AI 技术每日分析：DeepSeek 开源多任务智能体原型，拆分任务分发与监督

中国高技术产业发展促进会新质生产力工作委员会

博雅云创 & 中科创新驱动

2026 年 11 月 2 日

摘要

DeepSeek 开源了多任务智能体研究原型，用任务路由、独立执行会话和监督机制处理长时间工作。它的 Search、Fetch 工具也从 11 月 1 日起按次计费。开发者社区还在讨论 Apple Silicon 本地模型服务和代码架构可视化工具；商汤开放的循环层多模态模型，则提供了一个观察参数存储规模与实际计算深度如何分开设计的样本。这几件事落到开发者手里，问题都很具体：任务能否可靠完成，外部调用花多少钱，复杂代码还能不能看得明白。

Contents

1. DeepSeek 公开多任务智能体原型：任务分发、执行和监督各有分工 **2**
2. DeepSeek 搜索与抓取工具 11 月 1 日起计费：成功请求按次算 **2**
3. mlx-serve 受到关注：Apple Silicon 本地模型也想共用一套 API **3**
4. Floorplan 用 Claude 或 Codex 画代码地图：先弄清系统怎么连起来 **3**
5. 商汤循环层模型：9B 权重，前向计算执行 56 层 **4**
趋势观察 **5**
参考资料 **5**

1. DeepSeek 公开多任务智能体原型：任务分发、执行和监督各有分工

DeepSeek 在 GitHub 公布了 dsh-superagent-prototype 研究原型。按项目文档，Router 接收并分类请求，Worker 在相对独立的任务会话中执行，Watchdog 等监督机制观察任务状态和异常。把工作拆开后，并行任务、失败恢复和需要人介入的环节更容易单独处理，也不用把所有需求都塞进一段越拉越长的对话。文档还把操作授权说得很清楚：没有用户直接请求，也没有预先授予的规则，智能体就不应自行扩大行动范围。目前的代码与架构文档面向研究和工程试验，不能直接当作已完成安全认证的企业平台。[1][2]

做长任务时，上下文切换、状态管理、重复通知和失败恢复都可能拖住团队。这个原型把执行过程分成可观察的工作单元，代码修改、测试、资料整理分别推进时，谁做了什么、哪里出了问题，比较容易追踪。真要部署，还得隔离操作系统权限、限制工具调用，管住不同工作会话之间的共享文件和密钥。项目的任务分页等工程细节已经出现在公开问题列表里，社区正在给出反馈。[3]

2. DeepSeek 搜索与抓取工具 11 月 1 日起计费：成功请求按次算

DeepSeek 官方 API 文档写明，自北京时间 11 月 1 日零时起，在 DeepSeek Harness、Claude Code 和 Codex 等环境中经服务端调用的 Search 与 Fetch 接口按次计费。两项工具的中国区标价均为每 1000 次成功请求 5 元。Search 返回网页检索结果，Fetch 抓取网页正文并转成 Markdown。请求成功返回且有实际结果才计费；搜索为空、超时或请求失败不计费。计费此前已公告，11 月 1 日是正式生效的时间。[4][5]

单次调用不贵，持续运行的研究智能体却可能反复搜索、重试和抓取网页，还要支付模型阅读内容的 Token 费用。按一项研究任务成功调用 30 次 Search、20 次 Fetch 估算，仅工具请求费约 0.25 元，模型推理与后续处理另计。因此，除了 Token 消耗，业务平台还可以按任务记录检索次数、有效资料比例、重复抓取率和交付成功率。价格摆在台面上，开发者也能据此比较自建索引、第三方检索接口和本地知识库。

3. mlx-serve 受到关注：Apple Silicon 本地模型也想共用一套 API

11 月 1 日，开发者社区仍在讨论开源项目 mlx-serve。它使用 Zig、Swift 等技术，在 Apple Silicon 上提供本地模型服务；接口兼容 OpenAI、Anthropic 和 Ollama，覆盖聊天、编码、语音及部分多媒体模型工作流。项目仓库已有约 2300 个 GitHub 星标。这是持续迭代项目近期得到的关注，并不是 11 月 1 日首次发布。[6]

项目想解决的是接入问题：开发者在 Mac 上用统一服务接口调用本地模型，桌面应用就不必各自维护模型加载和推理逻辑。对个人开发者和小团队来说，Mac 既是工作站，也可能兼做轻量模型服务端；数据留在本地、离线可用和小规模使用成本，都是考虑因素。桌面智能体接入统一 API 后，切换云模型与本地模型也少了重复改接入代码的工作。不过，实际性能仍要看统一内存、模型量化精度、上下文长度和并发数量。仓库列出的功能和性能基准，最好在要使用的硬件上分别复测。

