arXiv:2602.01119v1 [cs.CL] 1 Feb 2026

Tendem：混合人工智能+人工平台

Toloka团队

2025年11月

摘要

Tendem 是一个混合系统，由人工智能处理结构化、可重复的工作，并在人类专家遇到模型失效或需要验证结果时介入。每项结果在交付给客户之前，都会经过全面的质量审查。

为了评估 Tendem 的性能，我们对 94 项现实任务进行了系列内部评估，并将其与纯人工智能代理以及由 Upwork自由职业者执行的纯人工工作流进行了比较。

结果表明，Tendem 始终能提供质量更高的输出，且周转时间更短。同时，其运营成本仍与纯人工执行相当。

在第三方智能体基准测试中，Tendem 的 AI 智能体（自主运行，无需人工参与）在网页浏览和工具使用任务方面的表现接近最先进水平，在前沿领域知识和推理方面也展现出强劲表现。

1 引言

我们推出 Tendem[1] ——一个将人工智能执行与人类专业知识相结合的混合执行平台，可简化专业和个人工作流程。任务由人工智能与人类专家协作完成：人工智能负责处理常规且节奏快的工作，人类专家则提供监督、确保上下文准确性并完善输出。如果人工智能达到能力极限，专家会接手工作，确保任务完成。该平台的架构包含一个用于任务编排、路由和质量保证的多层流水线，将自动推理与人工验证相结合。

与完全自动化的智能体工作流不同——例如嵌入在 ChatGPT[2] 和 Claude[3] 等广泛使用的助手中的智能体——Tendem 有意在不确定性高或影响重大的环节保留人在回路。纯自主代理往往会优先优化速度和成本，但在处理欠规范、多文档综合或非局部质量标准时可能会遇到困难。

相比之下，纯人工工作流（如自由职业者平台）或许能提供可追溯的工作履历评分与问责机制，但往往更慢，且成本波动更大。Tendem 瞄准高效节点：在模糊或对合规敏感的任务上，其质量优于纯 AI 系统，同时相对于纯人工执行，改善出结果时间与协调成本。

在本研究中，我们呈现一项对比研究的结果，该研究涵盖全自动流水线、纯人工工作流以及Tendem混合模型。我们观察到结果质量与一致性显著提升：与纯 AI 或纯人工工作流相比，Tendem 将更大比例的任务转化为可交付客户的成果。此外，相对于纯人工执行，Tendem 降低了运营成本，因此更高的质量并非以价格为代价。

该评估包含两个部分：一套由 94 项现实任务组成的内部任务集，以及一组公开智能体基准。我们衡量结果质量（人工评估）、时间和价格。

Tendem 在各项任务中实现了 74.5% 的高质量结果，相较纯人工基线，中位交付时间缩短了 53%。我们还在外部基准上评估了 Tendem 底层的 AI 智能体（无人工参与）。该智能体在网页浏览和现实世界助手任务方面接近最先进水平，在高难度知识方面也仍接近领先模型。这为我们的 Tendem 智能体提供了坚实的基础。

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

2 系统概述

Tendem 的架构围绕人工智能与人类专业知识之间的协作构建。每项任务都会经过一套结构化流程，明确人工智能何时独立执行、何时需要人工判断，从而确保工作流程高效，并产出高质量的交付成果。每个阶段都内置了检查和验证机制，避免小错误不断累积，并将人工注意力引导至最能发挥作用的环节。

2.1 角色

在设计系统时，我们重点关注以下三种角色：

客户／最终用户

客户首先概述问题并描述期望的结果，同时分享有效开展工作所需的背景材料或数据。客户期望在有限的时间内，以稳定的成本获得符合其标准的结果。但同样重要的是，他们希望清楚了解自己的数据在每个阶段如何被处理和保护。

自动化组件：Tendem 的 AI 智能体和工具

Tendem 的 AI 智能体 通过带有明确验证步骤门控的计划—行动—观察—验证循环，执行大部分常规及工具密集型工作（Yao 等，2023；Weng 等，2023；Manakul 等，2023）。工具包括网页浏览、适用于常见办公格式的文件输入／输出 、用于分析和图表制作的安全隔离*Python* 运行时，以及隔离的安全运行环境*bash*／命令行界面。所有操作均以最小权限运行并被记录，使每一步都可追溯且可重复。

