Tendem: A Hybrid AI+Human Platform

Toloka Team

November 2025

Abstract

Tendem is a hybrid system where AI handles structured, repeatable work and Human Experts step in when the models fail or to verify results. Each result undergoes a comprehensive quality review before delivery to the Client.

To assess Tendem's performance, we conducted a series of in-house evaluations on 94 real-world tasks, comparing it with AI-only agents and human-only workflows carried out by Upwork freelancers.

The results show that Tendem consistently delivers higher-quality outputs with faster turnaround times. At the same time, its operational costs remain comparable to human-only execution.

On third-party agentic benchmarks, Tendem's AI Agent (operating autonomously, without human involvement) performs near state-of-the-art on web browsing and tool-use tasks while demonstrating strong results in frontier domain knowledge and reasoning.

1 Introduction

We introduce Tendem[1] – a hybrid execution platform that joins AI execution with Human expertise that streamlines both professional and personal workflows. Tasks are completed through collaboration between the AI, which manages routine and fast-paced work, and a Human Expert who provides oversight, ensures contextual accuracy, and refines outputs. If the AI reaches its limits, an expert continues the work to bring the task to completion. The platform's architecture includes a multi-layered pipeline for task orchestration, routing, and quality assurance that integrates automated reasoning with human validation.

Unlike fully automated agentic workflows – e.g., agents embedded in widely used assistants such as *ChatGPT*[2] and *Claude*[3] – Tendem *deliberately* retains a human in the loop for uncertainty- or impact-heavy steps. Purely autonomous agents tend to optimize for speed and cost but can struggle with under-specification, multi-document synthesis, or non-local quality criteria.

By contrast, human-only workflows, such as freelancer platforms, might offer traceable work history ratings and accountability but tend to be slower and more variable in cost. Tendem targets efficient points where it outperforms AI-only systems on quality for ambiguous or compliance-sensitive tasks while improving time-to-result and coordination cost relative to human-only execution.

In this work, we present results from a comparative study involving fully automated pipeline, human-only workflow, and the Tendem hybrid model. We observe a substantial improvement in result quality and consistency: Tendem converts a much larger share of tasks into client-ready outcomes than either AI-only or human-only workflows. In addition, Tendem reduces operational cost relative to human-only execution, so higher quality does not come at the expense of price.

The evaluation comprises two components: an in-house set of 94 real-world tasks and a suite of public agentic benchmarks. We measure result quality (human evaluation), time, and price.

Tendem achieves 74.5% high-quality results across tasks and reduces median delivery time by 53% versus the human-only baseline. We also evaluate Tendem's underlying AI agent (without human involvement) on external benchmarks. The agent is close to state-of-the-art on web browsing and real-world assistant tasks – and remains close to leading models on hard knowledge. This provides a solid backbone for our Tendem Agent.

[1] `https://tendem.ai`
[2] `https://chat.openai.com`
[3] `https://www.anthropic.com/claude`

2 System Overview

Tendem's architecture is built around the interaction between AI and Human Expertise. Each task moves through a structured flow that defines when the AI acts independently and when human judgment is required to ensure an efficient workflow and high-quality deliverables. Checks and verifications are built-in at every stage, so small errors do not accumulate, and human attention is directed to the points where it makes the greatest difference.

2.1 Roles

We focus on the following three roles when designing the system:

Client / End User

The **Client** begins by outlining the problem and describing desired outcomes, sharing any background material or data needed to work effectively. Clients expect results that meet their standards, within a limited time-frame, and at a consistent cost. But just as importantly, they expect a clear view of how their data is handled and protected at every stage.

Automated Component: Tendem's AI Agent and Tools

Tendem's AI Agent executes the bulk of routine and tool-heavy work via a plan–act–observe–verify loop with explicit step gates for verification (Yao et al., 2023; Weng et al., 2023; Manakul et al., 2023). Tooling includes *web browsing*, *file I/O* for common office formats, a safe and isolated *Python* runtime for analytics and charting, and an isolated secure runtime environment *bash/CLI*. All actions run with minimal permissions and are recorded to make every step traceable and repeatable.

Quality assurance tools run continuously: spec conformance, unit/total reconciliation, citation matching, and lightweight self-consistency checks. When the system encounters uncertainty – whether from conflicting sources, failed checks, or a step that carries too much risk – it escalates to a Human Expert.

