

AI 技术每日分析：企业智能体进入生产系统，训练与评价体系同步升级

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 7 月 18 日

摘要

企业 AI 应用的重心继续从“接入一个模型”转向“让智能体稳定使用真实系统”。Smartsheet 披露其远程 MCP 服务器的生产架构，通过权限继承、工具审计、分层限流和面向模型的紧凑数据格式，把智能体接入项目管理系统；公司内部统计显示，接口优化累计节省超过 30 亿 Token。NVIDIA 与 Hugging Face 则把 NeMo Automodel 接入 Diffusers 生态，使图像和视频扩散模型能够在不转换检查点、不重写模型的情况下，从单机 LoRA 扩展到多机全量微调。语音 AI 的应用范围也在扩大，保加利亚社会企业 Stray Sheep 使用生成式语音制作手语课程和无障碍视频。OpenAI 同日提出以“每美元获得的有用智能”衡量企业 AI，评价重点由调用量转向成功完成任务、总成本和可依赖程度。

Contents

一、Smartsheet 把远程 MCP 推向生产：工具接口开始为智能体重新设计	2
二、NVIDIA 与 Hugging Face 打通扩散模型微调：一套训练框架覆盖单卡到集群	2

三、Stray Sheep 用生成式语音补足手语内容：小团队也能制作多层无障碍媒体	3
四、OpenAI 提出“每美元有用智能”：企业 AI 评价从 Token 价格转向任务结果	4
参考文献	4

一、Smartsheet 把远程 MCP 推向生产：工具接口开始为智能体重新设计

AWS 于 7 月 17 日披露，企业工作管理平台 Smartsheet 已在 AWS 上构建远程 MCP 服务器，使 Amazon Q、Claude Desktop 智能助手与外部 AI 客户端共用同一层 MCP 工具和权限体系，减少了为不同模型重复开发接口的工作。

这套系统的重点不是简单包装现有 API，而是针对智能体调用方式重新组织返回数据。Smartsheet 采用渐进式披露控制单次返回量，用严格 JSON Schema 约束参数，Token 数量可减少 35% 至 47%，累计节省超过 30 亿 Token。MCP 服务上线后的前四周，用户数量周环比增幅超过 87%。

智能体流量具有短时间连续调用、随后等待模型推理的突发特征。Smartsheet 因此使用 Fargate 自动扩缩容、每 15 分钟执行端到端金丝雀测试，并按用户而不是共享 IP 实施限流。管理员可以只开放只读工具，也可允许写入或破坏性操作；每次调用记录用户、组织、工具和结果，为审计和故障追踪保留链路。企业部署 MCP 的难点由“能否连接”转向“连接后是否可控、可测、可计费”。

二、NVIDIA 与 Hugging Face 打通扩散模型微调：一套训练框架覆盖单卡到集群

NVIDIA 与 Hugging Face 7 月 17 日发布 NeMo Automodel 和 Diffusers 的集成方案。开发者可以直接微调 Diffusers 格式模型，不需要把权重转换为另一套训练格式，也不需要为新模型重写完整训练脚本；微调后的检查点可以直接回到 Diffusers 推理管线或上传 Hub。该集成以 Apache 2.0 许可证开放。

首批配方覆盖 FLUX.1-dev、Qwen-Image、Wan 2.1、Wan 2.2 和 HunyuanVideo 等图像与视频模型，训练规模可以通过配置切换 FSDP2、张量并行、上下文并行和流水线并行，而不是修改模型代码。官方测试显示，Wan 2.1 1.3B 视频模型可在单张 40GB A100 上训练，更大的模型则可扩展到多 GPU 节点。

这项工作解决的是开源多模态模型的工程碎片化问题。模型发布速度很快，但企业往往缺少可复用的数据预处理、缓存、检查点、并行和部署流程。统一训练接口将缩短从基础模型到品牌视觉、行业视频和专用生成模型的适配周期，也让训练配方更容易进入版本管理和重复验证。

三、Stray Sheep 用生成式语音补足手语内容：小团队也能制作多层无障碍媒体

ElevenLabs 7 月 17 日介绍了保加利亚社会企业 Stray Sheep 的无障碍内容项目。保加利亚约有 12 万名听力受损人士，而该机构创始团队估计活跃手语翻译仅有 10 至 15 人。Stray Sheep 已向超过 1500 人教授保加利亚手语，并为公共活动和企业客户提供手语服务。

