, Bowen Shen, Chen Zhang, Chenhong He, Chiheng Lou, Fuli Luo, Gang Wang, Gang Xie, Hailin Zhang, Hanglong Lv, Hanyu Li, Heyu Chen, Hongshen Xu, Houbin Zhang, Huaqiu Liu, Jiangshan Duo, Jianyu Wei, Jiebao Xiao, Jinhao Dong, Jun Shi, Junhao Hu, Kainan Bao, Kang Zhou, Lei Li, Liang Zhao, Linghao Zhang, Peidian Li, Qianli Chen, Shaohui Liu, Shihua Yu, Shijie Cao, Shimao Chen, Shouqiu Yu, Shuo Liu, Tianling Zhou, Weijiang Su, Weikun Wang, Wenhan Ma, Xiangwei Deng, Bohan Mao, Bowen Ye, Can Cai, Chenghua Wang, Chengxuan Zhu, Chong Ma, Chun Chen, Chunan Li, Dawei Zhu, Deshan Xiao, Dong Zhang, Duo Zhang, Fangyue Liu, Feiyu Yang, Fengyuan Shi, Guoan Wang, Hao Tian, Hao Wu, Heng Qu, Hongfei Yi, Hongxu An, Hongyi Guan, Xing Zhang, Yifan Song, Yihan Yan, Yihao Zhao, Yingchun Lai, Yizhao Gao, Yu Cheng, Yuanyuan

