AI 技术每日分析：OpenAI 披露 AMD 旗舰 GPU IP 改进：最高 14% PPA 改善

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 19 日

摘要

OpenAI 公开了用前沿模型改进芯片设计的工程细节：Jalapeño 推理芯片的向量与矩阵单元实现了面积优化，与 AMD 合作的部分 GPU IP 设计改善幅度达到 14%，部分修改已被纳入制造或后续生产文件。Google Labs 则在 11 月 18 日推出 Stitch Loop 实验项目，让产品智能体接入 GitHub、运行日志与用户反馈，持续识别问题、编写修复代码并提交待审核的拉取请求。Google DeepMind 同日扩大 Genie 4 交互世界生成体验，NVIDIA 开始销售搭载最高 128GB 统一内存的 RTX Spark 紧凑桌面设备。模型能力延伸到工程工具、产品上线后的持续运行和本地执行环境，相关更新的价值可以通过最终的设计文件、代码修改和运行记录观察。

Contents

一、OpenAI 披露芯片设计进展：RTL 改写进入真实工程流程 **2**
二、Google Stitch Loop：把上线后的软件运营纳入 Agent 循环 **2**
三、Genie 4 扩展交互世界生成：实时编辑与原生音频 **3**
四、RTX Spark 紧凑桌面落地：本地智能体获得新硬件载体 **4**
五、SenseNova-Loop-9B 补充量化权重：共享层的端侧模型路线 **4**
趋势观察 **5**
参考资料 **5**

一、OpenAI 披露芯片设计进展：RTL 改写进入真实工程流程

OpenAI 于 11 月 18 日发布《Advancing chip design with OpenAI models》。官方披露，自 2025 年初便使用 GPT-4o 辅助自研推理芯片 Jalapeño 的设计，优化对象包括浮点加法器、乘法器、低精度点积电路及流水线。公司估算，模型提出的改动令向量单元标准单元面积降低约 13%、矩阵单元降低约 10%；其中一项乘法器重设计在综合环节节约约 15% 单元面积，并将延迟从两个时钟周期缩短至一个。相关优化已经经历 Jalapeño 芯片的实际硅片验证。

与 AMD 合作的工作聚焦 PPA，即功耗、性能、面积的组合优化。OpenAI 称模型通过改写 SystemVerilog RTL，使 AMD 旗舰 GPU 部分 IP 相对高度优化的基线取得最高 14% 的 PPA 改善，部分改动已经进入近期送制造的设计，另一些进入后续生产文件。Synopsys 工具用于检验 RTL 修改前后的功能等价性，并评估权衡。公开研究还对 PicoRV32 与 Ibex 两个开源 RISC-V 内核进行了 9,881 次评分运行。这里最可用于工业研发的经验，是把模型探索、综合仿真、形式校验、物理设计约束和工程师评审组成可追踪的工具链。

原　文：https://openai.com/index/advancing-chip-design-with-our-models/

二、Google Stitch Loop：把上线后的软件运营纳入 Agent 循环

Google Labs 于 11 月 18 日公开《Introducing Stitch Loop: A new way to build and evolve products》。这个新实验把软件仓库与运行指标连接起来，让智能体在产品交付后持续工作。团队可以接入 GitHub，并结

合 180 多种集成、MCP 服务或本地 CLI 上传的上下文，包括日志、用户行为、反馈和业务指标；系统围绕安全、延迟、无障碍等目标运行长期循环，提出改进建议、备选方案，在隔离云端环境中编写修改代码并创建拉取请求。截图、视频以及相应的证据可以随变更一起提供，最终合并由开发人员审核。

Stitch Loop 的四步过程包括建议、方案、构建和复盘。传统软件监控提供告警，研发团队再人工从日志进入代码；新实验试图把监测指标、代码上下文、修复实施与部署后的效果复核接到一条任务线上。Google 同时开放全球候补申请，符合条件的 Google AI Ultra 用户优先，因此本次属于产品实验与申请开放，不能写成所有企业已经全面部署。

原文:https://blog.google/innovation-and-ai/models-and-research/google-labs/stitch-loop/

三、Genie 4 扩展交互世界生成：实时编辑与原生音频

Google DeepMind 于 11 月 18 日宣布 Genie 4 进入新的 Project Genie 体验，面向符合条件的 Google AI Ultra 订阅者陆续开放。该体验支持使用自然语言实时修改交互环境、加入原生音频，并分享生成的世界；同时更新的 Genie 3 支持更长交互时段、全景接地和关联世界序列。对机器人、自动驾驶和工业仿真而言，世界模型研究提供了从静态图像走向可操作场景的参考，但本次公告所确认的是面向用户的交互世界生成体验，尚非工业设备仿真平台的交付通告。