Human Expert

A **Human Expert** is a professional who adds the kind of reasoning and contextual awareness that models cannot yet provide (Huang et al., 2025; Ji et al., 2023; Bender and Koller, 2020). They guide the system through uncertain steps, verifying, and refining the final result for real-world use. Experts are admitted through a series of tests and exams to to verify their expertise and ensure they undergo training and continuous QA.

In production workflow, Experts intervene at step gates (plan audit, plan step check, draft refinement, etc.) and perform the final offline QA pass as part of the layered-QA. The performance is tracked using QA and rework rates.

2.2 End-to-End Workflow

Tendem workflow consists of the following interconnected steps:

- **Client request**: the Client outlines what they want to achieve in plain language, adding any files or references the system will need to understand the task.
- **Clarify and formalize**: the AI Agent inspects files and requirements and asks targeted questions.
- **Plan with gated steps**: the AI Agent decomposes the task into steps, gating critical and high-risk ones under Human Expert supervision.
- **Routing and matching**: the system identifies a suitable expert based on the required skill set and generates a time estimate.
- **Hybrid execution**: the AI agent carries out the work under human supervision, with the Human Expert stepping in whenever the output needs adjustment or further development.
- **Online QA**: the system performs lightweight automated checks of both AI Agent and human edits, updating the plan and iterating further steps if needed.

- **Offline QA**: after task execution, the system conducts automated multi-step verification against the customer's requirements and attached materials, escalating uncertain cases to a human QA expert.
- **Finalization**: the system delivers the verified result to the Client or returns it for rework if material issues remain.

2.3 Target Advantages

Having a multi-component workflow helps us achieve several key advantages:

(1) **Intelligent Hybrid Execution:** Task decomposition and dynamic routing (AI for speed, Human Expert for judgment, and a verification loop between them) reduce coordination overhead and ensure every deliverable benefits from both machine efficiency and expert human judgment.

(2) **Multi-Layer Quality Assurance:** Each deliverable is first tested through automated systems that flag potential faults or signs of hallucination. From there, a human reviewer traces how the output was produced, checking its logic and alignment with the source material and requirements until it meets the standard expected for delivery.

(3) **Expert productivity uplift.** Routine subtasks are automated, allowing Human Experts to concentrate on high-impact decisions and polish, which shortens total cycle time and increases attention to detail.

(4) **Improved Client experience.** The system designed to have clear structured flow, targeted clarifications and visible progress, to make working with it feels straightforward and efficient.

3 Evaluation

To quantify price–quality trade-offs, we evaluate Tendem on in-house collected, economically valuable tasks that mirror real freelance platforms. We compare Tendem against AI agents and a freelance-marketplace worker baseline (subsection 3.1). This measures client-perceived outcomes on valuable real-life tasks.

To ensure transparent comparison, we also evaluate Tendem's underlying AI agent (without human intervention) on public benchmarks against existing models and AI-only agents (subsection 3.2). We use it as a proxy for assessing the strength of the backbone AI agent and its impact on the hybrid system's overall efficiency and performance.

3.1 In-house benchmark with real-world tasks

Public benchmarks probe core skills but underrepresent real-life settings that depend on a Client's personal preferences (ambiguous briefs, complex multi-stage tasks, input and files as deliverables).

To better assess system performance on *real-world, economically valuable remote-work projects* that reflect the complexity and practical constraints of professional task execution, we introduce a new in-house benchmark. It approximates real freelance-platform task distributions to mirror the conditions of professional work.

We assume that customers evaluate performance and successful task completion across three independent dimensions: **result quality**, **execution time**, and **price**.

Metrics

To compare Tendem's end-to-end performance against other AI agents and a freelance-platform baseline, we measure the following for each task:

1. **Result Quality**: each output is evaluated on a three-point scale (Bad, Mediocre, Good), with additional Decline label. We assess quality across three fine-grained quality criteria and overall result quality.

   - **Accuracy** – factual and numerical correctness, with no fabrications, sources, or outdated information; the presence of a reference does not compensate for an inaccuracy, it should be relevant.
   - **Completeness** – coverage of all required items, constraints, and deliverables in the Client's task so the result is self-contained and usable without follow-up, no items are missed, no part of the processing omitted.