质量保证工具持续运行：规范符合性、分项／总计核对、引文匹配以及轻量级自洽性检查。当系统遇到不确定性时——无论源于相互冲突的来源、未通过的检查，还是风险过高的步骤——都会升级交由人类专家处理。

人类专家

一位**人类专家** 是能够补充模型尚无法提供的推理与情境感知能力的专业人士（Huang 等，2025；Ji 等，2023；Bender 和 Koller，2020）。他们引导系统完成不确定的步骤，并验证、完善最终结果以供实际使用。专家须通过一系列测试与考试方可准入，以核验其专业能力，并确保其接受培训和持续的质量保证。

在生产工作流程中，专家在步骤门控（计划审计、计划步骤检查、草稿完善等）处介入，并作为分层质量保证的一部分执行最终的离线质量保证审核。绩效通过质量保证和返工率进行跟踪。

2.2 端到端工作流程

Tendem 工作流程由以下相互关联的步骤组成：

- **客户请求**：客户用通俗语言概述他们希望实现的目标，并添加系统理解该任务所需的任何文件或参考资料。
- **澄清并正式化**：AI 智能体检查文件和需求，并提出有针对性的问题。
- **带门控步骤的计划**：AI 智能体将任务分解为多个步骤，并在人类专家监督下对关键和高风险步骤实施门控。
- **路由与匹配**：系统根据所需的技能组合识别合适的专家，并估算所需时间。
- **混合执行**：AI 智能体在人类监督下开展工作；当输出需要调整或进一步完善时，人类专家会介入。
- **在线质量保证**：系统会对 AI 智能体和人工评估者的编辑内容进行轻量级自动检查，并在需要时更新计划、继续执行后续步骤。

• **离线质量保证**：任务执行后，系统会根据客户的要求和所附材料进行自动多步验证，并将不确定的情况升级交由人工质量保证专家处理。
• **最终定稿**：系统会将经过验证的结果交付给客户；如果仍存在重大问题，则退回返工。

2.3 目标优势

采用多组件工作流程有助于我们实现以下几项关键优势：

(1) 智能混合执行： 通过任务分解和动态路由（利用人工智能提升速度，由人类专家进行判断，并在两者之间建立验证闭环），减少协调开销，并确保每项交付成果都兼具机器效率与专家的人类判断。**(2) 多层质量保证：** 每项交付成果都会先由自动化系统进行检测，标记潜在故障或幻觉迹象。随后，人工审核员会追溯输出的生成过程，检查其逻辑以及与源材料和要求的一致性，直至达到交付标准。**(3) 专家生产力提升**。 自动化处理常规子任务，让人类专家能够专注于高影响力决策和润色，从而缩短总周期时间，并更加注重细节。**(4) 提升客户体验**。 该系统采用清晰、结构化的流程，提供有针对性的澄清和可视化进度，让用户觉得使用起来直观而高效。

3 评估

为量化价格与质量权衡，我们在内部收集的、具有经济价值且能反映真实自由职业平台情况的任务上评估 Tendem。我们将 Tendem 与人工智能代理及自由职业市场工作者基线进行比较（第 3.1 节）。这衡量了客户在具有价值的现实任务中感知到的结果。

为确保比较透明，我们还在公开基准上评估 Tendem 底层的 AI 智能体（不进行人工干预），并将其与现有模型和纯人工智能代理进行比较（第 3.2 节）。我们将其作为评估骨干人工智能代理能力的依据，并考察其对混合系统总体效率和性能的影响。

3.1 基于现实任务的内部基准

公开基准能够检验核心技能，但对依赖客户个人偏好的现实情境代表不足（例如模糊任务说明、复杂的多阶段任务，以及作为交付成果的输入内容和文件）。

为更好地评估系统在具有经济价值的现实远程工作项目 中的性能，并反映专业任务执行的复杂性和实际限制，我们推出了一项新的内部基准。该基准近似模拟自由职业平台的任务分布，以反映专业工作的实际情况。

我们假设，客户会从三个相互独立的维度评估表现和任务完成情况：**结果质量**、**执行时间**和**价格**。

指标

为了将 Tendem 的端到端表现与其他人工智能代理及自由职业平台基准进行比较，我们会针对每项任务测量以下指标：

1. **结果质量**：按照三点评分量表（差、中等、好）评估每项输出，并增加“拒绝”标签。我们依据三项细粒度质量标准和总体结果质量来评估质量。

- **准确性** –事实和数值是否正确，是否存在捏造信息、来源缺失或信息过时；仅仅提供参考资料并不能弥补不准确之处，参考资料本身也应与内容相关。
- **完整性**– 涵盖客户任务中的所有必需事项、约束和交付成果，确保结果完整且可直接使用，无需后续跟进；不遗漏任何事项，也不省略任何处理环节。

