Fal、Fireworks AI这类提供AI模型与服务器调用服务的初创企业营收持续走高。开发者借助它们快速运行模型，也引发投资者新一轮的投资热潮。

据两名知情人士透露，专注为谷歌Nano Banana等图像、视频生成模型提供推理服务的Fal，已和投资者洽谈新一轮融资，目标估值150亿美元。另有第三位知情人士称，谈判尚处于早期阶段，该公司有可能把目标估值上调至170亿至200亿美元区间。

无论最终估值落在哪个区间，本轮融资都将让这家成立五年的企业估值，较今年春季一轮融资的80亿美元近乎翻倍。一名知情人士表示，洽谈融资的背景是其年化营收攀升至8亿美元，相比3月水平翻了一番。

第三位知情人士称，按营收规模排名第一的推理服务商Fireworks AI也在考虑新一轮融资，若推进融资，目标估值可达300亿美元，相比7月公布的上一轮估值几乎翻倍。

Fireworks周五表示，公司已完成由阿特莱德斯资本牵头的员工股份出售，交易估值175亿美元，与7月公布的上一轮融资估值持平。截至7月，其年化营收已经达到10亿美元，较一年前增长四倍。目前暂无法获知其最新营收数据。

据报道，另一家推理服务商Baseten也在和投资者洽谈新一轮融资，目标估值260亿美元；同类服务商Modal正在洽谈融资，估值约150亿美元，约为四个月前一轮融资估值的三倍。

密集的融资洽谈，反映出这类提供前沿模型调用、协助客户定制开源模型的初创企业正处在行业红利期。

得益于开源AI技术的进步，开发者大量使用这类服务，开源模型成本往往低于OpenAI和Anthropic的闭源模型。自行搭建硬件来运行开源模型配置难度大，且芯片短缺推高成本，因此很多开发者倾向于通过简易应用程序接口调用模型，Fireworks与Fal都提供这类接口。

不过，随着OpenAI、Anthropic这类闭源模型厂商加大降价力度，推理服务商或将面临压力。代码初创企业Replit首席执行官阿姆贾德·马萨德周三表示，由于部分OpenAI模型价格大幅下调，公司近期使用开源AI的量相比年初有所减少。

此外，推理服务商向云厂商租赁算力的成本上涨，企业利润率存在被压缩的风险。这类推理公司过往毛利率大约在50%，低于顶尖软件服务商普遍70%以上的毛利率水平。

推理服务商认为，可以通过优化模型在芯片上的运行方式提升毛利率。同时提供推理服务与AI服务器租赁业务的Together
AI也在采购芯片服务器，通过自有数据中心对外出租。