

ElevenLabs 推出实时语音模型 Mini 和 Max

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 10 月 21 日

摘要

国际 AI 产品有两种走向：实时语音、搜索和业务动作被打包成完整的智能体服务；特定任务则被压缩进可在个人设备上运行的小模型。10 月 20 日，ElevenLabs 正式推出 Eleven Realtime v1 Mini 和 v1 Max 语音交互模型；独立团队 XHToken 公开 4B 参数的 Spark-X2.6 权重；小型开发团队 Labqoat 发布面向分类与决策的 qev-0.8b；OpenAI 披露收购药物研发流程智能体公司 Orchestra Bio；英伟达研究团队公开基于真实手术记录训练的交互式生成模拟器 Cosmos3-H-Dreams。这五项进展涉及实时交互、端侧成本、轻量决策、科研工作流和专业模拟。厂商的性能宣称仍需放到实际工程中检验。

Contents

一、ElevenLabs 发布两款实时语音模型：交互系统由组件拼接走向整体交付	2
二、XHToken 开放 Spark-X2.6-4B：端侧智能体继续强化工具调用与可控推理	2
三、Labqoat 发布 qev-0.8b：不生成文本的“小决策模型”找到细分用途	3
四、OpenAI 收购 Orchestra Bio：生命科学智能体从问答延伸至项目执行	3
五、英伟达公开手术交互生成模拟器：视觉逼真不是物理验证的终点	4

趋势观察	4
参考资料	4

一、ElevenLabs 发布两款实时语音模型：交互系统由组件拼接走向整体交付

ElevenLabs 于 10 月 20 日发布 Eleven Realtime v1 Mini 和 v1 Max, 均将语音识别、轮次判断、对话推理和语音输出整合为实时语音交互能力, 支持 90 多种语言, 并通过 ElevenAgents 及兼容 OpenAI Realtime 规范的 WebSocket 接口提供服务。Mini 主要面向成本敏感的陪伴、教育和一般助手, Max 针对客服、线索筛选和多步骤工具调用。官方引用的评测中, Max 在 Big Bench Audio 取得 98.8% 的正确率、平均首段语音等待约 0.57 秒; 在工具任务上, Max 和 Mini 的完成率分别为 87.1% 和 60.3%。这些数字来自特定测试集和配置, 不能直接套用到中文通话场景。[1][2]

语音智能体难在处理用户打断、长停顿、混杂口音, 以及调用外部系统后如何接续对话。整体模型省去了自行拼接识别、合成模块的部分工作。电话噪声、方言、敏感操作确认和端到端费用, 仍要逐项测试。对于实际客服, 还应保留人工接管、录音合规和工具权限审计。

二、XHToken 开放 Spark-X2.6-4B：端侧智能体继续强化工具调用与可控推理

开源团队 XHToken 公开 Spark-X2.6-4B 模型权重, 仓库标注 Apache 2.0 许可, 并提供 Transformers、vLLM、SGLang 及量化部署路径。与前代 X2.5 相比, 新版本保持约 40 亿参数规模, 主要通过后训练改善代码、工具使用、多步骤任务和长上下文表现, 并引入“关闭、低、高”三级思考预算, 供应用按任务难度调节推理开销。[3]

企业本地工具助手要频繁处理表单、目录检索和脚本修订, 未必每一步都要调用成本更高的远程大模型。分层设置推理预算或许能缩短响应时间; 标称“支持长上下文”, 却不能保证最大窗口下的召回率依然稳定。

官方榜单及与更大参数模型的比较尚需第三方使用相同硬件、量化、工具集进行复测，且仓库的自定义代码运行权限应先做安全审查。

三、Labqoat 发布 qev-0.8b：不生成文本的“小决策模型”找到细分用途

独立团队 Labqoat 于 10 月 20 日公布 qev-0.8b，以约 8 亿参数专门处理多选、是否、评分等决策问题，输出选项概率而不是长段自然语言。项目公开 CUDA、Apple MLX 及 GGUF 权重路径，标注 Apache 2.0 许可；开发者称其 8 位量化版本在苹果设备约占用 1.3GB 内存。官方自设的 Benchmark Heaven 测试得分为 77.1%，并报告在 L40S 上单次决策延迟中位数约 3.5 毫秒。数据集样本规模与执行环境由团队自行披露，不能当作行业独立认证。[4][5]

工单分流、邮件优先级、规则筛选和智能体动作候选排序，都可能用到这类模型。它处理的是大量重复判断，用低成本的本地概率估计来完成，并非替代复杂推理模型。概率输出还需校准与拒答阈值：置信度不足时应交给人工或更强模型，尤其不能直接批准支付、删除或权限变更。

