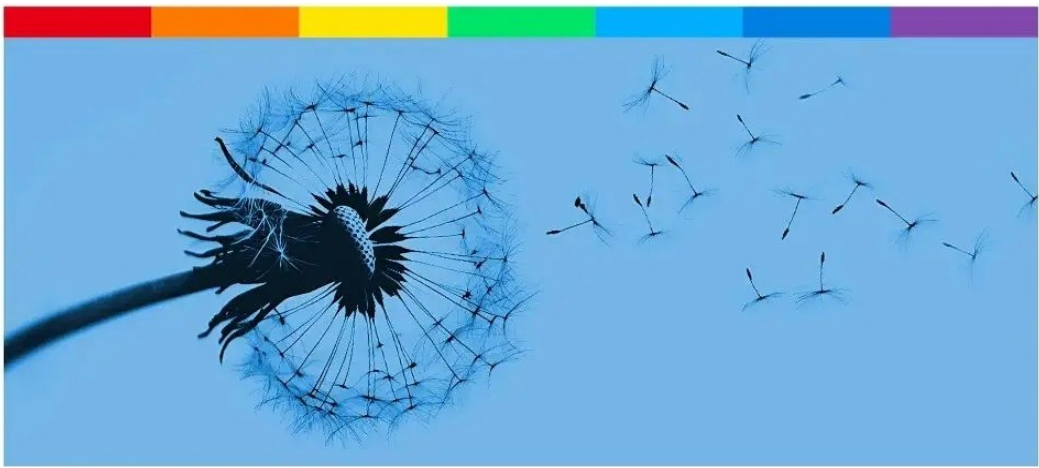


第一批 DeepSeek 开发者, 已经开始逃离了



爆火之后， 稳定成了刚需。

在 DeepSeek 服务繁忙的回复背后,不只是普通用户的焦急等待,当 API 接口响应突破临界阈值,DeepSeek 开发者的世界,也出现了持续震荡的蝴蝶效应。

1月30日,接入 DeepSeek 的 base 北京的 AI 开发者林森,突然收到程序后台报警,还没来得及为 DeepSeek 的出圈高兴几天,林森的程序便因为无法调用 DeepSeek,后台被迫瘫痪了3天。

一开始,林森以为这是因为在 DeepSeek 的账户余额不足导致的。直到2月3日春节假期过后返工,他终于接到了 DeepSeek 暂停 API 充值的通知。此时,尽管账户内余额充足,他也无法再调用 DeepSeek 了。

林森接到后台通知的第三天,DeepSeek 官方于2月6日正式对外发布公告,宣布暂停 API 服务充值。近半个月过去,截至2月19日,DeepSeek 开放平台的 API 充值服务仍未恢复正常。

在意识到后台瘫痪是由于 DeepSeek 服务器过载,而自己作为开发者,长达数天,却没有收到任何提前告知,更没有任何售后维护服务时,林森有种“被抛弃”的感觉。

“就像家门口有一家小店,你是老顾客,办了卡,一直和老板相处很好。突然有一天,小店被评为米其林餐厅,老板把老顾客抛在一边,之前办的卡也不认了。”林森形容道。

作为 2023 年 7 月开始部署 DeepSeek 的第一批开发者,林森为 DeepSeek 的出圈感到振奋,但如今,为了维持运转,他只能切换到 ChatGPT,毕竟“ChatGPT 虽然贵一点,但至少稳定。”

当 DeepSeek 从口口相传的小店变成网红打卡的米其林餐厅,更多和林森一样调用无门的开发者们,纷纷开始逃离 DeepSeek。

2024 年 6 月,小窗 AI 问答机在产品早期阶段就接入了 DeepSeek V2,让小窗合伙人娄池颇为惊艳的是,那个时间点,DeepSeek 是唯一能全文背诵《岳阳楼记》不会出错的大模型。因此,团队用 DeepSeek 承担了产品最核心的功能角色之一。