• **风格与格式–** 清晰度、结构和呈现质量，包括易读的组织方式（标题、列表、表格）、单位和符号表示的一致性、正确的语法和拼写、恰当的语气，以及正确的输出格式。

• **总体质量–** 端到端的用途适配度：答案符合用户意图，并且根据上述三个评估标准（分别独立评定，以权衡上下文因素和实际实用性）已可交付。

每项评估标准均按以下等级评分：

• **好** — 完全满足评估标准，可直接交付客户；最多只有不影响修改需求的小瑕疵。
• **中等** — 需要进行一些编辑才能满足评估标准。
• **差** — 存在重大缺漏；未经大量返工，不适合使用。
• **拒绝**— 由于安全过滤器或领域覆盖限制，任务遭到拒绝或未予尝试。

2. **时间**（延迟）：按从任务发布到开始实际执行的连接时间（如有）、执行时间直至最终结果的总时间计算，为两者之和。
3. **价格**：每项任务的实际成本，以人工劳动和基础人工智能用量之和计算，并以美元记录。

受评系统

在评估中，我们选择了常用的纯 AI 系统和纯人工系统。评估对象如下：

1. **ChatGPT Agent (OpenAI)**[4]。 集成于 ChatGPT 中的多步骤智能体助手，支持操作、工具使用和浏览。通过 ChatGPT 界面并使用默认设置进行评估。2. **Upwork**[5]。 通过 Upwork 平台雇用的自由职业者，他们遵循下述原则（最佳匹配、价格、工作成功率筛选）。每项任务均单独雇用自由职业者。我们没有指示他们使用或不使用人工智能。3. **Tendem**[6]。 Tendem 的生产版本。详细的系统概述见第 2 节。

数据集设计与收集

B构建一个具有代表性的真实世界客户任务基准测试涉及以下结构化流程 s:

• **用例抽样**。 从 Upwork[5]、Fiverr[7] 和 Freelancer[8] 中选取常见任务类别，以近似付费知识型工作的分布。
• **任务编写**。 根据抽样得到的类别，我们委托从业者创建新任务，并提供明确的验收标准及相关素材（例如电子表格、PDF 文件和链接）。
• **个人身份信息与合规性**。 任务设计为**不包含个人身份信息**，并避开受监管领域（例如医疗/法律建议）。

最终基准测试的领域/类别分布见表 1。

任务提交流程

所有第三方系统均通过其公开用户界面使用——手动粘贴任务文本——以反映终端用户体验并记录真实的连接延迟。在可用的情况下，使用最高订阅等级，以确保获得最高速率限制和最佳底层模型。我们记录了每项任务的创建时间、人工评估者开始工作的时间，以及交付最终结果的时间。

[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	类别	数量
销售	收集企业联系数据	14
	补全缺失字段（数据丰富）	6
	销售总额	**20**
运营	构建多步骤自动化工作流	1
	收集数据	10
	转换格式	2
	提取 PDF／文档／报告内容	3
	安排和管理预约及通话	3
	将原始数据整理为架构	6
	验证联系信息	3
	运营总计	**28**
营销	收集企业联系数据	4
	创建内容	7
	市场与竞争研究报告	10
	校对、分析并修正内容	3
	营销总计	**24**
分析	客户／用户访谈或反馈收集	1
	生成绩效仪表板与摘要	8
	市场与竞争研究报告	8
	运行探索性数据分析	5
	分析师总计	**22**
总计		**94**

表 1：内部基准领域与类别分布。

Upwork 上的 Freelancer 筛选

由于自由职业平台上的专家收费和技能水平各不相同，我们采用以下原则来指导选择，确保比较公平。

任务发布后收集提案；筛选时优先考虑候选人以往的人工评估经验、价格和工作成功率，并参考申请人的 Upwork 个人资料信息。我们倾向于选择工作成功率超过 80% 的最低价候选人，或前景良好的新手。最终选定执行者时不进行价格谈判——我们对报价进行评估，并原样接受最低的可接受报价。

