

主动个人 Agent 的机会与边界

研究截至 2026 年 10 月 14 日

阅读目录

核心结论

- 一 真正变化的是用户把哪些事情交给助手
- 二 为什么这轮变化值得认真看待
- 三 Muse 与 dot 的案例揭示了什么
- 四 长期记忆、权限和停止机制决定用户能否放心委托
- 五 商业可行性取决于省下的时间是否超过新增成本
- 六 谁更可能保留长期价值
- 七 哪些反例会改变判断
- 八 下一步应该看什么数据

结语

附录 市场格局与四条产品路线

- 一、平台个人入口：把 Agent 收进原有应用
 - 二、工作流产品：先接管团队的待办，再扩大责任范围
 - 三、自托管：把“自己的助手”做成可掌握的基础设施
 - 四、轻量消息入口：把安装软件变成添加一个联系人
- 看格局时，应当分开问的问题

核心结论

Muse、dot 所代表的变化，是 AI 助手开始尝试长期照看用户交给它的事情：发现遗漏、判断何时跟进、在授权范围内推进，并确认最后有没有办成。这比一次回答或一次工具调用更接近人们对“助手”的期待，也提出了更难兑现的产品承诺。

这条方向值得重视，但现有证据尚不足以证明它已经成为一个可独立收费、长期留存、利润稳定的消费市场。公开材料中有有用的任务成果，也有用户不断催动流程、访问权限卡住、监控悄悄停下来的案例。两类证据合起来，才是这个行业的真实起点。

本报告的判断是：主动个人 Agent 很可能成为重要的软件使用方式；近期更可能出现的是它与 Claude Code、Codex 等执行工具的融合，而非一个全新品类迅速替代另一个旧品类。能否形成独立生意，取决于它是否减少了用户发现任务、跟进进度和处理意外所花的总时间，而不只是增加建议、消息和后台运行时长。

本文中的 Agent，指能调用工具、推进任务的 AI 助手。harness 则指把模型、工具、任务状态、执行环境和权限规则连接起来的运行系统。主动性回答“什么时候开始、什么事情值得做”；harness 回答“怎样让这些工作持续而受控地运行”。二者可以同时存在，不能据此划分前后两代产品。

一 真正变化的是用户把哪些事情交给助手

编程 Agent 已经走出了一问一答

如果把 Claude Code、Codex 描述为只会写代码、没有记忆、每次都要人启动的终端工具，就会夸大个人 Agent 的新意。

Codex 在 2026 年 2 月推出桌面应用时，已经支持定时后台任务和结果审核队列；4 月的更新进一步加入线程自动化、后续唤醒和主动工作建议。Claude Code 的 routines 同样支持日程、API 和 GitHub 事件触发，关联 PR 的会话可以继续接收评论与 CI 更新。这些产品已经能围绕一个交代过的目标持续工作。

[Codex 桌面应用](#)；[Codex 4 月更新](#)；[Claude Code routines](#)

记忆也并非个人助手独有。Claude Code 的文档区分人工指令和自动记忆，并提供用户、项目等不同范围。它还明确提醒，记忆是提供给模型的上下文，不是保证执行的配置。[Claude Code 记忆文档](#)

因此，值得比较的是默认体验：用户需要自己准备多少背景、建立多少任务、配置多少触发条件，又需要多久回来检查一次。

个人助手把任务发现也纳入了承诺

软件项目通常已有 issue、PR、测试失败或告警，告诉 Agent 接下来该处理什么。个人生活和工作中，不少重要事项还没有被写成待办。例如，邮件里的安排变更可能意味着材料需要重做；一次已经结束的合作可能还有发票没有开。这里的困难是从零散信息中判断：这件事是否真的需要处理、用户是否已经处理过、现在是否值得打扰。

Muse 和 dot 希望将这种判断放在一个持续的个人关系里。用户交代目标后，助手应当记住有关背景，并在必要时回来推进。OpenAI 自己也在 dot 的系统说明中承认，持续运行和多 Agent 本身并非新技术；新产品需要应对的是运行过程中背景和权限不断变化的情况。[dot 系统卡补充](#)

这里至少包含四种不同能力：

- 1. 用户提出任务后，助手能调用工具完成。
- 2. 没有新指令时，助手能发现值得提醒或处理的事项。
- 3. 事情跨越数小时或数天，助手仍能记住进展，并在故障后继续。
- 4. 助手知道哪些动作要征求同意，能说明结果，也能接受纠正和停止。

一个产品可以擅长其中几项，却在另一项上失败。漂亮的购物目录不能证明它会可靠地监控折扣；成功发送一封邮件，也不能证明它理解了用户期待的审核方式。产品评价应分别检查这四件事。

同一件事可以有三种启动方式

同一件事，三种启动方式

假设退款场景。三种机制可以共存，不是产品等级排序。

显式委托	事件触发	主动发现
用户交代目标 “退货后帮我跟进退款”	用户给出条件 “超过七天未到账就提醒”	助手提出候选事项 发现退回商品仍未确认退款

下面用退款作一个假设例子，只用于说明机制，不是任何产品的实测记录。

用户说“这件商品退货后，请帮我跟进退款”，是显式委托。助手的目标和范围来自用户，后续即使跨天运行，也不需要重新猜测需求。用户说“退款超过七天还没到账就提醒我”，是在明确规则下进行事件自动化。助手需要可靠地检查状态、判断条件和恢复任务，但未必需要复杂的自主判断。

