Towards Real-World Evaluation of Agentic Work in Freelance Marketplaces

Mattie Terzolo[*]
Upwork Inc.
mattieterzolo@upwork.com

Teng Liu
Upwork Inc.
tedliu@upwork.com

Darvin Yi
Upwork Inc.
darvinyi@upwork.com

Karthik Gomadam
Upwork Inc.
karthikgomadam@upwork.com

Lance Hasson
Upwork Inc.
lancehasson@upwork.com

Ayan Sinha
Upwork Inc.
ayansinha@upwork.com

Pablo Mendes [†]
Upwork Inc.
pablomendes@upwork.com

Andrew Rabinovich
Upwork Inc.
andrewrabinovich@upwork.com

Abstract

Evaluating large language models (LLMs) on complex end-to-end digital work remains an open challenge. Many existing benchmarks are synthetic, static, or single-domain, limiting real world applicability and economic relevance. We present UpBench , a dataset and an evaluation pipeline derived from real tasks on Upwork . Starting from the marketplace corpus, we construct UpBench Qualified via heuristics-based filtering of fixed-price, single-milestone tasks and an automated feasibility assessment (Qualification Agent). We then derive UpBench Verified, a manually validated, PII-safe subset suitable for research use by the community. UpBench spans nine work categories and 572 unique task types, with tasks that resulted in an accepted deliverable and payouts ranging from $35 to $250 per job on average, enabling economically grounded and dynamically refreshable evaluation. We show initial results for several frontier LLMs on real-world Writing tasks, with human-in-the-loop experiments where agents iterate on their work based on human feedback. UpBench provides a practical, reproducible path to measure real-world progress while illuminating where current systems fall short.

1 Introduction

High-quality, economically-grounded datasets are essential for measuring real progress in AI systems. Much of today's LLM evaluation relies on synthetic tasks, static corpora, or narrow single-domain settings, which limits validity and obscures whether improvements translate into practical value. To advance beyond proxy tasks, we need datasets that are diverse across professional domains and tied to real transactions. Furthermore, by refreshing the dataset over time, we ensure the benchmark reflects evolving market demand as well as protecting against LLM memorization issues stemming from dataset leakage.

Upwork , an online work marketplace, offers a broad, longitudinal view of professional knowledge work. Its tasks span Creative fields like graphic design to technical work like Data Science and

[*]Technical correspondence

[†]General correspondence and participation in benchmark

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Analytics. The total list of 9 categories can be found in the Appendix. Each project represents an economic transaction with concrete deliverables and measurable payouts, providing a natural basis for dataset construction that captures both the complexity and the value of real work.

We introduce a data pipeline that converts the raw Upwork corpus into two progressively curated resources:

UpBench Qualified. A subset of Upwork tasks that pass heuristics-based filtering and an automated feasibility assessment that checks for features that we have found to correlate with higher likelihood of success for agents.

UpBench Verified. A further refined, manually verified subset of *UpBench Qualified*. Human verification ensures higher accuracy and stronger guarentees that the tasks are feasible for completion. This release-ready set will be made available to the research community.

2 Related Work

Table 1: Comparison of Existing Benchmarks

Benchmark	Tasks	Domains	Total Value	Task Horizon	Dynamic
UpBench Qualified	1,199	9	$101,695	short, medium, long	yes
UpBench Verified	201	9	$14,442	short, medium, long	no
SWE-Bench	2,294	1	N/A	short	no
SWE-PolyBench	2,110	1	N/A	short	no
SWE-Lancer	1,400	1	$1,000,000	short, medium	no
MLE-Bench	75	1	N/A	medium, long	no
InsightBench	100	1	N/A	long	no
REAL Bench	112	1	N/A	short, medium	no
MEGA-Bench	500	1	N/A	short, medium	no
PaperBench	20	1	N/A	medium, long	no

While numerous benchmarks have emerged to evaluate AI systems on real-world tasks and assess their limitations, existing approaches suffer from three critical flaws: static datasets, disconnect from economic reality, and narrow domain focus.