标注方案

我们从专业网络中招募了独立的人工质量保证专家，并对其进行入职培训，要求他们通过入门考试。评分者会收到关于如何应用评估标准（准确性、完整性、风格与格式、总体）的简明说明，以及基于示例的指导。为减少偏差，系统均经过匿名化处理，并在同一界面中接受评估。评估人员只能看到输入任务、所附素材和生成的输出。所有专家均获得具有竞争力的时薪。

对于每项任务，一名专家会针对全部 4 个质量等级指定标签（好、中等、差），并提供简短备注说明其决定的理由。如果系统拒绝完成任务，我们会记录为 拒绝 。在本实验中，我们对每项任务采用单人评分；但相当大比例的项目接受了抽查，以确保标注质量高。

我们未采用或验证作为评判者的大语言模型（Gu 等人，2025年）来评分，因此所有结果仅依赖人工评估（不过，这可以作为一项很好的后续实验）。

结果

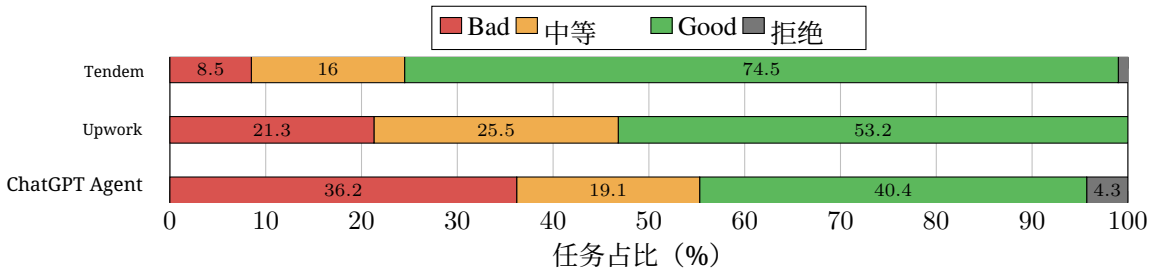

图 1：在内部基准测试中，各系统获得差/中等/好的任务占比。由人工评估者在不知晓系统名称的情况下评估每个系统。拒绝标签表示该系统的安全过滤器拒绝了该任务。

系统评估。 图 1 总结了各个受评估系统的差、 中等和好结果的分布。Tendem 的良好结果占比达到 74.5%，比 Upwork 高出 +21.3 个百分点（53.2%），比 ChatGPT Agent 高出 +34.1 个百分点（40.4%）。

Tendem 的较差结果也更少：差评占 8.5%，而 Upwork 为 21.3%，ChatGPT Agent 为 36.2%；这表明 Tendem 的表现更稳定、更可靠。

总体而言，在我们的内部评估中，**Tendem 的良好率显著高于所有基准系统**（Upwork 和 ChatGPT Agent）。对 Tendem 与最接近的基准系统 Upwork 的“好”结果占比进行单侧 z-检验，结果证实这项提升具有统计显著性（p 值为 $= 0.0012$）。

系统	总体				准确性	完整性	风格
	(% 好)	(% 中等)	(% 差)	(% 拒绝)	(% 好)	(% 好)	(% 好)
Tendem	**74.5** ± 4.5	16.0 ± 3.8	8.5 ± 2.9	1.1 ± 1.0	**74.5** ± 4.5	**81.9** ± 4.0	**70.2** ± 4.7
Upwork	53.2 ± 5.1	25.5 ± 4.5	21.3 ± 4.2	0.0 ± 0.0	63.8 ± 5.0	59。6 ± 5.1	59.6 ± 5.1
ChatGPT Agent	40.4 ± 5.1	19.1 ± 4.1	36.2 ± 5.0	4.3 ± 2.1	48.9 ± 5.2	48.9 ± 5.1	51.1 ± 5.2

表 2：在内部基准测试中，各系统在总体及三项质量标准上的差、 中等和好评任务占比。拒绝标签表示系统因安全过滤器或缺乏完成该任务的专业知识而拒绝了该任务。每个系统均由一名人工评估者进行评估，且未向其透露系统名称。列中最佳结果以粗体显示。