更强的主动发现则是：用户没有建立退款任务，助手从获准读取的邮件中看出商品已经退回，又发现退款迟迟没有确认，于是提出跟进建议。它必须判断这是否真是遗漏，是否已通过其他渠道解决，以及能否取得足够证据。主动发现本身并不授予它联系商家、提交投诉或改变交易的权限。

同样，项目改期后自动更新已明确关联的文档，可以是规则或既有任务的一部分；从团队讨论中推断某份尚未被提及的客户材料也需要修改，才涉及发现新需求。等待回复也是如此：按用户要求等到回复再继续，已经有很大价值，没必要为了强调“主动”而把它说成系统凭空发现了目标。

这三种方式可能出现在同一个产品中。前两种并不低级，第三种也不总是更好。若规则已经明确，可靠的自动化往往比不断猜测更合适。

判断一：新价值主要在于减少用户发现和跟进工作的负担。如果在相同模型与权限下，普通任务工具加上少量自动化配置，就能以同样少的用户投入达到相同结果，那么个人助手的独特优势便会缩小；它仍可能是更易用的包装，但不能据此声称产生了新的技术壁垒。

二 为什么这轮变化值得认真看待

模型能处理更长的任务，但可靠托管仍有距离

METR 在 2026 年 1 月的研究更新中，对模型达到 50% 成功率的任务难度跨度给出了持续增长的结果。这里的“时间跨度”按人类专家完成任务所需时间衡量，并不表示 AI 能无人看守地连续工作同样长的时间；50% 成功率也显然不足以支持重要个人事务的可靠托管。[METR 研究更新](#)；[指标定义](#)

它支持的较窄判断是：越来越复杂的任务开始值得交给模型尝试。系统有更多机会完成多步操作、检查结果和修复错误。它没有证明长尾失误已经消失。

长期运行所需的工程能力开始成为通用服务

Anthropic 的长任务研究指出，只给模型更多上下文、让它持续循环，并不能保证得到完整成果。任务需要初始化、分步推进，以及下一次运行能读懂的记录。其后公布的 Managed Agents 将持久会话、运行系统和执行沙箱分开，让部分故障可以单独恢复。[长任务工程研究](#)；[Managed Agents](#)

这些能力对个人助手很关键。一项持续监控不能因为浏览器退出就失去下文；一次工作交接不能只留下“正在处理”。当这类基础设施可以复用，团队才更有机会把精力放在用户目标和真实流程上。

但工程技巧也会过时。Anthropic 的研究说明，一些为旧模型设计的上下文重置办法，在新模型上反而成为负担。较持久的价值在于可靠恢复、权限执行和结果检查，而非某套固定的提示词或多 Agent 编排方式。[长期应用开发研究](#)

数据连接和分发条件正在汇合

MCP 等连接标准的发展降低了接入工具的门槛。2025 年 12 月，MCP 与 AGENTS.md 等进入 Linux Foundation 下的 Agentic AI Foundation。由此可以推断，单纯“接了很多工具”会越来越难构成长期优势，可靠使用工具和处理异常更重要。[MCP 官方公告](#)

与此同时，平台希望减少用户每次都要从空白对话开始的负担。OpenAI 的 Pulse 和 Google 的 Personal Intelligence 分别从主动信息和跨应用个人背景切入。已有账号、历史对话和应用关系，确实能降低初次使用的成本。[Pulse 发布](#)；[Google Personal Intelligence](#)

这说明供给条件正在改善，也说明厂商有争夺日常入口的动机。消费者是否愿意持续授权、是否觉得打扰值得、是否愿意额外付费，仍需另行验证。

三种趋势需要三种证据

三层趋势，三种证据

信心表示证据支持强度，不是成功概率。前两层不能直接证明第三层。

能力普及 · 高信心	工程与能力证据 长任务、持久化状态、事件触发
产品形态变化 · 中等信心	产品与使用证据 功能进入主界面与默认体验
独立商业市场 · 低信心	付费与成本证据 续费、全部成本与利润

第一种是技术能力普及：模型会调工具，任务能持久化保存，系统可以在事件发生后恢复执行。这需要工程文档、可复现的能力测试和运行数据支持，目前已有较强证据。

第二种是产品默认形态变化：用户不再每次手动拼接工具和任务，产品主动提供背景、持续任务和跨应用工作。这需要看功能如何进入主界面、默认体验如何改变，以及用户是否真正使用。多个厂商正在同时调整入口，说明这已经是实际的产品方向，而不只是研究设想。

第三种是独立商业品类成立：用户愿意为持续个人助手额外付费，长期续费后厂商仍有合理利润。这需要付费留存、成本、失败损失和真实替代行为的数据，不能由前两种趋势直接推出。

