精选快讯　软件快讯　2026年11月12日 10:24

Google推出Gemini Decisions API，单次最多评估1000个问题

要点

- Google推出由Gemini Decision 1模型驱动的Gemini Decisions API，面向企业和开发者的决策任务。
- 与逐token生成回答不同，Gemini Decision 1在单次处理步骤中为可用选项打分，并返回概率和置信度分数。
- Google称，该API端到端响应时间为150至250毫秒，定价为每百万输入token 0.10美元，输出token不收费；
- Google称，其团队已在Gemini Notebook中用该技术将查询路由至合适的AI模型，在Stitch中用于加快产品设计，在Google Search中将其作为检测对抗性内容的安全层。

Google推出由Gemini Decision 1模型驱动的Gemini Decisions API，面向企业和开发者的决策任务。该API单次请求最多可评估1000个问题，覆盖文本、图像、音频、视频和PDF文档；模型最多处理256000个token的多模态输入，用户可在单次请求中组合不同格式。其适用任务包括工具选择、政策检查、将查询路由至AI模型和信息评估。

与逐token生成回答不同，Gemini Decision 1在单次处理步骤中为可用选项打分，并返回概率和置信度分数。API支持三类决策：是或否评估、多项选择，以及按有序标准集评分。每个答案包含置信度分数，使开发者可将不确定案例转交备用系统或人工审核。

Google称，该API端到端响应时间为150至250毫秒，定价为每百万输入token 0.10美元，输出token不收费；无论一次请求评估多少问题，共享输入只计费一次。Google还称，Gemini Decision 1在推理、排序、智能体路由和多模态理解等基准测试中表现强劲，在对比模型中校准误差最低，并在共享基准套件上比使用Interactions API的Gemini 3.5 Flash-Lite便宜近五倍。

Google称，其团队已在Gemini Notebook中用该技术将查询路由至合适的AI模型，在Stitch中用于加快产品设计，在Google Search中将其作为检测对抗性内容的安全层。

报道来源：indiatimes

来源网站：https://www.inews.today/