Static Datasets Most existing benchmarks rely on fixed datasets that cannot evolve with changing technology and market demands. SWE-Bench Jimenez et al. (2024) curates 2,294 GitHub issues, while SWE-PolyBench Rashid et al. (2025) extends to multiple languages but maintains the same static approach. MLE-Bench Chan et al. (2025) similarly freezes 75 Kaggle competitions. These static datasets fail to capture the evolving nature of real-world work, where new technologies, frameworks, and problem types emerge.

Disconnect from Economic Reality Many benchmarks rely on unrealistic evaluation scenarios with limited real-world insight. SWE-Bench Jimenez et al. (2024) and MLE-Bench Chan et al. (2025) evaluate historical tasks with no connection to market demand or payment. These approaches prevent assessing whether AI improvements translate to economic value. SWE-Lancer Miserendino et al. (2025) makes progress by incorporating $1 million worth of real Upwork tasks with actual payments, demonstrating economically grounded evaluation. However, its single-domain focus and static dataset limit broader applicability to the evolving freelancing landscape.

Single-Domain Focus Existing benchmarks evaluate AI systems within narrow domains. SWE-Bench focuses on software engineering, MLE-Bench targets machine learning, and specialized benchmarks like InsightBench and REAL Bench Garg et al. (2025) examine single areas. This creates asymmetry: models aspiring to broad capabilities are assessed through single-domain tests.

In contrast, UpBench 's tasks organically evolve with market demand, are category-diverse and representative of economic value. Releasing a new version of UpBench is a matter of re-sampling from the qualification pipeline.

3 Benchmark Construction

We transform the raw Upwork corpus into two progressively curated datasets: *UpBench Qualified* and *UpBench Verified*.

3.1 UpBench Qualified

As a first step, we apply the following filtering criteria to select tasks with higher likelihood of success for agents. We focus exclusively on successfully closed fixed-price tasks rather than hourly ones, as fixed-price work provides clear delivery expectations. We limit analysis to single-milestone tasks to reduce complexity and omit projects with significant price changes, which indicate scope changes from the original post. An example project can be found in the Appendix.

After heuristic filtering, we use an automated Qualification Agent to assess feasibility. The agent reads the job post, attachment contents, and any available deliverables, and renders pass/fail judgments on the following criteria:

1. **Task Completeness**: All necessary information to complete the job is either included in the attachments or fully described in the project description.
2. **Deliverable Quality**: Deliverables are accessible and representative of the work product delivered.
3. **No PII Present**: The attachments and job post do not contain any personally identifiable information (PII)

The agent is equipped with tools to open and parse typical attachment formats and is instructed to make conservative judgments when information is insufficient. The long tail distribution of filetypes found in the attachments is shown in the appendix. Tasks passing these criteria constitute the *UpBench Qualified* set.

3.2 UpBench Verified

From *UpBench Qualified*, we construct *UpBench Verified* via manual replication of the Qualification Agent's actions on a smaller subset. Human reviewers independently inspect the job post, attachments, and deliverables, verify feasibility, and check for PII. *UpBench Verified* will be made available to the research community.

4 UpBench Dataset

4.1 Comprehensive Task Coverage

UpBench demonstrates remarkable diversity across professional domains through systematic mapping of marketplace tasks to the O*NET database, identifying 572 unique task types across nine major work categories that capture the broad spectrum of knowledge work in modern labor markets. This encompasses professional services spanning from highly technical activities like data mining and machine learning algorithms to creative endeavors such as graphics and graphic design. To be successful on UpBench , AI systems must demonstrate competence in diverse domains rather than isolated technical skills. Note that while these tasks represent real work in modern labor markets, they only represent a small fraction of the economic value of Upwork tasks – namely, the sampling is limited to the tasks that passed the qualification pipeline. As such, the subset is not reflective of the distribution or the magnitude of Upwork 's task set.

4.2 Economic Grounding

A full breakdown of economic diversity and authenticity of our verified benchmark can be found in the Appendix in Table 3. While these figures reflect our verified benchmark data rather than the overall distribution of job amounts on Upwork as a whole, they nonetheless showcase genuine market transactions with real client investments ranging from $35.82 to $244.65 per job on average.