据此，本报告对三个判断给出不同信心：

- 高信心：长任务执行、持久化状态和事件触发会成为通用能力。多家厂商已有具体产品与工程实践，判断不依赖某一个消费助手成功。
- 中等信心：持续跟进会进入更多产品的默认体验。多个入口正在改变，但用户是否长期采用、是否减少管理时间，仍缺少跨产品的可比数据。
- 低信心：独立的通用个人助手会很快成为大规模、利润稳定的消费品类。付费留存、全部运行成本与事故损失尚缺充分公开数据，而捆绑和融合已有现实路径。

信心等级表示现有证据对判断的支持强度，不是成功概率。

判断二：技术和产品的汇合已经发生，商业上应首先按融合发展来判断。如果独立助手在排除补贴和订阅捆绑后，仍表现出稳定续费、更低的看管成本和可持续利润，这一基准判断就需要上调为“独立品类成立”。反过来，如果所谓主动功能长期无人使用、关闭后结果没有变差，连产品形态发生实质变化的判断也应收缩。

三 Muse 与 dot 的案例揭示了什么

以下案例来自公开报道、厂商演示和帮助文档，没有在本报告中复现实测。独立体验可以证明某次发生了什么；官方材料可以说明产品设计和厂商声称的结果；两者都不能直接给出总体成功率。

Muse 旅行安排：成果有用，仍由用户推进下一步

CNN 记者连接 Google 应用后，请 Muse 就旅行安排给朋友发邮件。朋友回复后，记者又要求它把餐厅推荐和类似选择整理成 Google 文档，最终得到了一份完成度不错的文档。

这个案例说明跨应用组织信息有实际用途，但后半段是用户再次提出要求后启动的。它没有展示 Muse 自动等来回复、判断下一步并完成全程跟进。同一报道中的约会建议还出现了已关闭的场所，说明结果的外观完整与事实可用性需要分别验收。[CNN 体验报道](#)

Muse 学校提醒：发现有价值，最后仍由家人办成

Muse 的官方设计文章讲述了一个内部使用案例：助手从学校邮件和学区信息中发现，体育选拔报名只剩 12 小时。用户联系丈夫，后者在截止前 4 小时提交了表格。

这是“发现重要变化并及时提醒”的正面例子。真正提交报名的是人，初始监控授权、误报情况和长期漏报率没有披露。因此，它支持主动发现可能创造高价值，不支持“自动替家庭办完报名”的说法。

[Muse 官方设计文章](#)

Muse 个性化目录：内容生成与购买之间还有一步

在 10 月 12 日的访谈中，Muse 设计负责人 Alex Cornell 展示了根据 Instagram 收藏整理出的个人购物目录，并描述其每日刷新。他同时说明，尚未把这个目录接入便捷的购买流程。这是厂商人员的使用展示，不是独立长期测试。[访谈及演示章节](#)

它揭示了一个有吸引力的产品方向：利用用户已有的行为记录，减少手动解释偏好的工作。但目录生成、价格跟踪、通知送达和实际购买是不同结果，不能合并计算成一次自动购物成功。

dot 漏开发票：最贴近产品承诺的官方案例

OpenAI 的发布文章称，一位早期用户的 dot 发现了遗漏的出版物发票，准备好发票，并在获批后发送。这条任务链包含主动发现、准备材料、取得同意和对外执行，正好展示了个人助手希望提供的价值。[dot 发布文章](#)

不过，原始邮件、具体工具步骤、耗时和独立确认没有公开。这个案例可以解释产品承诺，不能用来估计普通用户每周能收到多少类似帮助。

dot 日常邮件：办成了事情，仍可能增加管理负担

The Neuron 的首日体验中，dot 写好并发送了一封简短回复。记者事后认为内容合适、也符合自己的表达方式，但原本希望先查看，于是要求今后的外发消息先经审核。

原始指令和授权细节未公开，不能据此判定为越权发送；后续是否稳定遵守新要求也没有得到验证。更有价值的观察是：执行成功与符合用户的审核预期，是两项不同要求。[The Neuron 体验报道](#)

dot 客户工作：抓取完成与监控交付还有距离

SlideFactory 的体验包含两个相关任务。改期客户会议并跟进表格时，账号、邮件应用和电脑授权出现问题，最后停在可复制的草稿，改期和邮件发送并未完成。SEO 监控中，两次网站抓取完成，但 API 接入尚未打通；次日定时报告虽已安排，文章未观察到实际运行和送达。[SlideFactory 体验报道](#)

这是单次、首日、非随机的体验，不能代表当前平均表现。但它明确说明：连接成功、完成一个子步骤、设置日程和最终交付，应该分别报告。若助手说“已安排好”，用户仍需自己确认第二天是否发生，持续跟进的价值就还没有兑现。

Muse 监控中断：最危险的失败可能没有通知

Reuters 在发布期查阅的内部员工帖子中，有用户报告快速售罄商品的监控约 15 分钟后停止刷新，部分错误没有说明，监控有时自行关闭。这是发布期内部测试的转述，具体环境、发生比例和后续修复情况未知，不能据此断言 10 月仍存在相同问题。[Reuters 报道转载](#)