但对于开发者而言,DeepSeek 虽好,稳定性却始终有所欠缺。

春节期间,不仅是 C 端用户访问繁忙,开发者们也时常无法调用 DeepSeek,团队决定选择几个已经接入 DeepSeek 的大模型平台同时调用。

毕竟,“现在已经有几十个平台有满血版 DeepSeek R1 了。”用这些大模型平台的 R1,配合 Agent 和 Prompt,也能满足用户的需求。

为了争抢 DeepSeek 外溢而出

的开发者群体,有头部云厂商开始面向开发者频频举办活动,“参与活动就免费送算力,如果不大批量调用,小开发者几乎能免费使用。”易标 AI 技术总监杨惠超表示。

不过,DeepSeek 热度当前,在第一批开发者们出逃之际,更多开发者仍在蜂拥而至,希望蹭到前者的流量红利。

郝鉴创业的项目,是通过调用 DeepSeek 的 API 来进行角色扮演的 AI 陪伴 APP,2月2日上线首周就获得了约3000名的活跃用户。

尽管有用户反馈 DeepSeek 的 API 调用时有报错,但已经有 60% 的用户希望郝鉴尽快推出安卓版。在郝鉴的社交媒体后台,每天都有至少几十个用户私信要下载链接,“搭建在 DeepSeek 上的 AI 陪伴平台”,无疑成了 APP 出圈的新标签。

经字母榜统计,在 DeepSeek 官网内收录的接入 DeepSeek 的各类 APP 名单,2025 年前 APP 名单仅有 182 行,如今已经扩展到了 488 行。

一面是 DeepSeek 成为“国产之光”爆火出圈,7 天涌入 1 亿用户,另一面,则是第一批部署在 DeepSeek 上的开发者,正因为过载人流导致的服务繁忙,而纷纷转投其他大模型。

对开发者而言,长时间的服务异常不再是简单的故障,而演变为代码世界和商业逻辑间的裂缝,他们被迫在迁移成本下进行生存演算,无论是涌入,还是逃出,开发者们都需要面对 DeepSeek 爆火带来的余震。

春节期间小程序后台被迫瘫痪三天后,至大年初六,为了保证程序正常运转的林森,离开已经部署一年多的 DeepSeek,转回了 ChatGPT。

即使 API 调用价格高了近 10 倍,但此时保证服务的稳定,成了优先级更高的选项。

值得注意的是,开发者离开 DeepSeek 转向其他大模型,并不如用户在 APP 内切换调用模型一般轻松。“不同的大语言模型,甚至是同一个语言模型的不同版本,对于提示词的反馈结果都有细微的差别。”即便林森仍在持续调用 ChatGPT,将所有关键节点从 DeepSeek 迁移 ChatGPT,并保证稳定且高质量的内容反馈,仍然花了他半天多的时间。

切换这个动作本身也许只需要两秒钟,但“更多开发者,换一个新模型要花上一个星期反复调整提示词,并重复测试。”林森告诉字母榜,

在林森这样的小开发者们看来,DeepSeek 服务器不足可以理解,只是如果能提前通知,能避免许多损失,无论是时间成本,还是

APP 维护成本。

毕竟,“登录 DeepSeek 开发者后台需要手机号注册,只需要一个短信,就可以提前告知开发者。”如今,这些损失将由曾在 DeepSeek 默默无名时就开始支持他们的开发者自己承担。

当开发者和某个大模型平台深度耦合,稳定性无疑成了不必宣之于口的契约,一个频繁波动的服务接口,足以让开发者重新审视对平台的忠诚度。

就在去年,林森在调用 Mistral 大模型(法国头部大模型公司)时,因为 Mistral 账单系统错误而重复付费,在他发出邮件后,Mistral 不到 1 小时就纠正了问题,并附上了 100 欧元的代金券作为赔偿。这样的应对,也让林森产生了更多信任。如今,他也将一部分服务迁回到了 Mistral。

易标 AI 技术总监杨惠超则在 DeepSeek V3 版本发布之后,就开始酝酿一场逃离。

不用 DeepSeek 来写诗或者吐槽,如果用 DeepSeek 来写标书呢?负责公司内 AI 标书项目的杨惠超,在 DeepSeek 推出 V3 版本后已经着手寻找替代方案。对他来说,在标书这样的专业领域,“DeepSeek 稳定性越来越不足。”

DeepSeek R1 版本火出圈的推理能力,并不吸引杨惠超。毕竟,“作为开发者,软件主要的推理能力是靠程序和算法,并不是太依靠模型的基础能力。底层即使用最老的 GPT 3.5,依靠算法纠正都可以产出一个很好的结果,模型只要回复答案稳定就可以。”