四、OpenAI 收购 Orchestra Bio：生命科学智能体从问答延伸至项目执行

Axios 于 10 月 20 日披露，OpenAI 已收购专注药物研发流程管理的初创公司 Orchestra Bio，计划将其团队和技术纳入面向生命科学的 Rosalind 产品体系。Orchestra 开发的智能体用于跟踪研究进度、临床与研发节点，以及提醒可能导致任务延误的事项；交易金额未公开。[6]

这起收购涉及科研流程管理：把实验结果、研究任务和下一步决策关联起来。它没有宣称 AI 已发现某种新药。科研团队要处理实验记录、项目里程碑和监管文档。智能体若参与协调，原始数据出处、责任人确认及

版本变更都要留下记录。收购完成也不代表该工具已在所有药企正式部署，更不能由并购行为推断临床成功率提升。

五、英伟达公开手术交互生成模拟器：视觉逼真不是物理验证的终点

英伟达研究人员 10 月 20 日介绍 Cosmos3-H-Dreams：利用手术视频与机器人运动学记录训练动作条件生成模型，尝试在控制器发出平移、旋转和夹持动作时，实时生成相应内窥镜场景。项目从此前桌面缝合任务扩展至临床胆囊切除场景，采用 Cosmos 3 与 FlashDreams 加速推理，团队称可在单张 RTX PRO 6000 上交互运行，并提供代码及研究说明。[7][8]

交互模拟比普通视频合成多一层要求：连续动作、器械位置和组织变化必须一致。研究者也明确指出，视觉相似并不意味着生物力学准确，误差可能随时间累积。现阶段，它更适合机器人研发和研究性测试，不能当作已获临床认证的训练或手术决策系统。后续需依靠真实数据留出验证、物理约束和专家盲评，才能判断迁移价值。

趋势观察

实时语音平台整合了原本分散的组件，小模型承担端侧决策；生命科学产品开始跟踪研究任务，交互模拟则需要同时经受动作和场景的检验。接下来要看的，是这些系统在具体权限、成本和验证条件下能否可靠完成任务。

参考资料

1. ElevenLabs 官方博客 | 《Introducing Eleven Realtime: a simpler way to build natural voice agents》 | 2026-10-20 | 用途：核验 Mini/Max 功能、可用渠道、语种、厂商测试条件及性能声明。

<https://elevenlabs.io/blog/realtime>

2. The Decoder | 《Elevenlabs releases two new speech-to-speech models designed for real-time AI conversations》 | 2026-10-20 | 用途：交叉核验两款模型定位、接口与语音评测数据。

<https://the-decoder.com/elevenlabs-releases-two-new-speech-to-speech-models-designed-for-real-time-ai-conversations/>

3. XHToken / Hugging Face | 《Spark-X2.6-4B: Advancing the Frontier of On-Device Agents》 | 2026-10-20 发布动态 | 用途：核验开放权重、Apache 2.0 许可、4B 规模和分级推理预算。

<https://huggingface.co/XHToken/Spark-X2.6-4B>

4. Labqoat 项目原站 | 《qev-0.8b: The tiny local decision model》 | 2026-10-20 | 用途：核验小模型的决策输出形式、硬件需求、原始测试规模与厂商数字。

<https://app.labqoat.com/e/qev/qev>

5. Labqoat / Hugging Face | 《qev-0.8b Model Weights》 | 2026-10-20 相关发布 | 用途：核对权重、量化版本和可复现部署入口。

<https://huggingface.co/labqoat/qev-0.8b>

6. Axios | 《Exclusive: OpenAI acquires biotech startup as AI race moves into drug development》 | 2026-10-20 | 用途：核验 Orchestra Bio 收购、Rosalind 整合和未披露交易价格。

<https://www.axios.com/2026/10/20/exclusive-openai-biotech-startup-drug-development>

7. NVIDIA 研究团队 / Hugging Face | 《NVIDIA Cosmos3-H-Dreams:

Bringing Clinical Surgery to Life Through Generative Simulation》 |
2026-10-20 | 用途：核验模拟器动作条件、训练数据、推理硬件与技术限制。

<https://huggingface.co/blog/nvidia/cosmos3-h-dreams>

8. NVIDIA / GitHub | 《Cosmos3-H-Dreams Research Repository》 |
2026-10-20 研究配套 | 用途：核对代码入口、研究实现和后续独立复现条件。

<https://github.com/isaac-for-healthcare/cosmos3-h-dreams>

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>