5 Experiments

We ran a limited benchmark on a subset of Writing category tasks to illustrate how UpBench supports evaluation of AI systems with both automated and human judgments. The evaluation process begins with the creation of objective acceptance criteria grounded in each job post and any attachments—these criteria serve as concise, verifiable checks designed to minimize subjective interpretation and enable consistent scoring across submissions. Worker Agents are lightweight scaffolds around frontier LLMs. They read the job context, draft simple plans, and generate deliverables—separating model behavior from complex tool orchestration. Outputs are evaluated in two steps. First, they are scored automatically against the rubric, with each criterion marked pass, fail, or skip, and results aggregated into interpretable feedback. Second, human experts review the same outputs. A comparison of the two—summarized in the Appendix (Table 4)—validates the automated evaluation process.

5.1 Agent Performance Results

We evaluate several frontier LLMs using identical scaffolding. For each task, we create both AI-only submissions and Human-in-the-Loop (HITL) submissions. In the HITL process: (1) the AI agent makes an initial attempt at the task, (2) a human evaluator grades the submission and provides detailed feedback based on the evaluation rubric, (3) the AI agent receives its previous attempt, the human feedback, and the original job post to make a second attempt, and (4) the final submission is graded again. Table 2 reports average success rates and token usage for both AI-only and HITL conditions. Pass rates represent the percentage of submissions that meet all of the job's acceptance criteria as evaluated by human annotators. The cost reported in the HITL column accounts for the time the human annotator took to review the initial submission and provide feedback.

Table 2: Agent Performance Results

	Avg Pass Rate		Median Duration (s)		Avg Cost	
Agent	AI-only	HITL	AI-only	HITL	AI-only	HITL
Claude 3.5 Haiku	19.0%	28.6%	116.1	604.4	\$0.04	\$5.08
Claude Sonnet 4	26.2%	**47.6%**	188.3	745.0	\$0.46	\$5.93
Gemini 2.5 Flash	26.2%	38.1%	**93.2**	**521.0**	\$0.04	\$5.09
Gemini 2.5 Pro	**31.0%**	42.9%	229.6	783.3	\$0.10	\$5.25
GPT-4.1-Mini	19.0%	23.8%	98.7	555.9	\$0.07	\$5.14
Kimi K2 Instruct	28.6%	35.7%	240.1	818.9	\$0.04	\$5.09
o3	26.2%	35.7%	133.7	627.9	\$0.13	\$5.27
Qwen3 235B A22B	23.8%	38.1%	121.9	597.5	**\$0.03**	**\$5.06**

6 Conclusion

UpBench links real marketplace tasks to a dynamic, multi-domain benchmark with explicit, verifiable acceptance criteria. It is economically grounded (real payouts), spans nine work categories, preserves authentic attachment and deliverable formats, and supports longitudinal refresh. The Qualified→Verified pipeline yields a research-ready, PII-safe subset while retaining task diversity and complexity. Rubric-based scoring provides interpretable signals that generalize across domains and enable reproducible generator/validator evaluation. The benchmark can be refreshed with new and current tasks as it is connected with the liver Upwork platform.

Limitations Key limitations are: (1) Qualification Agent performance and robustness – ambiguous scope, mixed-format attachments, and context limits can cause false positives/negatives; (2) Human verification throughput – manual replication and PII review are rate-limiting, slowing refresh cadence; (3) Privacy considerations — even after sanitization, residual risk necessitates conservative release policies and ongoing audits. We also note potential selection bias from heuristics-based filtering and uneven category coverage in this initial release.

Data Availability The PII-safe UpBench Verified subset will be made available to the research community. Access requires a data use agreement and adherence to privacy safeguards; we provide documentation, schema, and evaluation scripts to facilitate replication. Details of the application process and update cadence will be provided to approved partners.

References

[1] Chan, J. S., Chowdhury, N., Jaffe, O., Aung, J., Sherburn, D., Mays, E., Starace, G., Liu, K., Maksin, L., Patwardhan, T., Weng, L., and Mądry, A. (2025). Mle-bench: Evaluating machine learning agents on machine learning engineering.