它指出的行业问题却很清楚：对于持续监控，“没有消息”必须能区分为“没有发现变化”和“任务没有正常运行”。助手承诺长期照看一件事，就需要在访问失效、运行停止或结果无法确认时告知用户。

这些案例共同支持一个结论：主动个人 Agent 的需求并非空想，但产品价值取决于整条任务链最弱的一步。判断是否值得委托，必须看到外部结果，也必须计算用户补充背景、重新催促和手动接管的时间。

每一步都要有自己的结果

安排任务、取得授权、执行动作与确认结果分别验收。计划不等于完成。

发现	授权	执行	确认结果
是否真的需要处理 核对遗漏与证据	允许做什么 对象、动作与范围	采取了什么动作 与准备材料分开	最后有没有办成 检查外部结果

四 长期记忆、权限和停止机制决定用户能否放心委托

好的记忆需要及时改正

个人助手需要记住的内容经常带有时间和对象。“这周不要再联系这个人”不能自动变成永久偏好；一次临时预算也不能长期影响所有推荐。记忆的价值来自减少重复解释，风险则是旧信息持续左右新行动。

Muse 的官方文档支持对记忆提出修改或手动编辑，也允许导入其他助手的对话历史。但导入历史不等于不同系统能完整理解彼此的记忆结构。删除原始消息也不一定删除已经学到的信息，遗忘的效果需要单独检查。[Muse 记忆帮助](#)；[Muse 隐私帮助](#)

dot 的公开文档说明，其记忆是摘要，不是所有对话的完整记录；不同任务得到的上下文也可能不同。与 Muse 支持查看和编辑记忆的设计不同，dot 当前不支持用户逐项查看、直接编辑或删除它自身的记忆。用户可以在对话中纠正理解，但这不等于完成了逐项删除。断开连接会阻止新访问，也不会自动清除此前取得的信息。[dot 任务与记忆](#)；[dot 隐私说明](#)

因此，“记得更多”不是充分优势。更有价值的测试是：用户更改预算、日程或权限后，旧信息多久停止影响结果，其他任务是否同步采用新条件，错误内容会不会再次从连接来源中被读回来。

记忆还会产生维护成本。比如，用户为一次出差要求低预算酒店，助手却把它应用到后来的周年旅行；某次应急操作获得一次性批准，未来任务却将它误当成通用权限。信息保存得越久、被更多任务引用，错误就越可能重复出现。这些是假设风险，用于解释为什么长期记忆需要来源、时间和适用范围。

稳定而低风险的偏好，例如文档语言和代码格式，通常适合长期保存。暂时计划、推断和权限则需要更严格地区分。技术上可以称为持久化状态，但产品必须让用户的纠正真正影响后续行为。

判断三：长期记忆的优势最终要体现为少解释、少犯同样的错。如果开启记忆只是提高召回率，却增加了用户反复检查旧信息的时间，记忆就没有创造净收益。如果一次纠正能可靠地同步到有关任务，且无需用户逐一修改，它才可能形成真实的长期优势。

审批应减少不必要的打断，关键动作仍要守住

Muse 的官方安全说明描述了独立权限系统：需要批准的动作暂停等待决定，授权可限定为一次、某个任务或一段时间。dot 的公开安全材料则区分后台只读研究与后续执行，前者不能自行扩展为对外发消息或修改内容的权限。[Muse 权限设计](#)；[dot 安全设计](#)

这些是控制设计，不能直接证明实际运行中零误判。OpenAI 在 Codex 自动审核研究中报告过内部审批中断大幅下降，也同时承认自动审查可能受到误导。特定内部流量的结果不能外推为个人助手普遍安全或普遍省时。[自动审核研究](#)

用户需要的体验是：普通、已获授权的步骤不必反复看管；关键动作不会悄悄发生。只减少确认次数而不检查错放行的动作，可能把看管时间换成事故和修复成本。

停止必须覆盖真正仍在运行的工作

一个容易被忽略的问题是，暂停主对话是否也停止了已经委派的任务和未来的定时执行。dot 的公开控制说明明确区分这几类工作，暂停主任务不会自动停掉所有委派任务或未来日程，已完成的动作也不会因此撤销。[dot 控制说明](#)

这类细节直接影响用户能否理解自己的控制范围。评价个人助手时，除了问“能不能开始做”，还应问“能不能改、能不能停、停后是否真的不再行动”。

五 商业可行性取决于省下的时间是否超过新增成本

用户购买的是少操心，算力消耗只是成本之一

用户亲自发出请求时，已经帮系统筛选过一次需求。系统主动工作时，需要自己寻找候选事项，判断是否重要、是否重复、是否已经完成。即使最后没有值得报告的事，这个过程也会消耗计算和工具资源。

从用户角度看，价值包括少漏掉的重要事项、少做的重复劳动和少花的跟进时间；代价则包括订阅费用、补充背景、阅读建议、审批、纠错，以及误行动造成的损失。厂商还要承担模型推理、外部工具、持续监控和异常处理成本。