Tian, Yudong Wang, Zhen Tang, Zhengju Tang, Zhengtao Wen, Zhichao Song, Zhixian Zheng, Zihan Jiang, Jian Wen, Jiarui Sun, Jiawei Li, Jinlong Xue, Jun Xia, Kai Fang, Menghang Zhu, Nuo Chen, Qian Tu, Qihao Zhang, Qiying Wang, Rang Li, Rui Ma, Shaolei Zhang, Shengfan Wang, Shicheng Li, Shuhao Gu, Shuhuai Ren, Sirui Deng, Tao Guo, Tianyang Lu, Weiji Zhuang, Weikang Zhang, Weimin Xiong, Wenshan Huang, Wenyu Yang, Xin Zhang, Xing Yong, Xu Wang, Xueyang Xie, Yilin Jiang, Yixin Yang, Yongzhe He, Yu Tu, Yuanliang Dong, Yuchen Liu, Yue Ma, Yue Yu, Yuxing Xiang, Zhaojun Huang, Zhenru Lin, Zhipeng Xu, Zhiyang Chen, Zhonghua Deng, Zihan Zhang, and Zihao Yue. Mimo-v2-flash technical report, 2026a. URL https://arxiv.org/abs/2601.02780.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, M. C., Jianfeng Cai, Xinyuan Cai, Peizhou Cao, Yuxuan Cao, Ziwei Chai, Y. Charles, H. S. Che, Guanduo Chen, Guangyu Chen, Guanzheng Chen, Huarong Chen, Jia Chen, Jianlong Chen, Jun Chen, Kexin Chen, Peng Chen, Ruijue Chen, Wentao Chen, Xin Chen, Yang Chen, Yanru Chen, Yifei Chen, Yingjiang Chen, Yuankun Chen, Yujie Chen, Yutian Chen, Zhirong Chen, Dazhi Cheng, Yean Cheng, Jialei Cui, Jingbing Cui, Anqi Dai, Jiaqi Deng, Hao Ding, Rui Ding, Shaofeng Ding, Mengfan Dong, Mengnan Dong, Yuhao Dong, Yuxin Dong, Angang Du, Chenzhuang Du, Dikang Du, Jusen Du, Yulun Du, Yu Fan, Jing Feng, Qiulin Feng, Yichen Feng, Kelin Fu, Qiang Fu, Fuxuan Gao, Hongcheng Gao, Jingyue Gao, Tong Gao, Weijia Gao, Shangyi Geng, Jie Gong, Linhu Gong, Shengao Gong, Xiaochen Gong, Qizheng Gu, Yicheng Gu, Shuhao Guan, Haiqing Guo, Shiqi Guo, Xiang Guo, Zhengyan Guo, Beixi Hao, Wenxin Hao, Xiaoru Hao, Dailan He, Haotian He, Lehan He, Qi He, Weiran He, Xinran He, Xinyi He, Yibo He, Yunjia He, Chao Hong, Tiange Hong, Hao Hu, Jiaxi Hu, Ruikun Hu, Weiming Hu, Yangyang Hu, Zhenxing Hu, Liang Hua, Jinbin Huang, Ke Huang, Ruiyuan Huang, Siying Huang, Weixiao Huang, Yan Huang, Zhengjie Huang, Zhiqi Huang, Yulong Hui, Chaobo Jia, Yutong Jiang, Zhejun Jiang, Zuoyou Jiang, Wenyi Jin, Xinyi Jin, Yu Jing, Huanjun Kong, Guokun Lai, Aidi Li, Cheng Li, Chengyuan Li, Cong Li, Fang Li, Guanyu Li, Haoyang Li, Jia Li, Junxiong Li, Lei Li, Letian Li, Lincan Li, Weihong Li, Wentao Li, Xintong Li, Yang Li, Yishen Li, Yiwei Li, Yuxiao Li, Zhaowei Li, Zhaoxi Li, Zheming Li, Zhengxiao Li, Zhiyuan Li, Jiawei Lin, Xiaohan Lin, Yibo Lin, Zichao Lin, Ziyan Lin, Bill Liu, Boxiao Liu, Chuan Liu, Liang Liu, Shaowei Liu, Shudong Liu, Shuran Liu, Tianwei Liu, Weizhou Liu, Yangyang Liu, Yanming Liu, Yibo Liu, Yipeng Liu, Zhengying Liu, Zhiheng Liu, Enzhe Lu, Haoyu Lu, Linqiang Lu, Tingzhan Lu, Zhiyuan Lu, Aotian Luo, G. Luo, Junyu Luo, Yifan Luo, B. Lyu, Wenzhou Lyu, Shaoguang Mao, Yuan Mei, Xin Men, Minqing Ni, Yixuan Niu, Siyuan Pan, Shujun Peng, Zhangyang Qi, Ruoyu Qin, ZeChao Qin, Zeyu Qin, Haiquan Qiu, Jianxin Qiu, Jiezhong Qiu, Bowen Qu, Yuhao Qu, Zeyu Shang, Youbo Shao, Han Shen, Jincheng Shi, Juanfeng Shi, Lidong Shi, Shengyuan Shi, Wingchun Siu, Pengwei Song, Xiaoxi Song, Jianlin Su, Yunfeng Su, Zhaochen Su, Lin Sui, Jingsong Sun, Junyao Sun, Shaoning Sun, Shuzhe Sun, Tongyu Sun, Yujun Sun, Yunpeng Tai, Chuning Tang, Heyi Tang, Sirui Tang, Zecheng Tang, Chaoran Tian, Rongpeng Tian, Yu Tian, Wei Tu, Chensi Wang, Chuang Wang, Chunjie Wang, Dinglu Wang, Feng Wang, Hailong Wang, Haiming Wang, Hao Wang, Hao Wang, Huaqing Wang, Hui Wang, Jiayi Wang, Jinglong Wang, Jinhong Wang, Jiuzheng Wang, Linian Wang, Shaobo Wang, Shenzhi Wang, Shuyi Wang, Si Wang, Siyuan Wang, Tianfu Wang, Wenjue Wang, Xingran Wang, Xinmei Wang, Xinyuan Wang, Xusheng Wang, Yalin Wang, Yangkun Wang, Yao Wang, Yaoyu Wang, Yejie Wang, Yiqin Wang, Yucheng Wang, Yuzhi Wang, Zhaoji Wang, Zhaowei Wang, Zhengtao Wang, Zhenhao Wang, Zhongsheng Wang, Zifan Wang, Chu Wei, Ming Wei, Shouxin Wei, Zichen Wen, Fan Wu, Haoning Wu, Rucong Wu, Wenhao Wu, Xiaoxue Wu, Yingcong Wu, Yongqi Wu, Yuxin Wu, Zijian Wu, Xinglang Xian, Chenxuan Xiang, Yuye Xiang, Bocheng Xiao, Chenjun Xiao, Xin Xiao, Jin Xie, Xiaotong Xie, Yifeng Xie, Zhe Xie, Bowei Xing, Yiming Xiong, Baosheng Xu, Boyu Xu, Jiale Xu, Jianfan Xu, Jing Xu, Jinjing Xu, L. H. Xu, Qingtao Xu, Shuyao Xu, Suting Xu, Tiantian Xu, Tianxiang Xu, Weixin Xu, Xinran Xu, Yangchuan Xu, Ye Xu, Yueni Xu, Ziyao Xu, Haonan Xue, Junjie Yan, Yaoyao Yan, Fan Yang, Guangyao Yang, Hao Yang, Junwei Yang, Ruoyu Yang, Wenjie Yang, Xiaofei Yang, Xinyu Yang, Yi Yang, Yiling Yang, Ying Yang, Yuchen Yang, Zhen Yang, Zhilin Yang, Zian Yang, Zuhao Yang, Haotian Yao, Dan Ye, Haoran Ye, Wenjie Ye, Zhanbo Ye, Bohong Yin, Haoxiang Yin, Xietong Yin, Chengzhen Yu, Haozhen Yu, Longhui Yu, Shengnan Yu, Shuying Yu, Tianxiang Yu, Enming Yuan, Mengjie Yuan, Tongtian Yue, Wei Yue, Yang Yue, Dunyuan Zha, Haobing Zhan, B. H. Zhang, Dehao Zhang, Fei Zhang, Hao Zhang, Haoyuan Zhang, Huanyu Zhang, Jiapei Zhang, Jiaxuan Zhang, Jin Zhang, Kaiyi Zhang, Miaozhen Zhang, Puqi Zhang, Qinglei Zhang, Rong Zhang, Rui Zhang, Shaoshuai Zhang, Shiyi Zhang, Xiaobin Zhang, Xiaoyun Zhang, Y. Zhang,