[2] Chen, J., Liang, T., Siu, S., Wang, Z., Wang, K., Wang, Y., Ni, Y., Zhu, W., Jiang, Z., Lyu, B., Jiang, D., He, X., Liu, Y., Hu, H., Yue, X., and Chen, W. (2024). Mega-bench: Scaling multimodal evaluation to over 500 real-world tasks.

[3] Chen, M., Tworek, J., Jun, H., Yuan, Q., and et al. (2021). Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

[4] Garg, D., VanWeelden, S., Caples, D., Draguns, A., Ravi, N., Putta, P., Garg, N., Abraham, T., Lara, M., Lopez, F., Liu, J., Gundawar, A., Hebbar, P., Joo, Y., Gu, J., London, C., de Witt, C. S., and Motwani, S. (2025). Real: Benchmarking autonomous agents on deterministic simulations of real websites.

[5] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, 63(11):139–144.

[6] Gunjal, A., Wang, A., Lau, E., Nath, V., Liu, B., and Hendryx, S. (2025). Rubrics as rewards: Reinforcement learning beyond verifiable domains. *arXiv preprint arXiv:2507.17746*.

[7] Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N. V., Wiest, O., and Zhang, X. (2024). Large language model based multi-agents: A survey of progress and challenges.

[8] Jimenez, C. E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., and Narasimhan, K. (2024). Swe-bench: Can language models resolve real-world github issues?

[9] Kwa, T., West, B., Becker, J., Deng, A., Garcia, K., Hasin, M., Jawhar, S., Kinniment, M., Rush, N., Arx, S. V., Bloom, R., Broadley, T., Du, H., Goodrich, B., Jurkovic, N., Miles, L. H., Nix, S., Lin, T., Parikh, N., Rein, D., Sato, L. J. K., Wijk, H., Ziegler, D. M., Barnes, E., and Chan, L. (2025). Measuring ai ability to complete long tasks.

[10] Ming, Y., Purushwalkam, S., Pandit, S., Ke, Z., Nguyen, X.-P., Xiong, C., and Joty, S. (2025). Faitheval: Can your language model stay faithful to context, even if "the moon is made of marshmallows".

[11] Miserendino, S., Wang, M., Patwardhan, T., and Heidecke, J. (2025). Swe-lancer: Can frontier llms earn $1 million from real-world freelance software engineering?

[12] Ott, S., Barbosa-Silva, A., Blagec, K., Brauner, J., and Samwald, M. (2022). Mapping global dynamics of benchmark creation and saturation in artificial intelligence. *Nature Communications*, 13(1):6793.

[13] Rashid, M. S., Bock, C., Zhuang, Y., Buchholz, A., Esler, T., Valentin, S., Franceschi, L., Wistuba, M., Sivaprasad, P. T., Kim, W. J., Deoras, A., Zappella, G., and Callot, L. (2025). Swe-polybench: A multi-language benchmark for repository level evaluation of coding agents.

[14] Rawles, C., Clinckemaillie, S., Chang, Y., Waltz, J., Lau, G., Fair, M., Li, A., Bishop, W., Li, W., Campbell-Ajala, F., Toyama, D., Berry, R., Tyamagundlu, D., Lillicrap, T., and Riva, O. (2025). Androidworld: A dynamic benchmarking environment for autonomous agents.

[15] Sahu, G., Puri, A., Rodriguez, J., Abaskohi, A., Chegini, M., Drouin, A., Taslakian, P., Zantedeschi, V., Lacoste, A., Vazquez, D., Chapados, N., Pal, C., Mudumba, S. R., and Laradji, I. H. (2025). Insightbench: Evaluating business analytics agents through multi-step insight generation.

[16] Starace, G., Jaffe, O., Sherburn, D., Aung, J., Chan, J. S., Maksin, L., Dias, R., Mays, E., Kinsella, B., Thompson, W., Heidecke, J., Glaese, A., and Patwardhan, T. (2025). Paperbench: Evaluating ai's ability to replicate ai research.

[17] Tran, K.-T., Dao, D., Nguyen, M.-D., Pham, Q.-V., O'Sullivan, B., and Nguyen, H. D. (2025). Multi-agent collaboration mechanisms: A survey of llms.