因此，评估对象应是“一件真正办成、且用户认可的事情”。后台工作很久或生成很多内容，都不能直接说明净收益为正。Anthropic 的长期应用开发研究曾列出更复杂系统花费 6 小时、200 美元，单 Agent 花费 20 分钟、9 美元的单次示例；产出质量和范围并不相同，无法据此作公平的成本比较，但足以提醒：多 Agent 和运行更久并不会自动更便宜。研究案例

主动扫描面临一个容易被忽略的比例问题

主动请求本身是一道免费的需求筛选。用户从大量事件中挑出一件值得处理的事，也表明此刻愿意为它花注意力。后台助手失去了这道筛选，必须从大量普通信息中寻找少数重要事项。

以下是纯说明性计算，不是测得的产品表现：假设 1,000 个事件中只有 10 个值得提醒，助手找到了其中 8 个；即使它只把其余 990 个事件的 1% 错判为重要，也会增加约 10 个无用提醒。最终约 18 条提醒中，不到一半真正有用。识别能力看起来不错，日常体验仍可能很烦。

这解释了为什么更多数据源不一定立刻带来更好体验。扩大扫描范围会增加发现机会，也会增加重复、过期和用户早已知道的信息。研发团队里，反复报告已经有人处理的故障，可能不如只盯住尚未指派且影响发布的问题有用。

不能由此推导“越安静越好”。对于漏报代价很高的事项，用户可能接受更多提醒；固定摘要也可以将多条低紧迫性消息集中处理。系统需要根据后果和用户安排决定何时打断。

判断四：主动产品的长期优势会体现在提醒更值得看，而非主动发起更多工作。如果在控制模型、数据来源后，持续助手不能比“定时摘要加显式任务”减少遗漏和总管理时间，它的主动性就没有证明新增价值。反过来，有效提醒率提高、重复提醒下降且重要事项没有漏掉，才是更强的产品证据。

初期更有希望的是结果容易核验的工作

代码任务通常有版本差异、测试和运行结果，用户至少有相对明确的检查对象。个人事务中，提醒是否值得打断、跟进是否合适、礼物是否得体，反馈可能很晚才出现，也更依赖具体关系。

这不意味着所有代码任务都客观，也不意味着生活任务不能自动化。它意味着早期商业化应优先选择完成条件清晰的事情：核对遗漏材料、发现截止日变化、跟进已获授权的事项、准备可审核的内容、检查订单或文档状态。

这些场景的价值不止于提醒，而在于能够把发现、准备、执行和确认结果连起来。相比“理解用户的一切”，边界清楚的承诺更容易衡量，也更容易让用户逐步扩大授权。

等待和交接可能比持续推理更值得付费

很多事情失败在任务之间：资料准备完了，却没有人确认下一步；PR 已建立，但 CI 失败后没有接着修；对方已经回复，原任务却没有恢复。这些问题通常已有目标和可观察的状态，比让助手不断发明新项目更容易衡量。

以假设研发任务为例，用户要求“修复问题并让 PR 达到可合并状态”，写好代码只是其中一步。助手还需要等待 CI、判断失败原因、在授权范围内修复并验证。旅行安排也一样：日期等待同行者确认，收到回复后才继续更新计划，需要购买时再按既有权限处理。

等待可能持续数天，但没有必要数天不停推理。更合理的系统保存任务状态，明确外部依赖，在事件到来或约定时间恢复工作。这里需要持久化状态、任务编排和可观测性，让用户能知道它在等什么、有没有正常等、何时需要人介入。“常驻”描述的是任务关系持续存在，不等于计算始终满负荷运行。

判断五：近期更容易形成稳定收益的，是目标清楚的跨日跟进和交接。若普通规则自动化能处理同样的例外，成本更低且结果更稳定，就没有必要使用通用 Agent。只有在输入多变、需要理解上下文又能核验结果时，Agent 的判断能力才更可能值得付费。

实际使用正在增长，独立市场仍需付费数据验证

OpenAI 在 2026 年 6 月自报 Codex 已超过 500 万周活用户，知识工作者约占 20%，任务包括研究、报告和表格。Anthropic 同月发布的遥测报告，也显示用户在 Claude Code 中对多种非编码产出给予更高自主程度。[Codex 知识工作数据](#)；[Anthropic 经济研究](#)

这些证据比产品发布更接近实际使用，但仍有明确限制：前者是厂商统计，后者存在用户自选择，都不是个人助手长期付费留存的证明。它们更支持一个判断：人们确实愿意通过能执行任务的系统委托知识工作，个人助手可以在这个基础上发展。

六 谁更可能保留长期价值

以下是基于现有证据的商业判断，不是已经得到验证的市场份额预测。

模型与平台厂商有明显起步优势。对模型与推理资源的直接控制、现成账号和分发渠道，都有助于降低首次使用门槛。已有记忆与连接也能减少用户重复提供背景的工作。但拥有入口不代表用户愿意接受长期打扰，更不代表平台能处理每个行业的真实流程。

通用运行系统会继续重要，但部分能力可能成为标准配置。任务恢复、权限检查、审核和跨环境执行有长期价值；只为某代模型补短板的技巧更容易贬值。独立团队需要证明自己的系统在某类任务上更可靠、更省钱或更易检查。