在实际调用过程中,DeepSeek 在杨惠超眼中,似乎更像是一个聪明却会偷懒的“好学生”。

升级 V3 版本后,杨惠超发现,DeepSeek 对一些复杂问题有了更高的回答成功率,但稳定性却也攀升到了难以接受的程度,“现在问 10 条问题,至少有一条输出不稳定,在要求生成的内容之外,DeepSeek 往往喜欢自由发挥,额外生成和问题无关的内容。”

比如,标书中不允许出现错误字符,同时,大模型返回的结果,开发者们往往指定用 Json 结构(用指令每次调用大模型使得稳定返回固定字段)去输出数据,便于后续函数调用,但出现错误或者不准确,都会导致后续调用失败。

“DeepSeek R1,或许相对此前的 V3 版本,推理能力提升了很多,但是稳定性达不到商业化的水准。”在 @生产力 Mark 账号内,杨惠超提到。

作为 2024 年初,DeepSeek-coder 时期就加入的第一批用户,杨惠超并不否认 DeepSeek 是一个好学生,只是如今,为了保证生成

标书的质量和稳定性,杨惠超只能将目光转向国内其他更偏 B 端用户的大模型企业。

毕竟,曾经被称为 AI 界拼多多的 DeepSeek,正是凭借着性价比的标签,迅速聚集起一批中小 AI 开发者。但现在想直接稳定地调用 DeepSeek,就必须进行本地部署。“部署一个 DeepSeek R1,需要 30 万~40 万元的成本,如果用线上的 API 计算,30 万元我一辈子都用不完。”

既不够便宜,又不够稳定,调用无门的杨惠超们,正在批量离开 DeepSeek。

曾经,林森们是第一批坚定选择 DeepSeek 的人。

2024 年 6 月份,林森在开发自己的 AI 小程序少年听世界时,曾经对比了当时国内外几十家大模型平台。他需要用大模型每天处理数千条新闻,并且筛选、排序,找出适合青少年听的科技、自然新闻,并且对新闻文字进行处理。

这不仅要求大模型聪明,还得便宜。

涉及到每天数千条的新闻处理,对 token 的消耗极大,对独立开发者的林森来说,ChatGPT 模型很贵,只适合处理核心环节,对大量文本的快速筛选和分析,还要靠别的价格更低的大模型做支持。

同时,无论是国外的 Mistral、Gemini,还是 ChatGPT,调用都很繁琐:你需要在海外有一个具体的服务器,还要做中继站,同时需要用国外的信用卡购买 token。

林森便是通过英国朋友的信用卡,才能在 ChatGPT 账户完成充值。而一旦服务器在海外,API 响应速度也会有所延迟,这都让林森将目光投向国内,寻找一款 ChatGPT 平替。

DeepSeek 让林森颇为惊艳。“当时 DeepSeek 并不是最有名的,但却是反馈最稳定的。”以每 10 秒请求一次 API 调用为例,其他国内大模型 100 次内可能有 30% 的时候,不返回任何内容,但 DeepSeek 次次返回,并都能维持不逊色 ChatGPT 和 BAT 们大模型平台的回复质量。

而相比起 ChatGPT 和 BAT 们的大模型 API 调用价格来说,DeepSeek 真的太便宜了。

林森将大量新闻阅读和初步分析的工作交给 DeepSeek 后,发现 DeepSeek 的调用成本 10 倍低于 ChatGPT。经过指令优化后,每天调用 DeepSeek 的成本低到 2~3 元,“可能跟 ChatGPT 相比,它不是最好的,但 DeepSeek 的价格是极低的,对于我的项目来讲,它的性价比非常高。”

性价比,成了开发者们选择 DeepSeek 的首要原因。2023 年,杨惠超一开始将公司的 AI 项目从 ChatGPT 切换到了 Mistral,主要便是为了控制成本。随后 2024 年 5 月份 DeepSeek 推出 V2 版本,将 API 打到了 2 元每百万 token,这无疑是对其他大模型厂商的降维打击,这也成了杨惠超将公司做 AI 标书工具的项目切换到 DeepSeek 的缘起。