Yangkun Zhang, Ye Zhang, Yichi Zhang, Yikun Zhang, Yizhi Zhang, Yongting Zhang, Yu Zhang, Yutao Zhang, Yutong Zhang, Zheng Zhang, Zijing Zhang, Bin Zhao, Chenguang Zhao, Feifan Zhao, Jinglun Zhao, Jinxiang Zhao, Shuai Zhao, Wenshuo Zhao, Xiangyu Zhao, Xuanle Zhao, Yikai Zhao, Zijia Zhao, Haozhi Zheng, Huabin Zheng, Ruihan Zheng, Shaojie Zheng, Tengyang Zheng, Haofeng Zhong, Lei Zhong, Longguang Zhong, M. Zhou, Qiankang Zhou, Runjie Zhou, Ruozhang Zhou, Xinyu Zhou, Yiqiao Zhou, Zaida Zhou, Jinguo Zhu, Liya Zhu, Xinhao Zhu, Yangjunfeng Zhu, Yuxuan Zhu, Zhen Zhu, Chen Zhuang, Weiyu Zhuang, and Xinxing Zu. Kimi k3: Open frontier intelligence, 2026b. URL https://arxiv.org/abs/2607.24653.

United Nations. *International Standard Industrial Classification of All Economic Activities (ISIC), Revision 5*. Number No. 5/Rev.5 in Statistical Papers, Series M. United Nations, 2024.