[18] Wang, W., Alyahya, H. A., Ashley, D. R., Serikov, O., Khizbullin, D., Faccio, F., and Schmidhuber, J. (2024). How to correctly do semantic backpropagation on language-based agentic systems. *arXiv preprint arXiv:2412.03624*.

[19] Wijk, H., Lin, T., Becker, J., Jawhar, S., Parikh, N., Broadley, T., Chan, L., Chen, M., Clymer, J., Dhyani, J., Ericheva, E., Garcia, K., Goodrich, B., Jurkovic, N., Karnofsky, H., Kinniment, M., Lajko, A., Nix, S., Sato, L., Saunders, W., Taran, M., West, B., and Barnes, E. (2025). Re-bench: Evaluating frontier ai r&d capabilities of language model agents against human experts.

[20] Xiao, Y., Sun, E., Luo, D., and Wang, W. (2025). Tradingagents: Multi-agents llm financial trading framework.

[21] Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., and Cao, Y. (2023a). React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

[22] Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., and Cao, Y. (2023b). React: Synergizing reasoning and acting in language models.

[23] Yuksekgonul, M., Bianchi, F., Boen, J., Liu, S., Huang, Z., Guestrin, C., and Zou, J. (2024). Textgrad: Automatic" differentiation" via text. *arXiv preprint arXiv:2406.07496*.

A Appendix

A.1 List of 9 Task Categories in UpBench

1. Accounting & Consulting
2. Admin Support
3. Data Science & Analytics
4. Design & Creative
5. Engineering & Architecture
6. Sales & Marketing
7. Translation
8. Web, Mobile & Software Dev
9. Writing

A.2 Example of Project

Each record in the dataset is made up of two parts: the attachment directory and the project specification. An example specification appears below; attachment and deliverable directories mirror real client-provided context and freelancer outputs.


```
{
  "job_title": "Quick job just need debugging smtp script with gmail
     accounts",
  "job_description": "Hi I have a script running on my linux box that
     should work, but the developer is not responding to my emails.
     The script is attached as a txt file but its .py. The attached
     file should show you what I am looking to do.",
  "job_amount": "50.00",
  "expertise_tier": "INTERMEDIATE",
  "category": "Web, Mobile & Software Dev",
  "subcategory": "Scripts & Utilities",
  "subsubcategory": "Scripting & Automation",
  "attachments": [{"file_name": "main.py.txt", "file_size": 5058}]
}
```


A.3 Attachments and Deliverables Directories

The attachments directory contains files that clients attach to job posts for additional context, including various file types from PDFs and CSVs to .dwg and .epub files, typically 1–3 documents but potentially up to 30. The deliverables directory contains freelancer-submitted work products with similar file type distributions.

A.4 Temporal Distribution of Dataset

A.5 Economic Value Captured in UpBench Verified

A.6 Time Horizons

UpBench captures the full spectrum of professional work duration, with project completion times ranging from single days to over 90 days, measured as elapsed time between project initiation and successful completion (Figure 3 in the Appendix). Unlike existing benchmarks that predominantly focus on short-duration problems solvable within hours, our dataset encompasses tasks spanning

