AI 技术每日分析：Perplexity 发布 Decider v1.3 与 Flash v1.3，新增 Locate 内容抽取

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 14 日

摘要

11 月 13 日可核实的 AI 新进展集中在模型如何成为稳定的工程组件。Perplexity 在 11 月 12 日推出 Decider v1.3 和 Flash v1.3，为判断、路由、结构化字段提取提供专门模型；Salesforce 发表 SOV-Transcribe-1，在开放权重语音识别上给出一组性能数据；IBM 用 LLMON 语言和 Granite 4.1 预览模型组织指令、数据与工具结果；ggml-org 于 11 月 13 日发布 llama.cpp v1.0.0，明确公开接口的版本契约。这几条消息各自对应企业系统里的不同环节：判断与抽取、语音转写、上下文标记和本地推理接口。

Contents

一、Perplexity 更新 Decider：让判断和信息抽取脱离逐字生成 **2**
二、Salesforce 发布 SOV-Transcribe-1：七种语言的可自托管语音识别 **2**
三、IBM 推出 LLMON：让模型理解哪部分是指令、哪部分是数据 **3**
四、llama.cpp 跨入 1.0：本地推理组件开始建立版本契约 **3**
五、Anthropic 自托管 Managed Agents：智能体运行环境成为可部署产品 **4**
趋势观察 **4**
参考资料 **5**

一、Perplexity 更新 Decider：让判断和信息抽取脱离逐字生成

Perplexity 在 11 月 12 日发布 Decider v1.3 与更轻量的 Decider Flash v1.3。其任务定义很明确：接收文本、结构化内容或图像，针对命名问题返回标签的概率，用于分类、排序、路由及评测。新版本增添名为 Locate 的内容抽取能力，可按字段要求定位原文片段，返回字符区间和置信分数。这使其从“判断这个工单属于哪一类”，延伸到“抽出设备编号、订单金额和合同条款所在位置”。

官方资料显示，大版本基于 Qwen3.8-27B，Flash 建立在 Qwen3.5-4B 之上并通过蒸馏获得，BF16 权重规模分别约 49GiB 与 8.5GiB。企业由此可以按任务难度配置推理资源：高吞吐队列前置轻量分类器，复杂判断转入大模型，抽出的原文位置再交由规则与业务系统处理。Perplexity 介绍其模型通过 API、OpenRouter 及 Hugging Face 渠道提供；Hugging Face 模型页的访问状态需要按实际账户权限确认。工单分流和字段定位这样的任务，需要的是可检查的标签与原文位置，而不只是生成一段解释。

二、Salesforce 发布 SOV-Transcribe-1：七种语言的可自托管语音识别

Salesforce 主权 AI 团队 11 月 12 日在 Hugging Face 发布 SOV-Transcribe-1，采用 Apache 2.0 许可开放权重。模型卡给出的规模为总参数 26.6B、激活约 4.6B，结合 Whisper large-v3 音频编码器、Gemma 4 混合专家语言模型以及音频投影模块，面向非流式转写。其语言范围包括英语、法语、德语、西班牙语、意大利语、葡萄牙语与荷兰语，提供自然语言转写格式指令，以及专业名词、企业名称与产品术语的关键词提示。

Salesforce 公布，在 Artificial Analysis 评测中模型词错误率为 2.28%，

在其引用的非流式开放权重榜单中排第一；在 14 种英语口音的 AppTek 呼叫中心语料中平均词错误率为 8.21%。另一项针对姓名、电话、邮件与代码符号的 VoiceCodeBench 评测报告，完整实体无误的录音比例为 76.3%，关键实体召回为 93.4%。上述数字属于企业发布时的公开评测口径。结合企业工单、客服质检、售后音频归档，语音识别正在从“把声音变成文字”进入“高价值字段和专业术语可直接进入业务流程”的阶段。