Bertie Vidgen, Arjun Mann, Abby Fennelly, Jonathan Stanly, Lucas Rothman, Marcos Herrera Burstein, Julien Benchek, David Ostrofsky, Anirudh Ravichandran, Debnil Sur, Neel Venugopal, Alannah Hsia, Isaac Robinson, Calix Huang, Olivia Varones, Daniyal Khan, Michael R. Haines, Austin Bridges, Jesse Boyle, Koby Twist, Zach Richards, Chirag Mahapatra, Brendan Foody, and Osvald Nitski. Apex-agents. *ArXiv*, abs/2601.14242, 2026. URL https://api.semanticscholar.org/CorpusID:284909788.

xAI. Introducing grok 4.6, 2026. URL https://x.ai/news/grok-4-6.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiao-Chuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *ArXiv*, abs/2404.07972, 2024. URL https://api.semanticscholar.org/CorpusID:269042918.

Frank F. Xu, Yufan Song, Boxuan Li, Yuxuan Tang, Kritanjali Jain, Meng Bao, Zora Zhiruo Wang, Xuhui Zhou, Zhitong Guo, Murong Cao, Ming-Hsuan Yang, Hao Lu, Amaad Martin, Zhe Su, Leander Melroy Maben, Raj Mehta, Wayne Chi, Lawrence Keunho Jang, Yi-Qing Xie, Shuyan Zhou, and Graham Neubig. Theagentcompany: Benchmarking llm agents on consequential real world tasks. *ArXiv*, abs/2412.14161, 2024. URL https://api.semanticscholar.org/CorpusID:274822848.

Qianyu Yang, Yang Liu, Jiaqi Li, Junye Bai, Hao Chen, Kaiyuan Chen, Tiliang Duan, Jiayun Dong, Xiaobo Sharon Hu, Zixia Jia, Tao Peng, Yixin Ren, Ran Tian, Zaiyuan Wang, Yang Xiao, Gang Yao, Lingyue Yin, Ge Zhang, Chun Zhang, Jianpeng Jiao, Zilong Zheng, and Yu-Ting Gong. $onemillion-bench: How far are language agents from human experts? *ArXiv*, abs/2603.07980, 2026. URL https://api.semanticscholar.org/CorpusID:286376578.

Bowen Ye, Rang Li, Qibin Yang, Yuanxin Liu, Linli Yao, Hanglong Lv, Zhihui Xie, Chenxin An, Lei Li, Ling-Peng Kong, Qi Liu, Zhifang Sui, and Tong Yang. Claw-eval: Towards trustworthy evaluation of autonomous agents. 2026. URL https://api.semanticscholar.org/CorpusID:287209548.

Mengqi Yuan, Zilong Zhou, Xinzhuang Xiong, Weiming Wu, Jiayang Sun, Jiamin Song, Kaiqian Cui, Bowen Wang, Haoyuan Wu, Yitong Li, Dunjie Lu, Haikong Lu, Qi Zhen, Xinyuan Wang, Jiaqi Deng, Yuhao Yang, Cheng Chen, Boyuan Zheng, Alex Su, Xiao Yu, Hao Zou, Saaket Agashe, Xing Han Lu, Manpreet Kaur, Zhengyang Qi, Vincent Sunn Chen, Frederic Sala, Dayiheng Liu, Junyang Lin, Zhou Yu, Yu Su, Siva Reddy, Xin Eric Wang, Peng Qi, Tianbao Xie, and Tao Yu. Osworld 2.0: Benchmarking computer use agents on long-horizon real-world tasks, 2026. URL https://arxiv.org/abs/2606.29537.

Z.ai. Glm-5.3-flash: Frontier intelligence, flash cost, 2026. URL https://z.ai/blog/glm-5.3-flash.

Miaosen Zhang, Qi Dai, Yi-Fan Yang, Jian-Min Bao, Dongdong Chen, Kai Qiu, Chong Luo, Xin Geng, and Bai-Ning Guo. Magebench: Bridging large multimodal models to agents. *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1415–1427, 2024. URL https://api.semanticscholar.org/CorpusID:274581428.

Table 3: Primary work-domain definitions and counts in the 115 tasks.