其视频通常把手语翻译、字幕和旁白叠加在一起，使听障观众、阅读字幕的观众和听觉观众能够观看同一份内容。团队使用 ElevenLabs 文本

转语音生成课程、公益宣传、纪录片和培训材料的旁白，并从同一脚本生成其他语言版本。生成式语音降低了传统录音所需的演播、人员协调和多语种制作成本。

这类案例的重要性不在于模型参数，而在于 AI 能否降低小语种和小规模公益内容的生产门槛。无障碍产品仍需由听障群体、手语教师和专业翻译参与审核，语音合成只是内容链条中的一环，不能替代手语表达和文化语境。

四、OpenAI 提出“每美元有用智能”：企业 AI 评价从 Token 价格转向任务结果

OpenAI 7 月 17 日提出 “Useful Intelligence per Dollar” 评价框架，建议企业围绕四个问题评估 AI 系统：一项任务能否成功完成，完成任务的完整成本是多少，结果有多可靠，以及规模扩大后每一美元能否完成更多工作。其观点是，单纯比较每 Token 价格容易忽略重试、人工复核、延迟和返工；更强但单价更高的模型，可能因为成功率更高、返工更少而带来更低的单位任务成本。

在可靠性方面，框架建议区分“可直接使用”“需要修改”和“需要人工接管”三类结果，并在智能体获得写入和执行权限前明确数据访问范围、可修改系统和人工审批节点。

这套口径来自模型供应商，企业仍需独立核算，但它反映了采购逻辑的变化。模型排行榜和调用量不能直接说明经营价值；客服问题解决率、代码通过测试比例、合同审查准确率和每个成功任务的全成本，正在成为更有效的评价指标。

参考文献

- AWS: 《How Smartsheet built a remote MCP server on AWS》, 2026-07-17; 用途: 核验远程 MCP 架构、Token 节省、权限、限流与上线数据。
<https://aws.amazon.com/blogs/machine-learning/how-smartsheet-built-a-remote-mcp-server-on-aws/>
- Smartsheet Developers: 《Smartsheet MCP Server》, 持续更新; 用途: 补充 MCP 连接、工具目录和开发者使用方式。
<https://developers.smartsheet.com/>
- AWS Architecture Center: 《Guidance for Deploying Model Context Protocol Servers on AWS》, 持续更新; 用途: 补充分层防护、容器部署和网关设计。
<https://aws.amazon.com/solutions/guidance/deploying-model-context-protocol-servers-on-aws/>
- NVIDIA / Hugging Face: 《Fine-tune video and image models at scale with NVIDIA NeMo Automodel and Diffusers》, 2026-07-17; 用途: 核验集成方式、支持模型、许可证和训练性能。
<https://huggingface.co/blog/nvidia/scale-diffusers-finetuning-nemo-automodel>
- NVIDIA GitHub: 《NeMo Automodel》, 持续更新; 用途: 补充开源代码、训练配方和安装方式。
<https://github.com/NVIDIA-NeMo/Automodel>
- NVIDIA Documentation: 《Diffusion Fine-Tuning Guide》, 持续更新; 用途: 补充图像与视频模型微调配置。
<https://docs.nvidia.com/nemo/automodel/latest/guides/diffusion/>
- ElevenLabs: 《How Stray Sheep uses ElevenLabs to make Bulgaria more

accessible for Deaf people》，2026-07-17；用途：核验听障人口、手语教育规模和语音制作流程。

<https://elevenlabs.io/blog/how-stray-sheep-improves-accessibility-for-deaf-people>

- Stray Sheep：机构官网，持续更新；用途：补充保加利亚手语课程和无障碍服务背景。

<https://straysheep.bg/>

- OpenAI：《A scorecard for the AI age》，2026-07-17；用途：核验“每美元有用智能”、成功任务成本和可靠性指标。

<https://openai.com/index/a-scorecard-for-the-ai-age/>

- Hugging Face / Dharma AI：《Newer Models, Same Advantage》，2026-07-17；用途：补充专用小模型与通用模型采购逻辑的社区观察。

<https://huggingface.co/blog/Dharma-AI/newer-models-same-advantages>

联系我们，请扫描二维码



新质生产力工作委员会
官方公众号



工业智能算网
gyznsw.cn

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。

网站地址：<https://gyznsw.cn>