- **Style & Formatting** – clarity, structure, and presentation quality, including readable organization (headings, lists, tables), consistent units and notation, correct grammar/spelling, appropriate tone, and correct output format.
- **Overall Quality** – end-to-end fitness for purpose: the answer is aligned with the user's intent and ready to deliver given the three criteria above (assigned independently, to weigh contextual factors and practical usefulness).

Each criterion is graded using the following scale:

- **Good** — full criteria satisfaction, client-ready; at most nitpicks that do not require changes.
- **Mediocre** — requires some edits to satisfy the criteria.
- **Bad** — materially incomplete; not suitable without substantial rework.
- **Decline** — task refused or not attempted due to safety filters or domain coverage limits.

2. **Time** (latency): measured as connect time (if any) from task posting to active work, execution time to final result, and total time as the sum of both.

3. **Price**: the effective cost of each task, calculated as the sum of human labor and backbone AI usage, recorded in USD.

Evaluated Systems

For evaluations, we selected commonly used AI-only system and human-only system. The following were evaluated:

1. **ChatGPT Agent (OpenAI)[4].** A multi-step agentic assistant embedded in ChatGPT with actions/tool use and browsing. Evaluated via the ChatGPT UI with default settings.

2. **Upwork[5].** Human freelancers hired through the Upwork marketplace who follow principles described below (best-match, price, job-success filter). Freelancers are hired independently for each task. We did not instruct them to use or not to use AI.

3. **Tendem[6].** The production version of Tendem. The detailed system overview can be found in section 2.

Dataset Design and Collection

Building a representative, real-life benchmark of Client tasks involved the following structured process:

- **Use-case sampling.** Common tasks categories were selected from Upwork[5], Fiverr[7], and Freelancer[8] to approximate the distribution of paid knowledge work.
- **Task authoring.** Based on the sampled categories, we commissioned practitioners to author new tasks with clear acceptance criteria and attached assets (e.g., spreadsheets, PDFs, links).
- **PII and compliance.** Tasks were created to be **PII-free** and avoid regulated domains (e.g., medical/legal advice).

The final benchmark area/category distribution is presented in Table 1.

Task submissions protocol

All third-party systems were used via their public UIs – hand-pasting task text – to reflect the end-user experience and capture real connection latency. The highest subscription tier is used where available to ensure the highest rate limits and best backbone models. We captured timestamps when each task was created, when human work started, and when the final result was delivered.

[4] https://openai.com/index/introducing-chatgpt-agent/
[5] https://www.upwork.com/
[6] https://tendem.ai/
[7] https://www.fiverr.com/
[8] https://www.freelancer.com/

Area	Category	Count
Sales	Collect Business Contact Data	14
	Complete Missing Fields (enrichment)	6
	Sales Total	**20**
Operations	Build Multi-step Automation Workflows	1
	Collect Data	10
	Convert Formats	2
	Retrieve PDF / Document / Report Content	3
	Schedule & Manage Appointments & Calls	3
	Structure Raw Data into Schema	6
	Validate Contact Info	3
	Operations Total	**28**
Marketing	Collect Business Contact Data	4
	Create Content	7
	Market & Competitive Research Reports	10
	Proofread, analyse and correct content	3
	Marketing Total	**24**
Analysis	Customer / User Interviews or Feedback Collection	1
	Generate Performance Dashboards & Summaries	8
	Market & Competitive Research Reports	8
	Run Exploratory Data Analysis	5
	Analyst Total	**22**
Grand Total		**94**

Table 1: In-house benchmark area and category distribution.

Freelancer Selection at Upwork

Because freelance platforms offer experts with varying rates and skill levels, the following principles were used to guide selection to ensure fair comparison.

Tasks were posted and proposals collected; selection prioritized previous human experience, price, job success percentage, using information from the Upwork profile of applicants. We preferred the least expensive candidate with over 80% job success or promising newcomers. The performer was selected without price negotiation – bids were evaluated and the lowest acceptable offer was accepted as-is.