Work domain	Operational definition	n
RT: Retail and customer service	Product discovery, purchase, refund, exchange, merchandising, or customer-facing service workflows.	17
ER: Education, research, and culture	Teaching, learning, scientific analysis, archives, museums, or cultural-heritage work.	14
DI: Digital products and IT	Software interfaces, websites, dashboards, interactive media, and technical product workflows.	14
BM: Brand, media, and creative	Branding, advertising, photography, visual identity, and creative content production.	11
TE: Travel, venues, and events	Travel planning, venue selection, event access, itineraries, and event production.	10
EP: Energy, environment, and infrastructure	Environmental monitoring, energy systems, public infrastructure, maps, and risk-related operations.	10
RE: Buildings, property, and interior	Housing, property, architecture, interior furnishing, repair, and spatial planning.	9
ML: Manufacturing and logistics	Manufacturing, packaging, supply chain, cold-chain, quality, and operational logistics.	9
BF: Business, finance, and legal	Enterprise analysis, contracts, finance, compliance, and business operations.	9
FH: Food and hospitality	Restaurants, cafes, bakeries, menus, food service, and hospitality operations.	7
HS: Health, safety, and care	Clinical, health, safety, accessibility, and care-related decisions or services.	5

A Work-Domain and Work-Objective Definitions

We characterize each task along two complementary axes: the *work domain*, which describes the institutional and economic context in which the user goal arises, and the *work objective*, which describes the principal form of work needed to achieve that goal. These annotations are defined independently of the file format, software, or tools involved in task execution.

For work domains (Tab. 3), we use the International Standard Industrial Classification of All Economic Activities (ISIC Rev. 5) (United Nations, 2024) as an external reference for sectoral coverage. We do not adopt ISIC categories directly: benchmark tasks often cross adjacent economic activities, and several settings that are operationally similar for agent evaluation are separated in the standard classification. We therefore consolidate related activities into eleven benchmark-specific work domains that reflect the contexts represented in our tasks. For example, *Digital products and IT* groups computing, software, and digital-service workflows, while *Education, research, and culture* combines educational, scientific, archival, and cultural settings. A task is assigned according to the primary context of the user's goal rather than the organization type, artifact format, or application used.

For work objectives (Tab. 4), we define seven benchmark-specific categories describing the primary transformation from available information and actions to the requested outcome.

The two axes are intentionally non-redundant. The same objective may arise in different work domains, and tasks within the same domain may require different objectives. For example, a retail task may require either product *search, match, and select* or an *audit, verify, diagnose* workflow, while *design and create* tasks occur across branding, digital products, education, and hospitality. This separation allows the benchmark to report performance by both the context in which work occurs and the type of work the agent is asked to perform.

B Token-pricing assumptions for the resource analysis

Tab. 5 gives public on-demand rates (USD per million tokens). For each trial, cost is $[(I-H)p_{\text{input}} + Hp_{\text{cache}} + Op_{\text{output}}]/10^6$, where I, H, and O are total input, cached input, and output. The runner sums provider cache-hit counts across API calls and caps H at I; the exported uncached count

Table 4: Primary work-objective definitions and counts in the 115 tasks.

Work objective	Operational definition	n
AU: Audit, verify, diagnose	Check claims, detect inconsistencies or defects, attribute causes, and determine whether constraints are satisfied.	35
SE: Search, match, and select	Retrieve candidates, identify visual matches, compare alternatives, and choose an item or option.	17
DE: Design and create	Synthesize a new visual, document, interface, or presentation under semantic and visual constraints.	16
TR: Transform and reconstruct	Convert, redraw, rebuild, or recover information across visual, structured, and editable representations.	14
IM: Implement, repair, and interact	Modify or operate an interface, application, or interactive artifact and verify the resulting state.	13
PL: Plan and coordinate	Produce a plan, route, schedule, recommendation, or coordinated sequence of actions.	10
RP: Organize and report	Organize media or records and deliver a structured report, register, or evidence package.	10

equals $I - H$ in all nine model exports. Missing cache details are recorded as zero, which does not prove there were no cache hits. The right panel sums three trial costs per fully measured task, then plots their mean. The resulting comparison is token-based; monetary charges that depend on provider tariffs, cache policies, or separately billed services are not claimed to be recovered from these measurements.