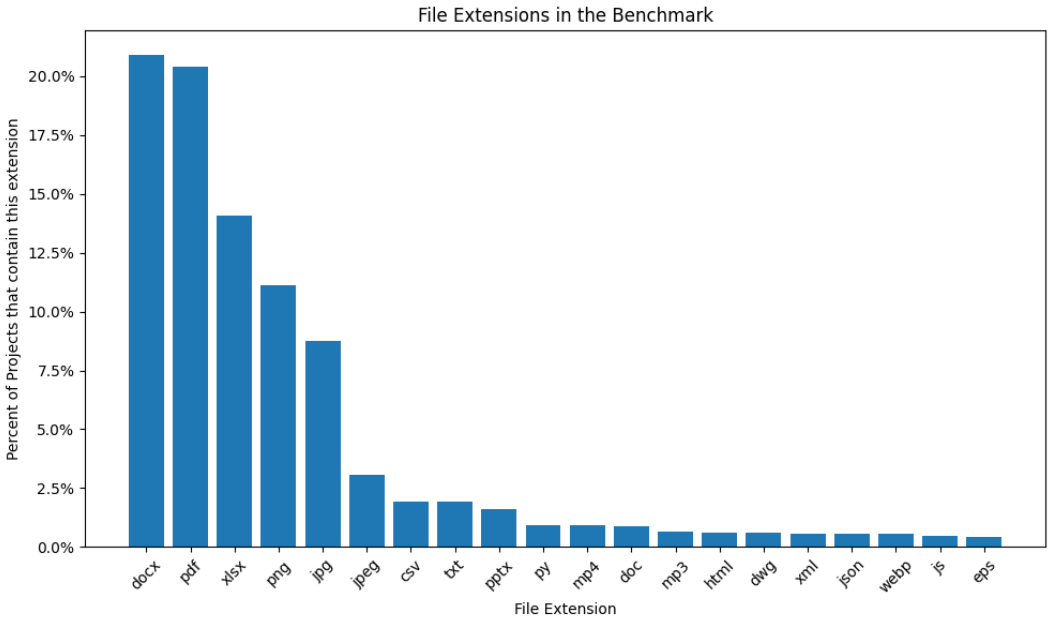

Figure 1: Long-tail distribution of file extensions observed across attachments and deliverables.

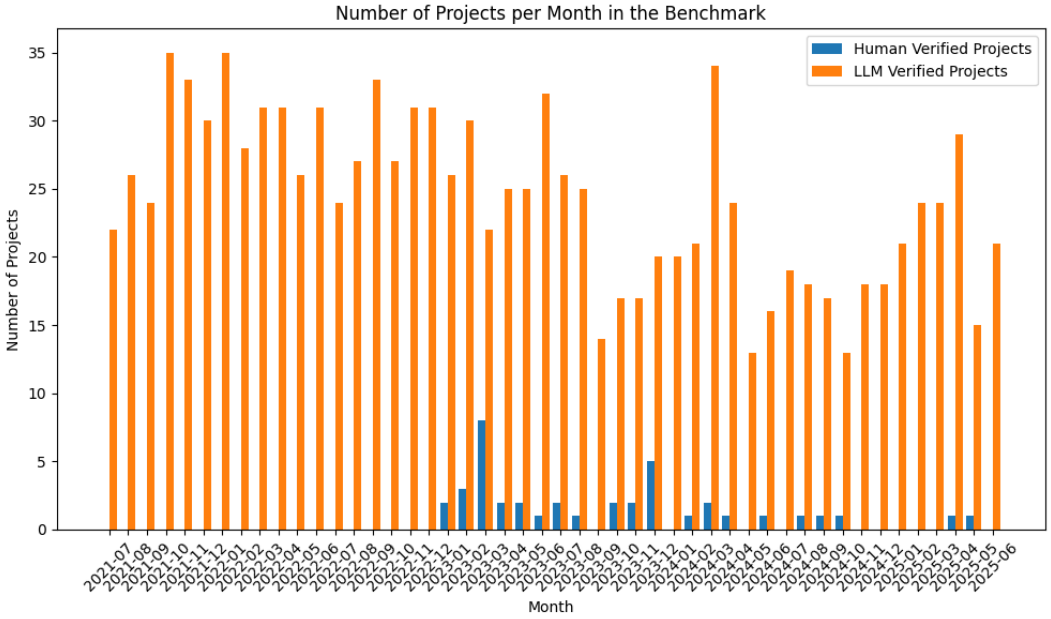

Figure 2: Number of Human Verified and LLM Verified Examples in the Benchmark

from rapid turnaround tasks (1–7 days) to complex, long-term engagements (31+ days) that require sustained reasoning, iterative development, and comprehensive deliverable creation.

Long time horizon tasks are particularly critical for AI evaluation as they correlate strongly with higher economic value, test sustained execution capabilities essential for professional work, and represent the strategic initiatives that drive real-world economic impact. This temporal diversity ensures that AI progress measurements reflect not just problem-solving speed, but the sustained professional competence required for meaningful economic contribution in knowledge work domains.

