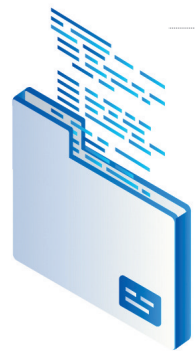


涨价开源换打法 大模型一周三更

8月第二周,中国大模型行业密集变阵。8月13日,DeepSeek确认DeepSeek-V4-Flash和DeepSeek-V4-Pro将自8月17日起上调价格,实行峰谷定价;同日,DeepSeek-V4-Pro正式版上线,并发布Harness(智能体运行框架)开发者预览版。一天后,智谱发布GLM-5.3,该版本与GLM-5.2采用相同基座模型,仅靠后训练便将编程能力提升50%;同日,阿里千问正式开源Qwen3.8系列模型,其中全新开源的270亿参数模型Qwen3.8-27B,家用显卡也能跑。

从“谁更便宜”转向“谁更不可替代”。DeepSeek用Harness对冲涨价可能带来的用户流失风险,智谱押注后训练撬动性能杠杆,阿里云继续以开源模型拉动云生态。绕不开算力等硬性成本时,路线就成了战场。AI模型聚合平台OpenRouter最新数据印证了大模型客户的选择:上周全球模型调用量排名前十的模型中,6个为国产模型。



8月第二周大模型行业核心动态

企业	时间	事件
DeepSeek	8月13日	V4-Pro正式版上线;Harness开发者预览版发布
DeepSeek	8月17日生效	API调价,实行峰谷定价
智谱	8月14日	GLM-5.3发布
阿里千问	8月14日	Qwen3.8-27B开源
阿里云	8月以来	Quick BI、Wan3.0、CosyVoice、Meoo、M890等密集更新
金山办公	8月13日—14日	灵犀专业版连续接入新模型

DeepSeek双线落子

蓝色鲸鱼、黑色鲸鱼——DeepSeek和DeepSeek Harness的官方微信公众号头像以颜色相区分,并在同一天发布内容,这对于数月才更新一次的DeepSeek来说实属罕见。

DeepSeek官方账号最新文章的标题是《DeepSeek-V4-Pro正式版上线》,“DeepSeek Harness团队”发布的是《DeepSeek Harness开发者预览版上线》。模型由预览版转为正式版是DeepSeek预告过的动作,上线Harness也有蛛丝马迹,但这两篇博文的阅读量依然迅速达到10万+。

所谓Harness,在当下的AI语境下指的是一套智能体运行框架。过去人类与AI的交互是线性的:你提问,它回答。现在进入了Agent(智能体)时代。Agent要能自主拆解任务、调用工具,甚至自我修正。在这个过程中,模型本身的“智商”固然重要,但运行的“环境和约束”决定了它是否可靠。Harness就是从对话到系统的跃迁。

Nextie(明日新程)创始人、“小冰之父”李笛向北京商报记者解释:“三年前,人们往往认为只要模型能力足够强,就自然可以得到完美的效果。但Claude Code的模型能力足够强,依然需要数十万行代码的Harness,才能让其在具体应用中发挥端到端的表现。此次DeepSeek的Harness,也是类似的具体行动。”

回到具体产品,DeepSeek Harness的设计思路是“一切皆插件”,采用插件式开放架构构建Agent Harness,即模型、工具、技能、会话、沙箱、存储、循环、调度、UI(用户界面)等所有Agent能力均由插件组合而成,可自由替换、灵活重组。DeepSeek Harness方面进一步介绍,开发者无需改动DeepSeek Harness的源码本身,就能以插件方式独立选择、替换或扩展其中任一能力。

按照这个逻辑,DeepSeek的模型只是DeepSeek Harness的插件之一。开发者不仅可以接入DeepSeek的模型,理论上也可对接市面上其他主流大模型。北京商报记者就此以用户身份向DeepSeek Harness团队留言,但截至发稿暂未得到回复。