质量标准分析。 表 2 不仅展示了总体任务评估，还分析了各系统在三项质量标准上的表现：

- 准确性。 Tendem 的准确性比 Upwork 高出 +10.7 个百分点（74.5% 对比 63.8%），这表明精心设计的混合工作流和分层质量保证能够增强事实依据和数值正确性。
- 完整性。 差距最大：比 Upwork 高出 +22.3 个百分点（81.9% 对比 59.6%）。原因之一可能是采用了带有分步门控的工作流程，可减少遗漏并确保符合验收标准。
- 风格与格式。 Tendem 比 Upwork 高出 +10.6 个百分点（70.2% 对比 59.6%），比 ChatGPT Agent 高出 +19.1 个百分点（70.2% 对比 51.1%），呈现效果更加整洁，可直接交付客户。

错误模式。 纯 AI 系统会出现：(i) 遗漏长规格说明中的内容；(ii) 在超出模型上下文限制时，综合分析流于表面；(iii) 伪造参考文献（Ji 等，2023；Huang 等，2025）。纯人工输出在各项评估标准上的质量参差不齐。Tendem 在步骤门控的不确定性较高时升级处理，从而大幅减少遗漏和伪造错误。

系统	价格 (美元)		时间 (小时)					
			连接		执行		总计	
	平均值	中位数	平均值	中位数	平均值	中位数	平均值	中位数
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

表 3：价格和时间指标，标准差显示为 ± 标准差。价格以美元计，时间以小时计。对于中位数，标准差通过自助法计算。ChatGPT Agent 是基于订阅的服务，因此没有直接价格。

价格和速度。 表 3 列出了每项任务的执行时间（分为连接时间和有效执行 时间）以及价格分析。

Tendem 的中位数价格为 32.0 美元，低于 Upwork 的 50.0 美元（-36%）；其平均价格则高于 Upwork，为 69.2 美元，而 Upwork 为 48.0 美元，这反映出右偏成本分布，其中少数案例需要大量人工介入。

在时间方面，Tendem 将总时间中位数从 35.0 小时（Upwork）缩短至 16.5 小时（-53%），连接速度更快（2.6 小时对比 4.6 小时，-42%），执行时间也更短（10.5 小时对比 26.8 小时，-61%）。正如预期，ChatGPT Agent 的延迟和价格都很低，但——如表 2 所示——在该数据集上的质量较低。

总体而言，**Tendem 的交付速度显著快于纯人工基线，中位数价格更低，质量也更高**。平均价格较高，是因为少数复杂任务升级处理，需要更多专家参与。

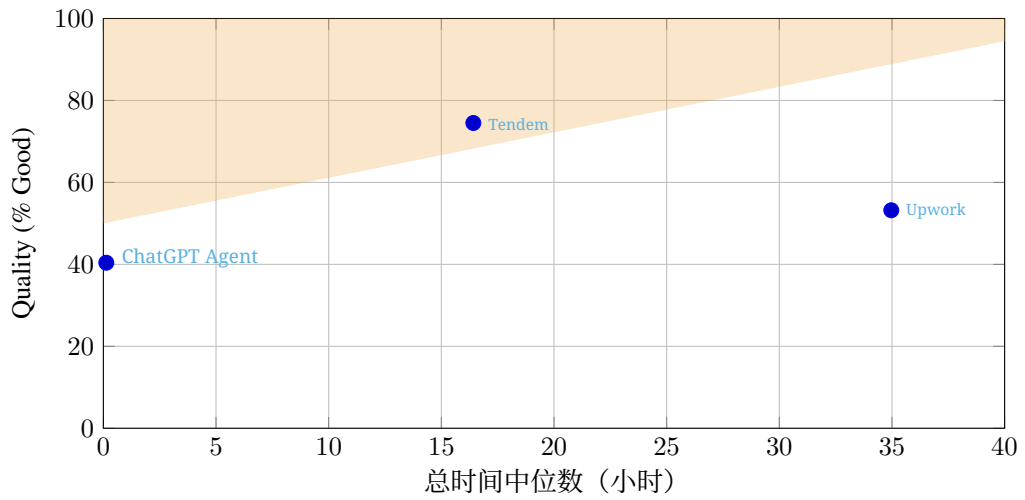

图 2：评估的系统质量与总时间中位数。浅橙色角落区域突出显示了理想范围（时间更短、质量更高）。

图 2 展示了质量与总时间之间的权衡。Tendem 处于有利区间（用时与 Upwork 相近或更短，质量却更高），单位投入时间所获得的质量最佳。虽然质量较低，纯 AI 系统仍处于次优效率前沿，因为它们能快速、低成本地完成部分任务，但许多输出质量不高，无法满足要求。