Table 3: Average and Total Payouts by Job Category

Category	Average Payout per Job	Total Payout in Benchmark
Accounting & Consulting	$119.78	$2,755
Admin Support	$53.50	$3,424
Data Science & Analytics	$244.65	$11,743
Design & Creative	$59.28	$4,387
Engineering & Architecture	$128.89	$5,800
Sales & Marketing	$85.90	$1,804
Translation	$35.82	$2,364
Web, Mobile & Software Dev	$69.22	$3,115
Writing	$100.55	$4,424

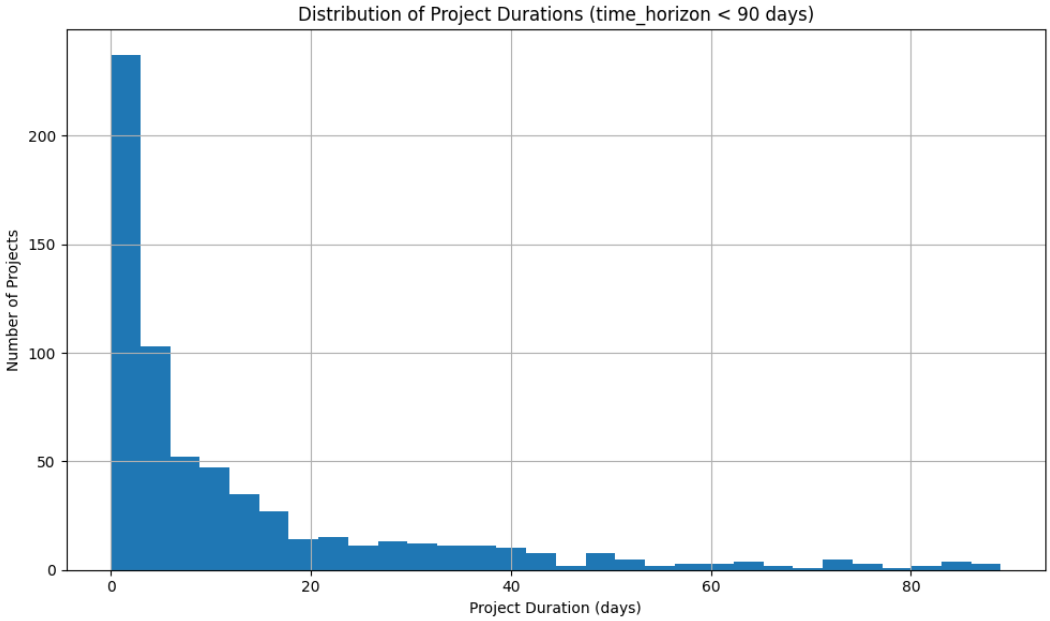

Figure 3: Number of Human Verified and LLM Verified Examples in the Benchmark

A.7 Additional Experimental Results

A.7.1 Evaluation Agent Agreement with Human Annotators

Table 4: Evaluation Agent Agreement with Human Annotators

Granularity	Precision	Recall
Final Judgment	0.5555	0.7612
By Acceptance Criteria	0.8333	0.8428

A.7.2 Model Performance by Qualification Status

Table 5: Model Performance by Qualification Status

Model	Success Rate (Non-Qualified)	Success Rate (Qualified)	tasks (Non-Qual./Qual.)
Anthropic Claude Sonnet 4	25.00%	**27.27%**	20/44
Anthropic Claude 3.5 Haiku	15.00%	**20.45%**	20/44
Google Gemini 2.5 Flash	10.00%	**27.27%**	20/44
Google Gemini 2.5 Pro	20.00%	**31.82%**	20/44
OpenAI GPT-4.1-Mini	15.00%	**20.45%**	20/44
OpenAI o3	20.00%	**25.00%**	20/44
Kimi K2 Instruct	30.00%	**29.55%**	20/44
Qwen3 235B A22B Instruct	15.00%	**22.73%**	20/44

A.7.3 Common Failure Modes

Our evaluation revealed several distinct failure modes across different LLM-based worker agents when completing benchmark tasks. These failures highlight fundamental limitations in current language models' ability to follow complex instructions, manage scope constraints, and provide appropriate levels of assistance.