同时,经过测试后,杨惠超发现,国内早已靠云服务在 B 端吃下市场的 BAT 们,“平台太重了”。

对于易标 AI 这样的初创公司来说,如果选择 BAT,会面临云服务的捆绑消费。对于只是简单调用大模型服务的杨惠超来说,无疑 DeepSeek 的 API 调用更加省事。

在迁移成本上,DeepSeek 也胜了一筹。

无论是林森还是杨惠超,初始的 APP 开发都是基于 OpenAI 的接口形式,如果切换到 BAT 们的大模型平台,都要把底层重新开发一遍。但 DeepSeek 兼容 OpenAI like 接口,切换大模型只需要修改平台地址,“1 分钟无痛切换。”

小窗 AI 问答机正式销售的第

一天就搭载了 DeepSeek,并将 5 个核心角色中语文和作文指导的角色交给了 DeepSeek 进行构建。

作为合伙人,娄池也在去年 6 月就被 DeepSeek 惊艳到。“DeepSeek 在中文理解上能力很棒,是那个时间点唯一全文背诵《岳阳楼记》不会出错的大模型。”娄池告诉字母榜,相比起其他大模型中规中矩、班味十足的文档式输出,用 DeepSeek 教孩子写作文,往往赢在了写作的想象力上。

在社交媒体风靡用 DeepSeek 写诗、写科幻小说之前,DeepSeek 的华丽文风,就让小窗 AI 团队眼前一亮。

对于开发者们来说,他们仍在期待 DeepSeek 恢复调用,眼下无论是迁移到 BAT 们部署了满血版 DeepSeek R1 的平台,还是转向其他大模型厂商,都似乎是“菀菀类卿”。

但竞争对手正在努力追平 DeepSeek 深度推理的出圈特长。

国内,近期百度、腾讯都陆续在自研大模型中加入深度思考能力;国外,OpenAI 也在 2 月紧急上新“Deep Research”,把推理大模型的思考能力用于联网搜索,并将对 Pro、Plus 和 Team 用户开放。谷歌人工智能实验室(Google DeepMind)也在 2 月发布了 Gemini 2.0 模型系列,其中 2.0 Flash Thinking 实验版本则是一款增强推理能力的模型。

值得关注的是,DeepSeek 仍然以文本阅读为主,但不管是 ChatGPT 还是 Gemini 2.0,在支持深度思考之余,都已经将推理能力引入多模态,支持视频、语音、文档、图片等多种输入模式。

对 DeepSeek 来说,在追赶多模态之余,更大的挑战还来自竞争对手在价格上的逼近。

在云平台部署侧,一众头部云厂商都选择接入 DeepSeek,一边分食流量,一边靠云服务绑定客户。对 DeepSeek 大模型的调用,某种程度上甚至成了绑定企业云服务的“赠品”。

百度创始人李彦宏近期提出,在大语言模型领域,“每 12 个月,推理成本就可以降低 90% 以上。”

在推理成本下降的趋势下,BAT 们的 API 调用价格持续走低已成必然,DeepSeek 的性价比优势,正在迎来大厂新一轮价格战的压力。

不过,大模型 API 价格战只是起步,面向开发者,大模型厂商们还拼起了服务。

林森接触过大大小小众多的大模型平台,让他印象深刻的是,某科技大厂会有专门的客户经理进行对接,无论是不稳定还是出现技术问题,都会主动和开发者联系。

而尽管作为一个开源的大模型平台,目标是为开发者提供更普惠的 AI 支持,DeepSeek 甚至在官网没有面向开发者出具发票的入口。

某头部大厂 AI 产品经理告诉字母榜,有的互联网公司 leader 坚持用 DeepSeek 替换原有大模型,完全不管替换模型重新调整 Prompt 所耗费的时间。同时,即便满血版的 DeepSeek R1,也有不少通用能力如 Function calling 等并不支持。

相比起用云服务跑通了 B 端服务场景的 BAT 们,在便捷度上,DeepSeek 仍然差着 AI 大厂一截。

只是 DeepSeek 的流量效应暂未褪去,赶潮人依然众多。

有分公司宣称接入 DeepSeek,只是开始调用 API,充值了几百块。

有的公司公告部署了 DeepSeek 模型,但其实只是让员工看了看 B 站教程,下载了一键安装包。在这波 DeepSeek 热潮里,泥沙俱下,鱼龙混杂。