三、IBM 推出 LLMON：让模型理解哪部分是指令、哪部分是数据

IBM Granite 团队 11 月 12 日发布 LLMON（LLM Object Notation），同时放出基于 Granite 4.1 的 3B 和 8B 预览模型及配套工具。LLMON 以明确的语义标签、保留标记与唯一 ID 组织上下文，把系统指令、任务输入、工具调用结果、数据对象分别标出，再以执行块描述需要调用的指令及对应输入。它为复杂智能体工作流增加了更像程序语言的结构层，而不是仅靠自然语言表述让模型自行推断每段内容的角色。

IBM 将这一设计用于控制工具结果与检索文本中的间接指令注入。其公布的 OPI 实验显示，相较所测试对照模型，在相关攻击样本上的抵抗表现有显著改善；同时提供训练方式和开源工具以便研究者复核。对于企业部署，直接价值在于把多轮上下文、工具回传与结构化数据纳入统一协议。随着模型与工具交换的数据越来越多，字段语义、来源和处理范围会成为影响系统维护成本的重要变量。

四、llama.cpp 跨入 1.0：本地推理组件开始建立版本契约

ggml-org 于 11 月 13 日公开 llama.cpp v1.0.0。过去项目频繁引入新模型结构、量化格式和多模态能力，使用 C/C++ API 的应用需要追随内部接口变化。这次版本升级更有工程意义的是明确 MAJOR、MINOR

和 PATCH 的兼容规则，配合对公共 API 的整理，让嵌入式、本地部署与企业私有推理产品可以基于明确版本组织升级。

新版本同时增加语音接口与若干模型适配，提供 OpenAI 风格的‘/v1/audio/speech‘端点，更新多模态基础设施与 Web 交互界面。工业本地化 AI 系统往往使用固定硬件、稳定的应用进程及严格的发布周期，推理底座的兼容性直接影响维护工作量。模型可以季度更新，底层运行库则需要形成可预测的版本节奏。把推理引擎当作长期基础设施管理，正在成为企业级 AI 技术栈的重要部分。

五、Anthropic 自托管 Managed Agents：智能体运行环境成为可部署产品

Anthropic 在 11 月 12 日的产品更新中介绍 Managed Agents 的自托管部署模式，允许企业在自有云或本地基础设施运行智能体，并保管凭据与会话历史。该变化面向有现成安全边界、审计系统和企业运维体系的组织。将智能体运行环境从单一云服务接口延伸到用户控制的基础设施，便于与既有身份系统、业务系统接口、任务记录与文件权限结合。

趋势观察

综合本期五项消息，AI 系统建设的能力分层已相当清晰：语音与视觉模型处理输入，专门的决策模型承担海量结构化判断，LLMON 一类协议描述任务和数据关系，本地推理引擎承担运行，智能体执行环境管理长期流程与状态。这些能力可以分别接入现有软件；实际采用时，还要逐项核对模型权限、接口兼容性和运行条件。

参考资料

1. Perplexity | Decider v1.3，2026-11-12 | 2026-11-12 | 编辑稿所列参考来源（未独立核验）

   https://www.perplexity.ai/hub/blog/perplexity-decider

2. Salesforce | SOV-Transcribe-1，2026-11-12 | 2026-11-12 | 编辑稿所列参考来源（未独立核验）

   https://huggingface.co/blog/salesforce-sovereign-intelligence/sov-transcribe-1

3. IBM Granite | LLMON，2026-11-12 | 2026-11-12 | 编辑稿所列参考来源（未独立核验）

   https://huggingface.co/blog/ibm-granite/llmon-an-llm-native-markup-language-now-in-granite

4. ggml-org | llama.cpp v1.0.0，2026-11-13 | 2026-11-13 | 编辑稿所列参考来源（未独立核验）

   https://huggingface.co/blog/ggml-org/llama-cpp-v1

5. Anthropic | Managed Agents 产品更新，2026-11-12 | 2026-11-12 | 编辑稿所列参考来源（未独立核验）

   https://claude.com/resources/product-updates

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