要点。
- **质量：** Tendem 的总体良好率比 Upwork 高 +21.3 个百分点（74.5% 对 53.2%），其中完整性的提升幅度最大（+22.3 个百分点）。
- **速度：** Tendem 将总时间中位数大幅缩短了 53%（16.5 小时对比 35.0 小时），这得益于更快的连接和执行时间。

• **价格：** 典型（中位数）价格比 Upwork 低 36%（32.0 美元对比 50.0 美元），但平均成本更高，这是由于少量升级处理、需要高度介入的任务所致。
• **定位：**对于模糊、多文档或规格密集的任务，混合分步门控可减少遗漏/捏造，**弥合纯 AI 系统仍然脆弱的最后一公里差距**。

3.2 外部基准（Tendem 的底层 AI 智能体）

Additionally，底层 AI 智能体——在无人工参与的情况下自主运行——被评估 ed 作为评估自动化层的实力及其相对于其他人工智能代理的表现。

基准测试

我们纳入的基准测试需满足以下条件：(i) 侧重现实任务而非玩具谜题；(ii) 具有较高难度，得分未达到≈100%的饱和水平；(iii) 符合我们的目标范围，即浏览/工具使用和事实推理；(iv) 不受环境限制（没有预先定义的工具实现）。

• **BrowseComp** （Wei 等，2025）：通过深入、多跳网页浏览获取可验证答案。
• **HLE（人类最后考试）**（Phan 等，2025）：高难度的多领域知识与推理 ng.
• **GAIA** （Mialon 等，2023）：人工创建的、涉及工具使用/浏览的助手任务。

设置与策略

本节中 Tendem 的所有运行均使用 Tendem 系统中的自主骨干人工智能代理，并进行独立评估。

Tendem 的人工智能智能体按照指定配置运行了所有基准任务，并使用官方基准评估脚本对结果进行评估。

评估遵循官方指标：BrowseComp、GAIA 和 HLE 使用精确匹配分数（Wei 等，2025年；Mialon 等，2023；Phan 等，2025年）。对于上下文基线，我们直接采用官方系统卡片和博客文章中的数值；不同来源的设置可能有所差异。选择这些数值是为了确保与 Tendem 进行公平比较。

结果

表 4 比较了 Tendem 的 AI 智能体与我们的基线。

Tendem 的 AI 智能体在浏览和工具使用方面表现出竞争力，相较于既有基线也表现强劲，但仍有提升空间。

系统	BrowseComp	HLE	GAIA
Tendem 的 AI 智能体 无需人工评估者参与	**71.0**	39.0	**78.2**
ChatGPT Agent	68.9	41.6	–
ChatGPT 深度研究	51.5	26.6	67.4
GPT-5 pro 配备工具	–	**42.0**	–
GPT-5 high 配备工具	54.9	35.2	–
o3 high 使用工具	49.7	24.3	–
Perplexity 深度研究	–	21.1	–
Manus	–	–	73.4
Flowith	–	–	78.1
HF Open Deep Research GPT-5 中等	–	–	62.8
HAL Generalist Agent Claude Sonnet 4.5	–	–	72.6

表 4：Tendem 与其他人工智能系统在智能体基准测试上的评估结果。所有数值均为百分比。Tendem 以外的数据由供应商报告；不同来源的测试设置可能有所不同。

Tendem 的 AI 智能体在**网页浏览/工具使用方面具有竞争力**（在 BrowseComp 上达到最先进水平，在 GAIA 上名列前茅），在**高难度知识方面接近领先模型**（HLE），但仍有提升空间。结果支持混合智能体模式：使用高质量智能体自动处理常规及工具密集型步骤，并在模型仍不够可靠之处升级交由人工判断。

3.3 评估局限性

内部基准反映了典型自由职业任务的分布，但未能充分代表更广泛的市场，也未涵盖高级编程、医学或法律咨询等专业工作。

此外，系统输出的风格可能会影响评估；不过，我们试图通过移除对被评估系统的任何提及、使用子量表评分标准来校准评分者，以及允许专家在相关情况下报告不确定性，来减轻这种偏差。

跨基准可比性并不完美：BrowseComp 使用实时网络；HLE 在设计上不允许浏览，但许多模型报告了使用网络工具的得分；GAIA 混合了工具使用与检索；HAL 报告具有成本意识，但对配置敏感（Wei 等，2025；Mialon 等，2023；Phan 等，2025）。