Labeling Protocol

We recruited independent human QA experts from our professional network and onboarded them, requiring passage of an entrance exam. Raters received concise instructions on how to apply the evaluation criteria (Accuracy, Completeness, Style & Formatting, Overall) and example-driven guidance. To reduce bias, *the systems were anonymized* and evaluated in the same interface. The evaluators saw only the input task, attached assets, and the produced outputs. All experts received a competitive hourly payment.

For each task, a single Expert assigned labels (*Good, Mediocre, Bad* for all 4 quality scales, and brief notes justifying the decision. We record a *Decline* when a system refused to complete the task. For this experiment, we used single-rater judgments per task; but a large fraction of items received spot checks to ensure high quality labeling.

We did not employ or validate LLM-as-a-judge (Gu et al., 2025) for scoring, so all results rely on human evaluation only (yet it can be a good follow-up experiment).

Results

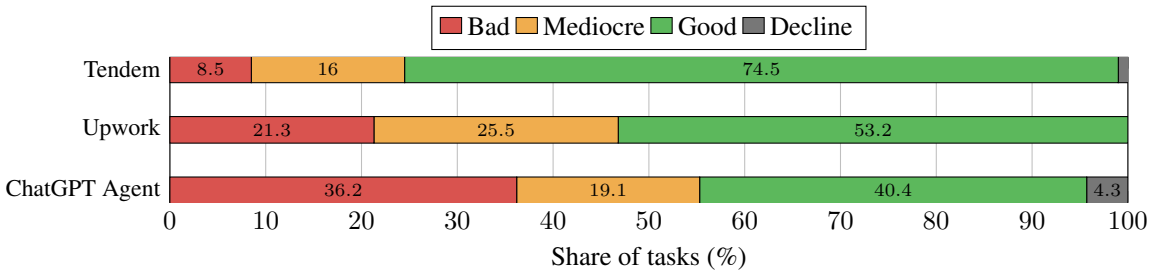

Figure 1: The fraction of tasks rated Bad/Mediocre/Good for each system on in-house benchmark. Each system evaluated by a human without system name indication. Decline label means safety filters of the system declined the task.

Systems Evaluation. Figure 1 summarizes the distribution of Bad/Mediocre/Good outcomes for each evaluated system. Tendem attains 74.5% Good, outperforming Upwork by +21.3 pp (53.2%) and ChatGPT Agent by +34.1 pp (40.4%).

Tendem also exhibits fewer weak outcomes: 8.5% Bad vs 21.3% (Upwork), and 36.2% (ChatGPT Agent); showing itself as a more stable and reliable performer.

Overall, **Tendem's Good rate is significantly higher than all baselines** (Upwork and ChatGPT Agent) on our in-house evaluation. A one-sided z-test comparing the share of 'Good' results between Tendem and the closest baseline Upwork confirms the improvement is statistically significant (p-value $= 0.0012$).

System	Overall				Accuracy	Completeness	Style
	(% Good)	(% Mediocre)	(% Bad)	(% Decline)	(% Good)	(% Good)	(% Good)
Tendem	**74.5** $\pm$ 4.5	16.0 $\pm$ 3.8	8.5 $\pm$ 2.9	1.1 $\pm$ 1.0	**74.5** $\pm$ 4.5	**81.9** $\pm$ 4.0	**70.2** $\pm$ 4.7
Upwork	53.2 $\pm$ 5.1	25.5 $\pm$ 4.5	21.3 $\pm$ 4.2	0.0 $\pm$ 0.0	63.8 $\pm$ 5.0	59.6 $\pm$ 5.1	59.6 $\pm$ 5.1
ChatGPT Agent	40.4 $\pm$ 5.1	19.1 $\pm$ 4.1	36.2 $\pm$ 5.0	4.3 $\pm$ 2.1	48.9 $\pm$ 5.2	48.9 $\pm$ 5.1	51.1 $\pm$ 5.2

Table 2: The fraction of tasks rated Bad, Mediocre, and Good for each system on the in-house benchmark for overall and three quality criteria. Decline label means system declined the task due to safety filters or lack of expertise for the task. Each system was evaluated by a human without system name indication. Bold font indicates best result in column.