Table 5: Public token rates used for cost estimation (USD per million tokens).

Model	Input	Cached input	Output	Provider source
DeepSeek-v4.1 Flash	0.30	0.006	1.20	DeepSeek
Gemini-3.8 Flash	0.75	0.075	3.75	Google
GLM-5.3 Flash	0.15	0.030	0.50	Z.AI
Grok-4.6	2.00	0.50	6.00	xAI
Kimi-K3 Max	3.00	0.30	15.00	Kimi
MiMo-v2.5	0.14	0.0028	0.28	Xiaomi
Qwen-3.8 Max	1.78	0.22	5.33	Alibaba Cloud
GPT-5.6 Sol	4.00	0.40	20.00	OpenAI
GPT-5.6 Luna	0.20	0.02	1.20	OpenAI

Table 6: Estimated total-token cost of three trials per task (USD). Each row uses tasks with three usable token records; completion in Fig. 7 still uses all 115 tasks.

Model	Mean	Median	P90	Max
DeepSeek-v4.1 Flash	0.37	0.33	0.66	1.53
Gemini-3.8 Flash	3.50	2.09	9.71	18.08
GLM-5.3 Flash	0.26	0.17	0.59	1.87
Grok-4.6	3.25	2.02	6.75	21.20
Kimi-K3 Max	4.79	3.41	8.99	42.18
MiMo-v2.5	0.05	0.04	0.08	0.46
Qwen-3.8 Max	2.73	1.61	5.97	28.04
GPT-5.6 Sol	2.52	2.13	3.90	15.79
GPT-5.6 Luna	0.12	0.11	0.18	0.32

DeepSeek uses peak rates (off-peak is half); Gemini uses introductory standard rates; Xiaomi uses overseas real-time rates. Qwen's US-region prices are CNY 12/1.5/36 for input/cache/output, converted at USD 1 = CNY 6.7489. Batch labels do not identify discounted API calls. Request lengths, cache writes/storage, tariff timing, and paid tool calls are unavailable, so these charges are outside this token-based comparison.

C Visual-Function Definitions and Annotation

We use task metadata, prompts, tags, and available tools to assign visual-demand labels. Each label has intensity 0 (absent), 1 (supporting), or 2 (central); the analyses in Sec. 4.4 use intensity 2. Tab. 7 defines the eleven reported functions and gives their central-demand counts. Functions can contribute jointly within a workflow. The resulting subgroups define workflow profiles under each visual demand and support demand-level comparisons of end-to-end success.

Table 7: Visual-function definitions and central-demand counts.

Function	Definition	n
Visual Grounding		
Visual content understanding	Identify visual entities and interpret their content, attributes, and relationships.	114
Constraint induction	Derive requirements for subsequent actions from visual evidence.	56
Visual Sensemaking		
Information-need recognition	Recognize visual information needed to resolve the task.	53
Active visual acquisition	Obtain relevant visual evidence through tools or interaction.	84
Cross-source synthesis	Combine and reconcile evidence across visual and nonvisual sources.	65
Visual state tracking	Maintain and update visual task state across actions.	30
Discrepancy diagnosis	Identify differences between observed and expected visual states.	49
Feedback adaptation	Revise actions in response to visual feedback.	28
Visual Realization		
Outcome specification	Specify the desired visual result under task constraints.	77
Visual construction	Produce or modify the required visual artifact or state through tools.	67
Visual verification	Check the resulting visual state against task requirements.	58

D Human Feasibility Check

As an external feasibility check, three human experts independently completed 30 tasks sampled from the Hard subset using the same task instructions and completion rubric. Their mean completion was 92%, with high agreement across raters. This check is separate from the model evaluation and indicates that the Hard tasks are executable by people under the stated workflow and grading conditions.

E Evaluation Protocol and Aggregation

Configurations. The nine canonical configurations correspond to the Native or Batch settings in the result exports. Harness comparisons use each model's canonical setting as its Base, with OpenClaw, OpenCode, and Hermes evaluated separately. All use the benchmark coverage and difficulty partition defined in Sec. 3.3.