4 安全、预期用途与局限性

范围外。 受监管或专家级建议（不过我们正在扩展领域覆盖）；高风险决策。

主要风险与缓解措施。

- **幻觉 / 捏造**。 步骤门控和必须提供来源的检查；当置信度较低时，离线质量保证会阻止交接。
- **规格漂移**。 在澄清过程中记录验收标准；并在在线质量保证中将其作为测试执行。
- **隐私**。 所有与 Tendem 合作的专家均在安全环境（虚拟机）中工作，并遵循严格的隐私协议。此外，我们会定期检查生产环境中的使用情况，实施访问控制，并审计专家。

5 发布与可复现性

我们发布内部收集的、包含具有经济价值的任务的基准测试。我们会在 GitHub https://github.com/toloka/tend em-evaluation 上公布每个系统的完整输入文本、输出和生成的文件。

6 结论

Tendem 是一种混合智能体，将自动化规划和工具使用与人工步骤门控及分层质量保障相结合。在一项包含 94 项任务的内部基准测试中，该测试模拟了具有经济价值的自由职业任务，Tendem 的质量和交付速度均优于纯人工和纯人工智能方法，良好结果占比达到 74.5%，与自由职业者相比，总时间中位数缩短了 53%。

观察到的质量提升源于完整的混合技术栈，而非任何单一组件。高质量的骨干人工智能代理负责执行常规步骤和工具密集型步骤，人类专家和质量保证专家则提供有针对性的审核与更新。

Tendem 的人工智能智能体评估表明，底层智能体在自主运行时，在浏览和现实世界助手任务方面具有竞争力，在高难度知识任务方面也与领先模型表现接近。这支持了这样的设计选择：将常规且工具密集型的步骤自动化，同时将需要大量判断的决策升级交由人类专家处理。

我们将继续改进智能体和混合组件，包括扩展工具集成、强化在线/离线质量保证信号、完善路由与升级流程，并将专家池从通才拓展至更多专业领域。

核心贡献者

Konstantin Chernyshev、Ekaterina Artemova、Viacheslav Zhukov、Maksim Nerush、Mariia Fedorova、Iryna Repik、Olga Shapovalova、Aleksey Sukhorosov、VladimirDobrovolskii、Natalia Mikhailova、Sergei Tilga

参考文献

E. M. Bender 和 A. Koller。《迈向自然语言理解：数据时代的意义、形式与理解》。载于 D. Jurafsky、J. Chai、N. Schluter 和 J. Tetreault 编，计算语言学协会第*58*届年会论文集，第5185–5198页，线上，2020年7月。计算语言学协会。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 和 J. Guo。《关于作为评判者的大语言模型的调查》，2025年。URL：https://arxiv.org/abs/24 11.15594。L. Huang、W. Yu、W. Ma、W. Zhong、Z. Feng、H. Wang、Q. Chen、W. Peng、X. Feng、B. Qin 和 T. Liu。《大语言模型幻觉综述：原理、分类体系、挑战与开放问题》。*ACM*信息系统汇刊，43(2)：1–55，2025年1月。ISSN 1558-2868。doi：10.1145/3703 155。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 和 P. Fung。《自然语言生成中的幻觉综述》。*ACM*计算调查，55(12)：1–38，2023年3月。ISSN 1557-7341。doi：10.1145/3571730。URL：http://dx.doi.org/10.1145/3571730。P. Manakul、A. Liusie 和 M. J. F. Gales。《SelfCheckGPT：用于生成式大语言模型的零资源黑盒幻觉检测》，2023年。URL：https://arxiv.org/abs/2303.08896。G. Mialon、C. Fourrier、C. Swift、T. Wolf、Y. LeCun 和 T. Scialom。《GAIA：通用人工智能助手基准测试》，2023年。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 等。《人类最后考试》，2025年。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 和 A. Glaese。《BrowseComp：面向浏览智能体的简单而富有挑战性的基准测试》，2025年。URL：https: //arxiv.org/abs/2504.12516。Y. Weng、M. Zhu、F. Xia、B. Li、S. He、S. Liu、B. Sun、K. Liu 和 J. Zhao。《通过自我验证提升大语言模型的推理能力》，2023年。URL：https://arxiv.org/abs/2212.09561。S. Yao、J. Zhao、D. Yu、N. Du、I. Shafran、K. Narasimhan 和 Y. Cao。《ReAct：在语言模型中协同推理与行动》，2023年。URL：https://arxiv.org/abs/2210.03629。