Analysis of Quality Criteria. In Table 2 we go beyond overall task evaluation and analyze performance across three quality criteria:

- *Accuracy.* Tendem improves Accuracy vs. Upwork by +10.7 pp (74.5% vs 63.8%), indicating that designed hybrid workflow and layered QA strengthens factual grounding and numerical correctness.
- *Completeness.* Largest gap: +22.3 pp vs Upwork (81.9% vs 59.6%). One of the reasons might be a workflow with step-gates that reduce omissions and enforce acceptance criteria.
- *Style & Formatting.* Tendem gains +10.6 pp vs Upwork (70.2% vs 59.6%) and +19.1 pp vs ChatGPT Agent (70.2% vs 51.1%), yielding cleaner, client-ready presentation.

Error patterns. AI-only systems exhibit (i) omissions on long specs, (ii) shallow synthesis beyond model context limits, (iii) fabricated references (Ji et al., 2023; Huang et al., 2025). Human-only outputs show uneven quality along all criteria. Tendem substantially cuts omission/fabrication errors via escalation at step gates with high uncertainty.

System	Price (USD)		Time (hours)					
			Connection		Execution		Total	
	Average	Median	Average	Median	Average	Median	Average	Median
Tendem	69.2 ± 9.3	32.0 ± 5.3	4.8 ± 5.7	2.6 ± 0.6	20.9 ± 24.3	10.5 ± 3.8	25.7 ± 24.3	16.4 ± 2.9
Upwork	48.0 ± 3.3	50.0 ± 3.8	14.5 ± 27.4	4.6 ± 0.5	38.3 ± 54.3	26.8 ± 3.9	52.7 ± 58.8	35.0 ± 6.8
ChatGPT Agent	–	–	0	0	0.2 ± 0.2	0.1 ± 0.01	0.2 ± 0.2	0.1 ± 0.01

Table 3: Price and timing metrics with standard deviation shown as ± std. Prices are in USD; times are in hours. For median values std is calculated using bootstrap. ChatGPT Agent is subscription-based service, hence no direct price.

Price and Speed. Table 3 provides execution time (split by *connection* and effective *execution* time) and price analysis per task.

Tendem's median price is 32.0 USD versus Upwork's 50.0 USD (-36%); its average price is higher at 69.2 USD versus 48.0 USD for Upwork, reflecting a right-tailed cost distribution with a few high-touch cases.

When it comes to time, Tendem reduces median total time from 35.0 h (Upwork) to 16.5 h (-53%), with faster connection (2.6 h versus 4.6 h, -42%) and execution (10.5 h versus 26.8 h, -61%). ChatGPT Agent exhibits very low latency and low price as expected, but – as shown in Table 2 – delivers lower quality on this dataset.

Overall, **Tendem delivers results significantly faster than the human-only baseline and at a lower median price, with higher quality**. The higher average price reflects a small number of complex, escalated tasks that required additional expert involvement.

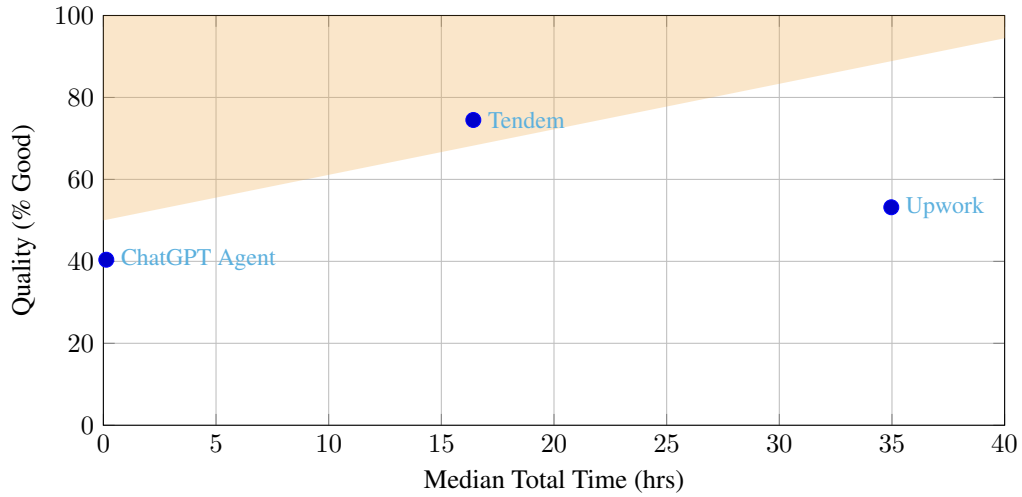

Figure 2: Evaluated system quality vs median Total Time. Light-orange corner highlights the desirable region (lower time, higher quality).