涨价与开源的同题问答

如果将模型涨价和发布Harness拼在一起,业内人士从中看到了商业化焦虑。

8月17日,DeepSeek执行新的收费标准,实行峰谷定价。高峰时段为北京时间9时至12时、14时至18时(其余为空闲时段),高峰价格为空闲时段的两倍。

DeepSeek-V4-Flash空闲时段,百万Tokens输入(缓存命中)价格为0.05元,百万Tokens输入(缓存未命中)价格为1.5元,百万Tokens输出价格为4.5元;高峰时段上述价格均翻倍,分别为0.1元、3元、9元。

DeepSeek-V4-Pro空闲时段,百万Tokens输入(缓存命中)价格为0.15元,百万Tokens输入(缓存未命中)价格为4.5元,百万Tokens输出价格为13.5元;高峰时段上述价格均翻倍,分别为0.3元、9元、27元。

“基础大模型涨价是大模型从跑马圈地到进入商业化深水区的大趋势,也逐渐成为行业共识,”李笛告诉北京商报记者,“模型有商业价值不等于有合理的商业模式,涨价是



模型厂商不断感知社会接受度、调整商业模式,最终达到供需平衡的过程。”

早期模型厂商的价格战甚至免费体系,一再拉低用户对大模型价格的心理预期。“在这种情况下,模型厂商单纯调价,往往在短期内表现为阵痛,造成用户流失。”李笛认为,“模型企业往往不能只调价,还必须同时提供比过去更深入的服务,Harness就是提质手段之一。大模型时代的今天,我们使用的是最新技术,但商业模式却非常陈旧——过去30年以边际成本递减为基本前提奠定的商业模式,并不适用于大模型时代。”站在整个行业角度,即使有些厂商没有涨价,也在回答同一个问题:模型之外,还能靠什么挣钱?

8月以来,阿里云从算力、模型到产品均有新动态:瓴羊企业级智能分析产品Quick BI完成AI-native架构升级,发布全新AI Pro版本;视频生成模型Wan3.0启动公测;上线一站式AI语音生产平台CosyVoice Studio;AI应用创作平台Meoo(秒悟)团队版正式全量上线;灵骏真武M890超节点实例正式上线,首批在乌兰察布地域开售;正式开源Qwen3.8系列模型。

智谱的“后训练”杠杆

“圈地”的方式不止开源和涨价两种。智谱选择了一条更省成本的路:不换基座,聚焦后训练。

8月14日,智谱发布GLM-5.3。与GLM-5.2相比,基座模型没变,通过后训练Scaling(规模化)提高了模型的智能上限。编程方面,GLM-5.3在内部自建基准评测Z.ai Code Bench中较GLM-5.2提升50%,在Terminal Bench 3.0、Agents’ Last Exam(CLI)等公开基准测试中取得开源模型第一;网络安全方面,在白盒代码审查与漏洞发现等安全任务中,GLM-5.3表现与Mythos 5相当,展现出面向网络安全防御场景的潜力。

据智谱介绍,上述提升全部来自后训练。基于IndexShare、SAO以及持续演进的新一代slime框架,智谱在与GLM-5.2完全相同的基座上高效推进强化学习。“可能我们还远未开发出这个基座的智能上限。”智谱在公告中强调。

效果如何?当天,金山办公旗下AI办公智能体灵犀专业版即首发接入智谱GLM-5.3模型。在介绍此次合作时,金山办公提到了GLM-5.3的代码和网络安全能力。值得注意的是,灵犀在前一天已首批接入DeepSeek-V4-Pro正式版,同样提到了代码能力。

这不是巧合。Agent时代,能对话只是门槛,能写代码才是模型撬开客户钱包的硬指标。

文渊智库创始人王超向北京商报记者表示:“模型厂商的商业模式远没有确定,到底是靠C端(用户端)、B端(企业端)还是生态,都没确定。”在他看来,“和海外同行相比,国内大模型的算力成本更低,但推出的模型效果比同样成本的模型更好。”从某种程度看,智谱的策略正是这一判断的注脚。

OpenRouter最新数据显示,上周全球AI大模型调用量前五名中,除GPT-5.6 Luna外均为中国模型;前十名中有6个国产模型,分别是DeepSeek-V4-Flash-0731、Hy3、DeepSeek-V4-Flash-0423、GLM-5.2、MiMo-V2.5、DeepSeek-V4-Pro。