Success metrics and trial accounting. We retain three budgeted trial slots per task, using trial identifiers 1–3. Pass metrics use the recorded binary success outcomes of the task-specific graders. Missing or errored trials count as failures, and additional reruns do not increase the trial budget. Let $y_{t,1}$ indicate success on the first budgeted trial of task t, and let c_t be the number of successful trials among the three budgeted trials. We compute Pass@1 as $\frac{1}{N}\sum_t y_{t,1}$, Pass@3 as $\frac{1}{N}\sum_t \mathbf{1}[c_t > 0]$, and Pass^3 as $\frac{1}{N}\sum_t \mathbf{1}[c_t = 3]$. Pass^3 therefore measures observed success in all three trials.

Completion. Comp. measures partial task completion using the task-specific grading rubric together with a visual LLM judge of the final output. The rubric score captures satisfaction of task-specific requirements, while the visual judge evaluates visual requirements of the produced artifact. These signals are combined according to the task-specific grader and normalized to a common completion scale. We average completion over the three trial slots and then equally over tasks. This preserves a common denominator across configurations while retaining partial progress on tasks that do not fully pass.

Table 8: Maximum-turn terminations under the canonical 100-turn budget. For DeepSeek and Gemini, we additionally report terminations after increasing the budget to 300 turns. Each setting contains 345 trials over 115 tasks.

Model	100-turn	300-turn
DeepSeek-v4.1 Flash	128 (37.1%)	45 (13.0%)
Gemini-3.8 Flash	232 (67.2%)	64 (18.6%)
GLM-5.3 Flash	0	–
Grok 4.6	10 (2.9%)	–
Kimi-K3 Max	0	–
MiMo-v2.5	0	–
Qwen-3.8 Max	0	–
GPT-5.6 Sol	2 (0.6%)	–
GPT-5.6 Luna	0	–

F Additional Turn-Budget Results

Tab. 8 reports the number of trajectories terminated by the interaction limit under the canonical 100-turn setting. Turn exhaustion is highly concentrated in DeepSeek and Gemini: 128 of 345 DeepSeek trials (37.1%) and 232 of 345 Gemini trials (67.2%) reach the maximum number of turns. In contrast, Grok and GPT-5.6 Sol terminate in only 10 and 2 trials, respectively, while no max-turn termination is observed for the remaining models.

We therefore rerun DeepSeek and Gemini on the same 115 tasks with the interaction budget increased from 100 to 300 turns. DeepSeek's Pass@3 increases from 27.0% to 33.9%, while its `max_turns` terminations decrease from 128 to 45 of 345 trials. Gemini's Pass@3 increases from 26.1% to 31.9%, with max-turn terminations decreasing from 232 to 64 trials. Despite these gains, both remain below GPT-5.6 Sol's 41.7% Pass@3 under the canonical 100-turn budget. These results show that turn exhaustion accounts for a meaningful fraction of failures for DeepSeek and Gemini, but increasing the interaction budget alone does not close the performance gap. The 300-turn reruns are reported separately and are not included in the canonical results in Tab. 2.

G Completion and Repeated-Trial Reliability

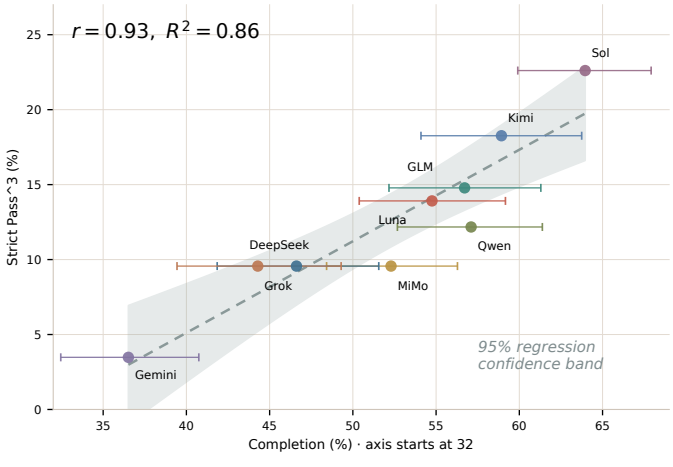

Figure 8: Completion versus Pass^3 across the nine canonical configurations. Horizontal whiskers show task-bootstrap 95% confidence intervals for completion; the shaded region shows the 95% confidence band for the fitted mean relationship.