Figure 2 shows the trade-off between quality versus total time. Tendem occupies a favorable region (higher quality at a similar or lower time than Upwork), showing best quality per hour spent. Despite lower quality, AI-only system lies on the sub-efficient frontier, while they can solve some tasks fast and cheap – many outputs are low quality and fail to meet requirements.

Takeaways.

- **Quality:** Tendem improves overall Good rate by +21.3 pp vs Upwork (74.5% vs 53.2%), with the largest gain in *Completeness* (+22.3 pp).
- **Speed:** Tendem cuts median total time by an impressive 53% (16.5 h vs 35.0 h), driven by faster both connection and execution times.

- **Cost:** Typical (median) price is 36% lower than Upwork (32.0 vs 50.0 USD), while average cost is higher due to a small number of escalated, high-touch tasks.
- **Positioning:** For ambiguous, multi-document, or spec-heavy tasks, hybrid step-gates reduce omissions/-fabrication, **closing the last-mile gap where AI-only systems remain brittle**.

3.2 External Benchmarks (Tendem's underlying AI agent)

Additionally, the underlying AI agent – operating autonomously without human involvement – was evaluated to assess the strength of the automation layer and its performance relative to other AI agents.

Benchmarks

We include benchmarks that (i) reflect *real-world tasks* over toy puzzles, (ii) are *hard*, not saturated at ≈100%, and (iii) match our target scope of browsing/tool use and factual reasoning (iv) are not *environment-constrained* (no predefined tool implementations).

- **BrowseComp** (Wei et al., 2025): deep, multi-hop web browsing for verifiable answers.
- **HLE (Humanity's Last Exam)** (Phan et al., 2025): hard multi-domain knowledge and reasoning.
- **GAIA** (Mialon et al., 2023): artificially created assistant tasks with tool-use/browsing.

Setup and policy

All Tendem runs in this section use the autonomous backbone AI agent from the Tendem system, evaluated independently.

Tendem's AI agent was run on all benchmark tasks using the specified configuration, with official benchmark evaluation scripts applied to the results.

Evaluation follows official metrics: exact match score for BrowseComp GAIA and HLE (Wei et al., 2025; Mialon et al., 2023; Phan et al., 2025). For contextual baselines, we use official system cards/blogposts values as-is; setups might differ across sources. We chose it to ensure a fair comparison with Tendem.

Results

Table 4 compares Tendem's AI Agent against our baselines.

Tendem's AI Agent is competitive on browsing/tool use and performs strongly relative to established baselines, yet leaves room for improvement.

System	BrowseComp	HLE	GAIA
Tendem's AI agent without human involvement	**71.0**	39.0	**78.2**
ChatGPT Agent	68.9	41.6	–
ChatGPT Deep Research	51.5	26.6	67.4
GPT-5 pro with tools	–	**42.0**	–
GPT-5 high with tools	54.9	35.2	–
o3 high with tools	49.7	24.3	–
Perplexity Deep Research	–	21.1	–
Manus	–	–	73.4
Flowith	–	–	78.1
HF Open Deep Research GPT-5 medium	–	–	62.8
HAL Generalist Agent Claude Sonnet 4.5	–	–	72.6

Table 4: Evaluation of Tendem and other AI systems on agentic benchmarks. All values are percentages. Non-Tendem numbers are vendor-reported; setups might differ across sources.

Tendem's AI Agent is **competitive on web browsing/tool-use** (state-of-the-art tier on BrowseComp; top on GAIA) and **close to leading models on hard knowledge** (HLE), yet still has a room for improvement. Results support the hybrid agent: high-quality agent to automate routine/tool-heavy steps and escalate judgment where models remain brittle.

3.3 Evaluation limitations

The in-house benchmark reflects the distribution of typical freelance tasks but does not fully represent the wider market nor include specialist work such as senior-level programming, medical, or legal-advice tasks.

Also, the style of the system outputs might affect evaluations; however we try to mitigate this bias by removing any reference to the evaluated system, using sub-scale rubrics to calibrate raters, and allowing a expert to report uncertainty where relevant.