4. Floorplan 用 Claude 或 Codex 画代码地图：先弄清系统怎么连起来

开源工具 Floorplan 近期也在开发者社区受到关注。项目说明称，它可以调用 Claude 或 Codex 阅读代码库，把文件、业务模块、前后端请求和

外部服务画成可交互的关系图, 再用普通语言解释它们如何配合。GitHub 仓库已有约 1600 个星标, 还提供在线示例。它是持续传播的开源工具, 并非当天发布的大型商业产品。[7]

代码生成快了，新加入项目的工程师、交接团队，甚至原作者，未必能马上说清一次页面操作会经过哪些服务、数据库和权限。能点开查看的系统地图，可用于代码审查、变更影响分析和新人入职培训。但这张图是模型读源码生成的；碰到复杂动态调用或运行时配置，仍要拿测试、链路追踪和真实部署配置核对。写代码之外，团队也需要持续看懂已经写下的系统。

5. 商汤循环层模型：9B 权重，前向计算执行 56 层

商汤开源团队 10 月 30 日公开了 SenseNova-Loop-9B-Preview，相关代码和模型说明近期仍在研究者中受到关注。按官方介绍，模型约有 90 亿存储参数；32 组不同的解码器层中，24 层循环执行两次，前向计算总计经历 56 层执行。它支持文本和图像输入，训练上下文长度最高 262144 个 Token，采用 Apache-2.0 许可。开发文档还介绍了混合注意力和两路残差流机制。这是 10 月 30 日发布的项目，本期把它作为模型架构路线的研究案例，而非当天新发布的消息。[8][9]

复用权重可以让存储容量和计算深度分开设计：权重占用减少，有利于模型加载与显存配置；层重复执行，也会增加计算和时延。团队披露的能力指标，还需要结合硬件、吞吐和具体任务复现。选开源模型时，企业不能只看参数量，显存占用、任务完成率、每秒 Token 和工具调用稳定性也得一起算。

趋势观察

从 DeepSeek 的任务调度和工具计费，到本地推理、代码地图，开发者要处理的是可靠性、成本、部署和可维护性。AI 应用做连续任务时，能否恢复失败任务、控制检索次数、看清代码依赖，比一段演示里的顺畅回答更值得在工程现场逐项检查。

参考资料

1. DeepSeek ／ GitHub ｜《dsh-superagent-prototype: Towards a More Autonomous Agent System》｜ 2026-10-31 公开，2026-11-02 查阅｜项目架构、研究原型定位
   https://github.com/deepseek-ai/dsh-superagent-prototype

2. DeepSeek ／ GitHub ｜《000 Philosophy》｜ 2026-10-31 前后更新｜任务授权、人员注意力与执行原则
   https://github.com/deepseek-ai/dsh-superagent-prototype/blob/main/docs/000-philosophy.md

3. DeepSeek ／ GitHub Issues ｜项目公开问题列表｜ 2026-11-01 查阅｜项目早期工程反馈及生产化边界
   https://github.com/deepseek-ai/dsh-superagent-prototype/issues

4. DeepSeek API Docs ｜《网页搜索与抓取》｜ 2026-11-01 计费生效，2026-11-02 查阅｜接口功能、价格和计费条件
   https://api-docs.deepseek.com/zh-cn/guides/search/

5. 澎湃新闻｜《DeepSeek：网页搜索与抓取单价为每 1000 次成功请求 5 元，11 月 1 日起生效》｜ 2026-10-18 ｜生效日期及原公告交叉核查

https://www.thepaper.cn/newsDetail_forward_34276589

6. ddalcu ／ GitHub ｜《mlx-serve: Native LLM inference server for Apple Silicon》｜持续更新，2026-11-01 关注｜本地推理、API 兼容与开源社区动态

   https://github.com/ddalcu/mlx-serve

7. haydenschmitty ／ GitHub ｜《Floorplan: See how your app actually fits together》｜持续更新，2026-11-01 关注｜代码地图、项目结构解释与开发者工具

   https://github.com/haydenschmitty/floorplan

8. OpenSenseNova ／ GitHub ｜《SenseNova-Loop-9B-Preview》｜ 2026-10-30 ｜循环层架构、权重与执行层数

   https://github.com/OpenSenseNova/SenseNova-Loop

9. SenseNova ／ Hugging Face ｜《SenseNova-Loop-9B-Preview Model Card》｜ 2026-10-30 ｜模型权重、部署要求与许可

   https://huggingface.co/sensenova/SenseNova-Loop-9B-Preview