北京商报记者 魏蔚

半年报艰难盈利 香飘飘第二曲线未稳

七年来首次在半年报中实现盈利,香飘飘交出了一份备受关注的成绩单。

近日,香飘飘发布2026年半年度报告。上半年公司实现营业收入12.77亿元,同比增长23.31%;归母净利润1085.97万元,上年同期亏损9739.09万元,实现扭亏为盈。这是香飘飘自2020年以来首次半年报盈利。不过,市场对其盈利成色及“双轮驱动”战略的实际效果仍存分歧。

“双轮驱动”是香飘飘自2018年正式确立的核心经营战略。公司在2018年年度报告中首次明确提出基于“冲泡+即饮”的双轮驱动战略,并多次在公告及投资者交流中强调,冲泡业务与即饮业务作为该战略的“双轮”,战略地位均非常重要。其中,冲泡业务是公司的“基本盘”,即饮业务旨在打破冲泡产品“冬季热饮”的局限,拓展全年消费场景。

据半年报数据,公司扣非净利润仍亏损147.97万元,非经常性损益合计1233.93万元,主要由政府补助719.33万元和金融资产公允价值变动收益892.22万元构成,占净利润比重较高,盈利主要依赖非经常性损益支撑。

中国食品产业分析师朱丹蓬在接受北京商报记者采访时表示,香飘飘转型创新升级效果并不明显,“品牌老化、产品老化、渠道老化、客户老化、团队老化是其硬伤”。他认为,香飘飘核心市场在三四五线城市,而明前特级龙井轻乳茶、Meco鲜果茶等新产品主打一二线城市,品牌力明显不足。

分业务看,公司“双轮驱动”战略下的两大板块表现分化明显。报告期内,冲泡业务实现销售收入6.68亿元,同比增长

57.7%,受益于春节旺季窗口拉长及渠道修复;即饮业务实现销售收入5.79亿元,同比下降约2.1%,即饮业务占营收比例从2025年全年的约35%提升至46%。即饮业务被公司定位为“第二成长曲线”,公司在半年报中坦承,即饮业务仍处于市场投入期,仍需较多时间和资源投入,公司产品单一的状况将会在一段时间内持续存在。

冲泡业务层面,香飘飘推出两条革新路线——品质升级路线推出“原叶现泡”轻乳茶系列;功能养生路线围绕“药食同源”推出“古方五红”“古方八舒”等新品。据香飘飘统计,古方五红上市仅半年,精准购买用户便突破100万人次,复购率达37%,为品牌开辟了新的增长空间。

北京商报记者观察到,目前香飘飘天猫旗舰店中销量最高的产品为古方五红暖乳茶,其次是香咖豹手冲拿铁和低GI即饮轻乳茶,经典款杯装茶饮已不在旗舰店售卖,客服回应称“老款的经典奶茶目前店铺暂时没有库存”。

业务分化之外,区域市场的失衡同样值得关注。据半年报数据,香飘飘华东大本营增长稳健,但华南区域销售收入3811.74万元,同比下降21.64%,是所有区域中唯一

下滑的。与此同时,报告期内华南地区净减少16家门店(新增6家、减少22家)。

北京商报记者就盈利质量、“双轮驱动”战略执行效果及下半年规划等问题,通过邮件及电话向香飘飘发送采访申请,截至发稿未获回复。

从公开信息来看,香飘飘在产品布局上的调整并未停止。据香飘飘此前公告,公司于2025年12月对经营范围进行变更,新增“保健食品生产”及“保健食品(预包装)销售”两项业务。香飘飘方面当时向北京商报记者表示,此举是“常规经营范围的补充,为公司未来产品创新预留更灵活的空间”。