Table 9: Visual-access ablation for Kimi-K3 on the same 115 tasks. Rows denote the multimodal outcome and columns the text-only outcome.

	Text-only	
Multimodal	Pass	Fail
Pass	19	**13**
Fail	0	83

Completion and Pass^3 are strongly associated across configurations (Pearson $r = 0.93$, Spearman $\rho = 0.93$; Fig. 8). Both metrics use the same scoring pipeline, and the fit summarizes their aggregate relationship. They capture different aspects of task outcomes: completion credits partial progress, while Pass^3 requires a passing result on every trial. Success coverage and repeatability also distinguish Qwen from Kimi: Qwen passes 43 tasks at least once and 14 on all three trials, whereas Kimi's corresponding counts are 39 and 21.

H Visual-Access Ablation

To assess whether successful workflow completion materially depends on visual access, we evaluate Kimi-K3 under a text-only condition on the same 115 tasks. The task instructions, environment, tools, execution harness, and interaction budget remain unchanged; the only intervention is that image inputs are withheld at the model interface throughout execution, while textual and structured observations remain available.

Removing visual access reduces Pass1 from 27.8% (32/115) to 16.5% (19/115), a decline of 11.3 percentage points. Importantly, the paired task outcomes change only in one direction: 13 of the 32 tasks solved with multimodal access fail under the text-only condition, while no task that fails with multimodal access becomes successful after images are removed. Thus, 40.6% of the workflows successfully completed with multimodal access are no longer solved without visual information. This provides direct evidence that visual access materially contributes to successful completion on a substantial subset of Agentic-WorkBench workflows.

I Human Validation of Task Graders

We conduct a small-scale human study to assess whether the task-specific graders agree with independent human judgments of task outcomes. We randomly sample 24 tasks from Agentic-WorkBench and, for each task, sample four agent outcomes from the evaluation runs, yielding 96 outcomes in total. Three human annotators independently evaluate each task without access to the model identity, grader score, or grader success label.

For each task, annotators are shown the original user request together with the final observable outcome needed to assess completion, including the produced artifact and/or resulting service state. They rank the four anonymized outcomes from best to worst according to how completely and correctly each outcome satisfies the user request, including both substantive and visual requirements. Ties are allowed when outcomes are judged indistinguishable. Annotators additionally provide a binary judgment of whether each outcome successfully completes the task. The evaluation focuses on final task outcomes rather than agent trajectories, matching the outcome-based grading protocol used in the benchmark.

We compare grader scores with the human consensus ranking using task-level Kendall's τ_b and pairwise ordering agreement, and measure agreement among human annotators using Kendall's W. For binary task success, we report the agreement rate between the grader and the majority human judgment together with Cohen's κ. As shown in Tab. 10, the graders show strong agreement with independent human judgments. The mean grader–human ranking correlation reaches $\tau_b = 0.70$, with 84.7% pairwise ordering agreement across outcome pairs. Binary success judgments agree on 86.5% of the 96 outcomes ($\kappa = 0.71$), while human annotators show high internal consistency with a

Table 10: Human validation of task-specific graders on 24 randomly sampled tasks. Three annotators independently evaluate four anonymized agent outcomes per task.

Tasks	Outcomes	Human W	Grader–human τ_b	Pairwise agree.	Success agree.	κ
24	96	0.81	0.70	84.7%	86.5%	0.71

mean Kendall's W of 0.81. These results indicate that the task-specific graders largely preserve both the relative ordering of outcome quality and the task-success judgments made by human evaluators.