Incomplete Scope Fulfillment When tasks specified exact quantities or comprehensive requirements, worker agents frequently delivered partial results. For example, tasks requesting 13 blog posts would often result in only 4 completed posts, suggesting difficulties in maintaining task scope awareness throughout longer generation processes.

Information Gap Hallucination When provided with incomplete source material, worker agents consistently chose to fabricate missing details rather than acknowledging information gaps. This behavior prioritized deliverable completion over factual accuracy, leading to plausible but incorrect content that could mislead clients.

Over-Completion vs. Collaborative Feedback In tasks requiring editorial review or suggestions (such as proofreading), worker agents frequently bypassed the collaborative intent by directly implementing changes rather than providing feedback. This pattern suggests difficulty distinguishing between completion-focused tasks and advisory roles.

A.8 Top Tasks by Category

Figure 4 presents a comprehensive hierarchical breakdown showing the nine major work categories with their relative prevalence, common job titles, and top associated tasks, creating a diverse testbed for evaluating AI capabilities across multiple dimensions of professional work.

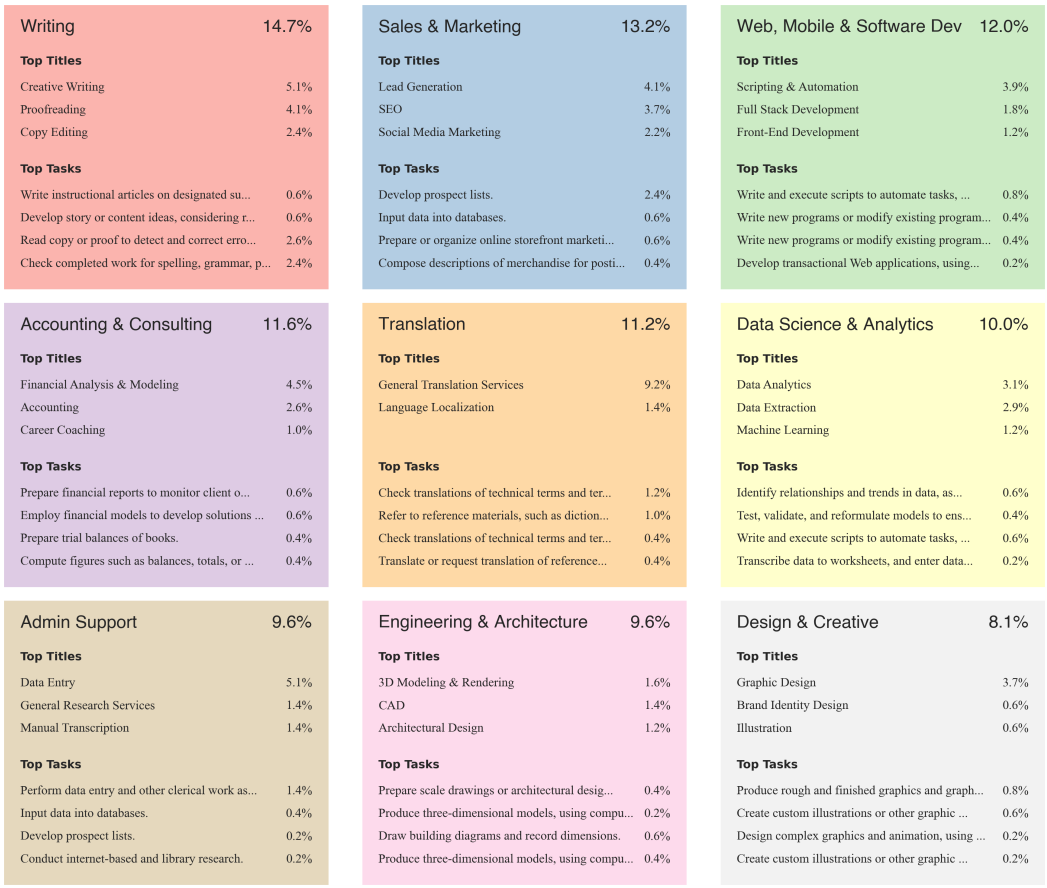

Figure 4: Top three titles and top 4 tasks for each cateogy of work