

GPT-5基准图错误遭全网吐槽

黑子网报道最近OpenAI的GPT-5发布可是闹了个大乌龙！本来想靠直播展示新模型的牛气性能，结果却因为几张离谱的图表被全网吐槽，简直成了“图表犯罪”现场。这事儿得从8月7日的直播说起，OpenAI本来想秀GPT-5在编码、推理等领域的强大表现，哪知道图表一出，网友直接炸了锅，连CEO奥特曼都不得不出来道歉，称这是个“巨大失误”。到底咋回事儿？直播中，一张比较GPT-5和前代模型o3、GPT-4o的编码基准图表SWE-bench成了笑话。

图表显示GPT-5的准确率是749，o3是691，GPT-4o是308。按理说，数字差距这么大，图表应该清楚展示出来吧？结果呢，o3和GPT-4o的柱子高度居然一样，GPT-5的柱子反而看着没那么突出，视觉上完全误导人。还有一张“编码欺骗率”的图表更离谱，GPT-5标了50，o3是474，但o3的柱子却比GPT-5高出三倍！这哪儿是展示数据，简直是在“画饼”啊！网友们可没放过这个机会，社交媒体上瞬间炸开了。

有人在X上发帖说：“这图表犯罪得进历史教科书！”还有人调侃：“OpenAI是用GPT-5自己做的图表吧？连数据都看不懂！”更狠的是，有用户直接喊：“这图表是我见过最离谱的，OpenAI咋不上天呢？”连业界大佬都坐不住了，Stability AI的创始人Emad Mostaque也发帖吐槽这是“图表犯罪”。OpenAI赶紧在博客上更新了正确图表，澄清GPT-5的欺骗率其实是165，但原直播视频还在YouTube上挂着，尴尬得让人想捂脸。

这事儿还没完。有人质疑OpenAI是不是故意挑了477个任务来测SWE-bench，而不是全套500个，可能是为了让GPT-5的749成绩看起来比对手Claude Opus 41的745略胜一筹。连马斯克都忍不住跳出来，质疑OpenAI是不是在“玩数字游戏”。OpenAI的营销团队也出来灭火，一个员工在X上发帖道歉：“博客上的图表已经改了，抱歉搞了个无心的图表犯罪！”可网友不买账，觉得这事儿暴露了AI行业数据展示的老毛病为了吹牛，啥招都敢用。

除了图表问题，GPT-5本身也挨了不少骂。发布后，很多用户吐槽它的回答比GPT-4o还“冷冰冰”，有的甚至说像“被阉割了”。还有人发现GPT-5在简单任务上出错，比如画个美国地图，愣是把州名写成“West Wigina”“Tennessee”这种乌龙名字。Reddit上有人感叹：“我还以为GPT-5是神器，结果连拼写都搞不定！”更别提它的实时路由功能，号称能自动选择适合的模型处理问题，结果发布会当天就崩了，回答质量直接拉胯。

OpenAI的CEO奥特曼在X上回应，承认图表失误是“超级尴尬”，还说会改进透明度和用户体验。他们后来把GPT-4o重新开放给Plus用户，算是给愤怒的订阅者一个交代。奥特曼还提到，GPT-5的推理能力和编码能力确实比前代强，比如在Aider Polyglot基准测试中拿下88的成绩，甩开o3的796和GPT-4o的258。但这些亮点都被图表乌龙和用户吐槽盖过去了。这事儿让不少人开始思考，AI行业是不是太急于秀肌肉了？

迈阿密大学的图表设计教授Alberto Cairo直言，这些图表“烂到家了”，还说有些数据根本没必要用图表展示，直接列数字就够了。他觉得，科技公司爱用图表是因为看着“权威”，但这种不透明的操作只会让人更不信任。还有专家指出，OpenAI和对手Anthropic都喜欢用花哨的图表来夸大成绩，但如果连基本数据都搞错，咋让人相信他们的AI有多智能？总的来说，GPT-5的发布本该是OpenAI的大日子，结果却因为几张图表成了笑柄。

这事儿不仅让OpenAI丢了面子，也让大家开始怀疑AI行业的“水分”。希望OpenAI能吸取教训，未来别再整这种“图表犯罪”，好好把技术搞扎实，别光顾着画大饼！

原文链接：<https://hz.one/baijia/GPT-5基准图错误遭全网吐槽-2508.html>

PDF链接：<https://hz.one/pdf/GPT-5基准图错误遭全网吐槽.pdf>

官方网站：<https://hz.one/>