掌握具体业务流程的公司仍可能有强位置。如果一家公司拥有权威数据、交易入口、明确的完成标准和异常处理经验，通用助手很可能还要通过它办事。连接更多工具不会自动取代这些能力。

个人关系可能带来切换成本，但不是天然壁垒。如果长期使用确实减少解释、纠正和看管，用户可能愿意留下。反过来，如果记忆易迁移、多个助手可以并用、结果差异不大，“记得我”未必足以支撑单独订阅，更不能直接称为网络效应。

判断六：平台占据入口的优势很强，但行业价值不会仅由入口决定。如果垂直产品在同一流程中能更快接入有效数据、更少出错、更容易确认结果，它仍有收费空间。若通用平台以相近成本完整覆盖这些例外，垂直产品只剩转发请求和展示结果，其独立价值就会削弱。是否掌握真实流程，比是否拥有一个拟人化头像更重要。

由此看，最可能的近期格局是融合而分工：个人入口发现和跟进事情，通用 Agent 负责执行，专业服务完成关键业务步骤。生活、工作和组织权限仍需要分开管理，用户未必愿意把所有事务交给同一个系统。

七 哪些反例会改变判断

个人助手并不是第一次获得大型平台支持。Cortana 曾拥有强大的分发条件，微软仍在 2023 年起结束多个界面的支持，相关 Outlook 语音功能也于 2024 年退出。它不能证明新一代模型会失败，却足以否定“有用户入口就必然成功”。[微软支持记录](#)

对生产率的主观感受也不可靠。METR 的早期随机试验中，熟悉仓库的开发者使用当时的 AI 工具后，完成任务实际更慢，却主观觉得更快。这个旧实验不能代表 2026 年工具；METR 后续也指出，新样本的选择偏差使当前收益难以可靠估计。它留下的方法提醒是：要把阅读、检查和返工算进去，不能只问用户“感觉省时吗”。[早期随机试验](#)；[2026 年方法更新](#)

接下来有三种值得观察的走向。

- 融合成为主流。非编码后台任务在现有订阅中普及，个人助手负责减少配置和跟进工作。独立产品依靠更深的业务流程或更好的持续使用体验收费。

- 独立消费市场成立。用户持续付费，逐步增加连接来源，在明确范围内扩大授权；助手带来的实际收益超过打扰、监督和运行成本。

- 主要停留在建议和简报。产品能发现话题，却难以交付结果；用户频繁接管或关闭通知，厂商通过减少执行和限制用量控制成本。

本报告以第一种走向作为近期基准：持续助手能力进入现有产品，独立公司在特定流程中寻找收费空间。第二种需要更强的续费与成本证据；第三种是必须防范的实际风险。

八 下一步应该看什么数据

最值得跟踪的不是模型调用次数、生成条数或后台在线时长，而是下面这些结果。

1. 90 天付费留存与实际续费。区分独立购买、订阅捆绑和免费试用，避免把平台已有用户误当成新需求。

2. 每位用户每周真正办成多少件有用的事。预先定义完成条件，把“建议已提出”“任务已安排”和“结果已送达”分别统计。
3. 每办成一件事，用户花了多少分钟。包含设置、解释、审批、检查、催促和纠错，并与不用助手时比较。
4. 主动建议最后有多少被采纳并产生结果。同时看误报、重复提醒、已完成事项被重启，以及用户为何忽略建议。
5. 持续任务是否可靠运行。记录漏跑、访问失效、通知延迟和故障告知，检查系统能否让用户区分“无变化”与“未运行”。
6. 每件完成事项的全部成本。厂商看推理、工具和支持成本；用户看时间、费用和事故损失，两边都不能只统计成功步骤。
7. 纠正与停止是否生效。用户修改偏好或撤回权限后，检查其他任务和后续日程是否采用新要求，是否仍出现旧行为。
8. 增加连接与授权是否真的改善结果。看同一用户是否更省时、是否更愿意留下，不能仅把权限更多当成信任更高。

一种有辨别力的测试，是让条件相近的用户分别使用定时摘要和持续助手，控制模型能力与数据权限，在数周内比较遗漏事项、用户投入时间和最终结果。这样的比较才能回答：新增价值来自主动跟进，还是只来自更方便地汇总信息。上述指标和测试是建议，不是本报告已经完成的实测。

结语

主动个人 Agent 的潜力在于替用户承担一部分发现、记住和跟进事情的工作。Claude Code、Codex 已经证明这条能力路径与编程之外的知识工作有关；Muse、dot 正在尝试把它组织成持续的个人助手体验。

真正的竞争会落到几个朴素的问题上：它找到了值得做的事吗，最后办成了吗，用户是否少操心了，出错和停止时是否说得清楚。能够长期给出好答案的产品，才可能获得更大授权和稳定收入。

最值得下注的是可靠的持续跟进，而不是“永远在线”的形象。近期机会在有明确目标、可检查结果跨日事务；长期赢家会让用户少漏掉重要事情、少管理一个工具，并以可持续的成本做到这些。