在即饮业务仍处投入期、冲泡业务尝试通过“药食同源”打开新空间的双重背景下,香飘飘的战略方向已逐渐清晰。然而,方向与落地之间仍有一段距离。朱丹蓬认为,加大即饮业务是香飘飘未来唯一的出路,但新茶饮行业竞争加剧,即饮业务在模式、团队管理、促销方面均面临挑战,下一步需聚焦降本增效。“现在消费者讲究品牌的调性和推广触达率,这方面香飘飘还有欠缺。未来如何让新生代接受这个品牌是关键。”

北京商报记者 董亮 实习记者 鲁芹

七鲜落地AI智能无人咖啡店

8月16日,京东集团旗下现制饮品品牌七鲜咖啡首家24小时智能化无人咖啡店落地北京银河SOHO。不同于传统咖啡店,店内所有咖啡饮品从研磨、萃取到制作、出杯全流程均由机器人执行。北京商报记者现场计时观察,具身智能机器人从取杯到放置入柜通常花费20至30秒。

从现场来看,这家“无人咖啡店”的智能化主要由前端的具身智能机器人服务和后端的全自动化生产线构成。收到订单后,店内的全自动化生产线会完成咖啡制作,并将咖啡运至出货窗口,由一台等候在此的具身智能机器人夹持并送至取餐柜中。

在咖啡制作速度上,据七鲜咖啡相关负责人介绍,单杯饮品最快约20秒即可完成制作,30秒内出杯,整条生产线最多可同时制作7至8杯饮品。开业当天,七鲜咖啡还进行了“一小时内由机器人咖啡店供应最多杯咖啡”的吉尼斯世界纪录挑战,最终以202杯挑战成功。

据了解,消费者通过自助点单系统下单后,七鲜咖啡银河SOHO店内智慧大屏将全程展示每一杯咖啡从落杯、落冰、现制到入柜的全过程。北京商报记者扫码发现,门店已提供由机器人制作的咖啡、茶饮、气泡水、益生菌果汁饮料等多种类饮品,共上架超10款单品,包括鲜奶拿铁、顶榴美式、经典美式等。目前,该店商品定价在15至20元,开通京东外卖超级月卡后优惠至6至13元价格区间,如顶榴美式原定价17.9元,月卡价为9.7元。

在标准化出品基础上,七鲜咖啡还提供极具个性化的DIY定制体验。顾客下单前可自由调节甜度、咖啡浓度、果汁浓度等,且支持1%到100%的精细调节。后续,七鲜咖啡还将根据口味偏好实现AI个性化推荐,可结合个人健康情

况精准建议当日的咖啡因摄入量,提供更加科学、健康的饮品。

七鲜咖啡相关负责人表示,目前七鲜咖啡银河SOHO店主攻到店服务,后续将陆续开通各平台的外卖服务。

就七鲜咖啡24小时智能化无人咖啡店的落地情况来看,七鲜咖啡似乎并不只是为了给消费者提供一种“机器人做咖啡”的新鲜体验,而是试图将AI深度应用到饮品研发、门店运营等多个环节。例如通过AI捕捉消费者偏好等数据为饮品研发提供参考,或24小时监测门店的运转情况等。

七鲜咖啡方面表示,京东入局咖啡茶饮赛道,是由于看到了新生产模式下的机遇:依托京东AI模型与具身智能实现咖啡口味的个性化与定制化,以七鲜咖啡线下门店为载体,赋能咖啡茶饮业务,满足用户日常生活需求。

在如今的现制咖啡市场,瑞幸、库迪等品牌通过快速扩张和低价策略占据大量市场份额,星巴克等品牌则占据着更高客单价的市场,主流茶饮品牌也在上架咖啡产品。同时,消费者对于价格、产品更新和消费体验的要求也在不断提高。

对于2025年推出的七鲜咖啡而言,落地24小时智能化无人咖啡店,一方面是提供传统咖啡店较难满足且日益增长的24小时咖啡茶饮服务,有效满足消费者从早到晚想喝就喝的需求;另一方面,依托AI技术与具身智能,实现更高的制作效率、更稳健统一的出品以及更低的门店经营成本。

上述负责人表示,未来将根据门店的实际运营情况持续迭代店型、设备和产品,并进一步探索规模化复制。

北京商报记者 何倩 毛思怡