Cross-benchmark comparability is imperfect: BrowseComp uses live web; HLE disallows browsing by design yet many models report score with web tools; GAIA mixes tool use and retrieval; HAL reports are cost-aware but configuration-sensitive (Wei et al., 2025; Mialon et al., 2023; Phan et al., 2025).

4 Safety, Intended Use, and Limitations

Out-of-scope. Regulated or expert-level advice (yet we are expanding domains coverage); high-stakes decisions.

Primary risks and mitigations.

- **Hallucinations / Fabrication.** Step gates and source-required checks; offline QA blocks handoff when confidence is low.
- **Specification drift.** Acceptance criteria captured during clarification; enforced as tests in online QA.
- **Privacy.** All experts working with Tendem operate in a secure environment (VM) and follow strong privacy protocols. Additionally, we perform regular production-use inspections, enforce access controls, audit experts.

5 Release and Reproducibility

We release our in-house collected benchmark with economically valuable tasks. We publish full input text, output and produced files for each of the systems – on GitHub `https://github.com/toloka/tendem-evaluation`.

6 Conclusion

Tendem is a hybrid agent that combines automated planning and tool use with human step gates and layered quality assurance. On a 94-task in-house benchmark that mirrors economically valuable freelance tasks, Tendem delivers higher quality and faster turnaround than both human-only and AI-only approaches, achieving 74.5% Good outcomes and a 53% reduction in median total time compared to freelance workers.

The observed quality gains arise from the full hybrid stack rather than any single component. High-quality backbone AI agent executes routine and tool-heavy steps and Human Experts and QA specialists provide targeted review and updates.

Tendem's AI agent evaluation shows that the underlying agent operating autonomously is competitive on browsing and real-world assistant tasks, and remains close to leading models on hard knowledge, supporting the design choice to automate routine, tool-heavy steps while escalating judgment-heavy decisions to Human Experts.

We will continue improving both Agentic and Hybrid components by expanding tool integrations, strengthening online/offline QA signals, refining the routing and escalation pipeline, and broadening the expert pool from general to more specialized domains.

References

E. M. Bender and A. Koller. Climbing towards NLU: On meaning, form, and understanding in the age of data. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online, July 2020.

Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.463. URL https://aclanthology.org/2020.acl-main.463/.

J. Gu, X. Jiang, Z. Shi, H. Tan, X. Zhai, C. Xu, W. Li, Y. Shen, S. Ma, H. Liu, S. Wang, K. Zhang, Y. Wang, W. Gao, L. Ni, and J. Guo. A survey on llm-as-a-judge, 2025. URL https://arxiv.org/abs/2411.15594.

L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, Jan. 2025. ISSN 1558-2868. doi: 10.1145/3703155. URL http://dx.doi.org/10.1145/3703155.

Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, Mar. 2023. ISSN 1557-7341. doi: 10.1145/3571730. URL http://dx.doi.org/10.1145/3571730.

P. Manakul, A. Liusie, and M. J. F. Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models, 2023. URL https://arxiv.org/abs/2303.08896.

G. Mialon, C. Fourrier, C. Swift, T. Wolf, Y. LeCun, and T. Scialom. Gaia: a benchmark for general ai assistants, 2023. URL https://arxiv.org/abs/2311.12983.

L. Phan, A. Gatti, Z. Han, N. Li, J. Hu, H. Zhang, C. B. C. Zhang, M. Shaaban, J. Ling, S. Shi, et al. Humanity's last exam, 2025. URL https://arxiv.org/abs/2501.14249.

J. Wei, Z. Sun, S. Papay, S. McKinney, J. Han, I. Fulford, H. W. Chung, A. T. Passos, W. Fedus, and A. Glaese. Browsecomp: A simple yet challenging benchmark for browsing agents, 2025. URL https://arxiv.org/abs/2504.12516.

Y. Weng, M. Zhu, F. Xia, B. Li, S. He, S. Liu, B. Sun, K. Liu, and J. Zhao. Large language models are better reasoners with self-verification, 2023. URL https://arxiv.org/abs/2212.09561.

S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao. React: Synergizing reasoning and acting in language models, 2023. URL https://arxiv.org/abs/2210.03629.