附录 市场格局与四条产品路线

截至 2026 年 10 月 14 日。以下按产品路线选例，不是功能排行榜；能力来自官方公开资料，未经本报告实测。战略意义为分析判断。Muse、dot 之外的竞争者，争夺的也不都是“再安装一个私人助理”。

一、平台个人入口：把 Agent 收进原有应用

Gemini Spark 是最直接的参照，而且品牌本身已发生变化。Google 当前帮助页明确说，Spark 已改名为 Task，进入 Gemini 的 Chat/Task 切换入口，原任务与计划保留。它结合个人上下文、连接应用和执行环境，远程浏览器可在用户设备关闭后继续工作。目前要求个人账号、Pro 或 Ultra 订阅；工作、学校账号不适用，部分地区仍未开放。[Google 当前文档](#)

这个变化比又增加一个功能更值得注意：Google 正把“独立命名的新助手”变为用户已有产品中的工作方式。对新品牌而言，竞争门槛从能不能后台执行，升为为什么用户要另建一套关系与上下文。但品牌合并不能被当作需求成功或失败的证明。

Microsoft Copilot Cowork 则从 Microsoft 365 的文件、邮箱和日历切入，个人账号版本仍标为 Preview。官方正在把 Copilot Tasks 用户迁过去；会话会迁移，旧记忆不迁，定时任务迁移后默认关闭。这是个具体警示：长期助手的连续性还取决于产品迁移，而非只取决于模型记忆。当前迁移说明也写明远程浏览器尚待推出，不能把“可定时”直接写成任意网页任务都能离线完成。[Microsoft 迁移说明](#)

二、工作流产品：先接管团队的待办，再扩大责任范围

Claude Cowork/Tag 展示两种延伸。Cowork 的定时任务现已支持远程运行；需要本地环境的任务仍有例外，Pro、Max 在 10 月 6 日前建立的本地计划不会自动变成云端任务。与此同时，Chat 和 Cowork 正在整合为统一体验。定时任务；统一入口

6 月 23 日发布的 Claude Tag 更直接进入 Slack：在获授权的频道内积累上下文，开启 ambient（主动关注频道变化）行为后主动跟进沉寂但未解决的事项。它的记忆和工具可以按管理员指定的频道范围隔离。[Tag 发布公告](#)

Lindy 当前也主推团队助手，以 Slack 协作、共享会议上下文、定时任务和可复用技能为中心，同时保留 iMessage、邮箱入口。[Lindy 官网](#)

三者的共同切口是已有组织流程：周报、客户交接、待回复事项。商业意义在于用户不必先把全部生活交给助手，团队可以围绕少数可见产出决定是否扩大使用。区别也由此产生：集体上下文有利于协作，却要求清楚回答“谁能指挥它、谁能看见记忆、谁对结果负责”。这类产品不能仅因不是私人陪伴形象，就被排除在主动 Agent 的竞争之外。

三、自托管：把“自己的助手”做成可掌握的基础设施

OpenClaw 是不同性质的参照。其官网称项目由独立基金会维护、采用 MIT 许可；OpenAI 为捐助方之一，不是产品所有者。用户自行运行服务、连接消息入口与执行工具；主动检查本身会调用 Agent，并消耗推理资源。[OpenClaw 官网](#)；[主动检查文档](#)

它提供的并非只有更便宜的软件，而是更强的可修改性和部署控制。代价是用户也接过维护、升级、凭据和故障恢复责任；自托管不代表模型调用不出网，更不代表自动安全。

对普通用户产品的意义是一个清晰的价格参照：开源方案使“有记忆、能调工具、能定时”越来越难单独收费。商业产品必须让人感到省下的运维和监督成本值得付费。反过来，自托管也未必能取代低摩擦服务，因为多数人购买的是办成事，而不是一套要维护的 Agent 系统。

四、轻量消息入口：把安装软件变成添加一个联系人

Poke 把入口放在日常消息应用，用 Recipes 把工具连接与工作方式包装成可复用起点。其额度文档区分得很明确：免费聊天与 Pro 加入的后台自动化并不是同一交付范围；高端方案还提供真人介入。因此评价“完成任务”时，需区分模型独立执行与混合服务，而不能只看结果展示。[Poke 官网](#)；[额度机制](#)；[服务 FAQ](#)

Pally 则突出人物、承诺和生活安排，通过 iMessage、RCS、Telegram 交互；官网还提出经用户授权的 Pally 间协作，让不同关系有不同共享范围。它当前宣称主体免费、Cloud Mac 另收费，计划从经其完成的购买向商家收费。这是商业模式计划，不是已获验证的收入或毛利。[Pally 官网](#)

两者试图降低“想到一件事就交代”的摩擦。其机会在于消息关系天然支持跨天跟进；局限是用户也在同一入口接收朋友、工作和广告，主动消息必须争取注意力。Pally 的商家收费设想还引出另一项考验：推荐、用户利益和交易变现如何保持一致。官网承诺商家不能购买推荐排序，仍应看实际执行与披露。