研究环境的构建涉及场景几何、动作后状态变化、视角一致性和事件连续性。工业应用可能把这类模型用于场景原型、展示或训练数据准备，再与多物理场求解器、设备参数和真实试验数据衔接。这些工程环节有明确的专业分工，产品本身的公开能力和工业应用的推演在时间与性质上分别记录。

原文:https://blog.google/innovation-and-ai/models-and-research/google-deepmind/genie-4-now-available-project-genie/

四、RTX Spark 紧凑桌面落地：本地智能体获得新硬件载体

NVIDIA 于 11 月 18 日宣布，ASUS、Dell 和 MSI 的 RTX Spark 紧凑桌面产品开始销售，其他厂商产品将在后续推出。官方介绍的配置包括 20 核 Grace CPU、Blackwell RTX GPU、6,144 个 CUDA 核心，最高 128GB 统一内存与 10GbE 网络能力；厂商将其定位为本地运行模型、持续运行个人智能体、视频图像处理及小规模微调的专用 Windows 桌面。NVIDIA 表示，这类设备能够承载约 120B 参数级模型的特定工作流，并通过 Switchyard 在本地和云端模型之间选择执行位置。

对有本地数据约束的工程团队，这样的终端提供了一种有吸引力的部署位置：本地文档、代码和设计资产可留在现场，模型通过 CLI、MCP 和已有开发环境完成检索或处理。实际可运行模型的参数规模、上下文长度与吞吐仍然取决于量化方案、缓存和任务性质。11 月 18 日的新事件是几家厂商正式开售紧凑桌面，不把未来型号预告记为已经上市。

原 文：https://blogs.nvidia.com/blog/rtx-spark-desktop-local-ai-available-now/

五、SenseNova-Loop-9B 补充量化权重：共享层的端侧模型路线

SenseTime 此前于 11 月 17 日正式发布 SenseNova-Loop-9B，11 月 18 日其 Hugging Face 模型卡新增 AWQ 和 GGUF 权重下载链接。这一更新属于部署资源的新增，不重复把模型首发包装成 11 月 18 日新闻。模型以约 90 亿参数采用循环复用中间层权重，并结合 mHC 连接，使计算

路径加深而不按同样比例增大参数。官方表示 4bit 权重约占 4.9GB，部分部署配置的显存要求低于 10GB。模型在编程、工具调用及推理任务上的表现属于提供方评测数据。

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

趋势观察

这条路线适用于企业将通用任务拆分到小型岗位模型的探索。是否选用循环模型，需要结合工具调用成功率、提示词长度、吞吐与现场硬件测算。开放模型权重和量化格式让实际测试更容易开展，对开发者的价值比一个没有可重复环境的综合分数更具体。

参考资料

1. OpenAI ｜ Advancing chip design with OpenAI models ｜ 2026-11-18 ｜用户原稿参考资料，未另行核验
   https://openai.com/index/advancing-chip-design-with-our-models/

2. Google Labs ｜ Introducing Stitch Loop: A new way to build and evolve products ｜ 2026-11-18 ｜用户原稿参考资料，未另行核验
   https://blog.google/innovation-and-ai/models-and-research/google-labs/stitch-loop/

3. Google DeepMind ｜ Genie 4 is now available ｜ 2026-11-18 ｜用户原稿参考资料，未另行核验
   https://blog.google/innovation-and-ai/models-and-research/google-deepmind/genie-4-now-available-project-genie/

4. NVIDIA ｜ NVIDIA RTX Spark Compact Desktops Out Now ｜ 2026-11-18 ｜用户原稿参考资料，未另行核验

https://blogs.nvidia.com/blog/rtx-spark-desktop-local-ai-available-now/

5. SenseTime/Hugging Face | SenseNova-Loop-9B | 2026-11-18 | 权重更新

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

机构介绍与关注

新质生产力工作委员会

中国高技术产业发展促进会新质生产力工作委员会，专注于推动工业人工智能、智能制造、数字化转型等前沿技术发展，为企业提供政策解读、技术咨询和产业对接服务。

工业智能算网

专注于工业人工智能、新质生产力、工业软件 CAE、智能制造等前沿技术。提供每日动态分析、技术趋势解读、解决方案分享，推动工业智能化转型。

关注高促会新质生产力工委会公众号

关注工业智能算网平台

https://gyznsw.cn