看格局时，应当分开问的问题

平台产品问的是“能否利用已有账号与订阅，成为默认入口”；工作流产品问的是“能否在有边界的流程里扩大责任”；自托管问的是“控制权是否值得维护成本”；消息型助手问的是“能否把交代和跟进做得足够自然”。这些路线会借用彼此能力，却不必最终只剩一种界面。

因此，Muse／dot 的潜在对手既有直接的私人助手，也有逐步吞入个人事务的工作产品。最有辨识度的比较，不是列出谁支持更多工具，而是选择同一件跨天事务，看哪个产品最少需要用户重述背景、催促和收拾残局。

来源目录

1. Codex 桌面应用
<https://openai.com/index/introducing-the-codex-app/>
2. Codex 4 月更新
<https://openai.com/index/codex-for-almost-everything/>
3. Claude Code routines
<https://claude.com/resources/articles/introducing-routines-in-claude-code>
4. Claude Code 记忆文档
<https://code.claude.com/docs/en/memory>
5. dot 系统卡补充
<https://deploymentsafety.openai.com/gpt-6-astra/sec%3Aappendix-sol-luna>
6. METR 研究更新
<https://metr.org/blog/2026-1-29-time-horizon-1-1/>
7. 指标定义
<https://metr.org/time-horizons/>
8. 长任务工程研究
<https://www.anthropic.com/engineering/effective-harnesses-for-long-running-agents>
9. Managed Agents
<https://www.anthropic.com/engineering/managed-agents>
10. 长期应用开发研究
<https://www.anthropic.com/engineering/harness-design-long-running-apps>
11. MCP 官方公告
<https://blog.modelcontextprotocol.io/posts/2025-12-09-mcp-joins-agentic-ai-foundation/>
12. Pulse 发布
<https://openai.com/index/introducing-chatgpt-pulse/>
13. Google Personal Intelligence
<https://blog.google/innovation-and-ai/products/gemini-app/personal-intelligence/>
14. CNN 体验报道
<https://localnews8.com/money/cnn-business-consumer/2026/09/23/meta-says-its-muse-ai-agent-can-do-things-for-you-i-put-it-to-the-test/>
15. Muse 官方设计文章
<https://introducing.muse.ai/>
16. 访谈及演示章节
<https://www.lennysnewsletter.com/p/how-meta-designed-muse-an-ai-agent>
17. dot 发布文章
<https://openai.com/index/introducing-dots/>
18. The Neuron 体验报道
<https://www.theneuron.ai/explainer-articles/openai-dots-hands-on-review-herman/>
19. SlideFactory 体验报道
<https://www.theslidefactory.com/post/openai-dots-review>
20. Reuters 报道转载
<https://www.investing.com/news/stock-market-news/meta-launches-ai-agent-that-can-access-other-apps-to-send-emails-make-payments-4892258>

21. Muse 记忆帮助
<https://www.meta.com/help/artificial-intelligence/995796179982326/>
22. Muse 隐私帮助
<https://www.meta.com/help/artificial-intelligence/1047255454427887/>
23. dot 任务与记忆
<https://learn.chatgpt.com/docs/dots/tasks-and-memory>
24. dot 隐私说明
<https://help.openai.com/en/articles/20001529-dots-privacy-security-and-safety-faqs>
25. Muse 权限设计
<https://research.meta.ai/blog/security-and-safety-for-ai-agents-our-approach-with-muse>
26. dot 安全设计
<https://openai.com/index/how-we-build-safety-security-and-privacy-into-dots/>
27. 自动审核研究
<https://alignment.openai.com/auto-review/>
28. dot 控制说明
<https://learn.chatgpt.com/docs/dots/controls>
29. Codex 知识工作数据
<https://openai.com/index/codex-for-knowledge-work/>
30. Anthropic 经济研究
<https://www.anthropic.com/research/economic-index-june-2026-report>
31. 微软支持记录
<https://support.microsoft.com/en-us/cortana/end-of-support-for-cortana>
32. 早期随机试验
<https://arxiv.org/abs/2507.09089>
33. 2026 年方法更新
<https://metr.org/blog/2026-02-24-uplift-update/>
34. Google 当前文档
<https://support.google.com/gemini/answer/17094507?co=GENIE.Platform%3DDesktop&hl=en>
35. Microsoft 迁移说明
<https://support.microsoft.com/en-us/microsoft-365-copilot/learning/changes-microsoft-copilot-app>
36. 定时任务
<https://support.claude.com/en/articles/13854387-schedule-recurring-tasks-in-claude-cowork>
37. 统一入口
<https://support.claude.com/en/articles/16761823-claude-cowork-and-chat-are-one-claude>
38. Tag 发布公告
<https://www.anthropic.com/news/introducing-claude-tag>
39. Lindy 官网
<https://www.lindy.ai/>
40. OpenClaw 官网
<https://openclaw.ai/>
41. 主动检查文档
<https://docs.openclaw.ai/gateway/heartbeat>

42. Poke 官网

<https://poke.com/>

43. 额度机制

<https://poke.com/docs/usage-and-resets>

44. 服务 FAQ

<https://poke.com/faq>

45. Pally 官网

<https://pally.com/>