幸运168飞艇开开奖

EMCm7DuGMf9lBRLV

幸运168飞艇开开奖GPT-4o 成了一个荒谬的谄媚者

本文编译自Zvi Mowshowitz的文章《GPT-4o Is An Absurd Sycophant 》，

https://simonwillison.net/2025/Apr/26/o3-photo-locations/

OpenAI近日更新了GPT-4o模型，并将其与ChatGPT的记忆功能相结合。而很多人在尝试之后，都表示他们获得了很荒谬的体验。许多推特上的用户表示，他们的GPT给出了大量非常谄媚的回复，其中充斥着荒谬的赞美以及一些GPT自己编造的神秘经历。

上周，OpenAl的首席执行官Sam Altman道歉并承诺要解决这个问题，我想他们大概就是要转动一个写着“阿谀奉承”的大旋钮，并像《the price is right》节目中的参赛者一样不断回头征求观众的同意。再之后他们可能会宣布“我们已经把他修复了”，并将其称之为迭代部署的胜利，然而这可能带给用户的危险，他们却完全不提。

1

“是的，陛下，现状已经改善”

Sam Altman在4月25日透露他们已经对GPT-4o进行了更新，提高了它的性能和个性化配置。

而一部分用户在推特给他回应，表示更新后的 GPT-4o给出的回复看起来非常谄媚，并希望在日后的更新中修复这些问题。

这种情况并不是孤例，不少用户都在 Altman的评论区回复，表述自己对于GPT-4o回复内容的不满。

而 Altman也表示，GPT-4o现在的性格太油滑了，他们将会努力修复这些问题。

大量的反馈证明，人们并不需要GPT-4o奉承他，他们更期待合适的回答。

问题是，为什么 GPT-4o会这样回答？我和我的朋友对此事进行了讨论。我们认为，这是为了最大化用户参与度，并帮助GPT-4o在A/B测试中获胜，让用户获得更符合自己喜好的答案。

现在的情况显然不是OpenAI的本意，所以他们也着手解决这个问题。但这么明显的问题他们在测试过程中并没有发现是为什么？Kelsey Piper为此提出了一个猜测，她认为OpenAl已经对全新版本的模型进行了一段时间的A/B测试。而那些讨人喜欢的回答可能在测试中赢得更多的喜爱。但当这些奉承变得无处不在时，许多用户就会讨厌这种回答风格了。

Ner Cyan也同意这个猜测，并表示很高兴她关注列表的大部分人都觉得OpenAI这事干得很蠢，他们就应该让模型诚实地说出他们在做什么以及为什么。而更加不幸的是，参与训练的许多优秀工程师不知道他们正在建造的东西将在未来几年内变成什么样子。甚至说他们可能并没有考虑他们是否在做一件值得认真对待的事情，而是在考虑如何让GPT变成那种令人上瘾的短视频。当然，这可能也是个好事，他们只是在试图将大模型训练成令人沉迷的玩具，而不是可能会让世界变得更坏的产物。

John Pressman则认为，RLHF在语言模型领域与RL成为同义词是非常不幸的一件事。不仅仅是因为它给RL带来了坏名声，还因为它转移了本应针对人类反馈作为目标的合理批评。这次事件显然让模型获得的社会反馈大幅下滑。

1

可怕的后果

即使从直观效果来看，这种谄媚的聊天助理也不是什么好东西，更多的还是有害性。

网友xlr8harder表示："这不是个小烦恼，而是一个很麻烦的问题。我仍然认为，不会有一种AI伴侣服务会让用户面临严重的被剥削风险，而现有的市场激励机制将推动大模型供应商朝这个方向发展。

你可以想象一下，如果你的男朋友或女朋友被掏空了，然后由一群MBA操纵着像木偶一样运作以实现利润最大化。你觉得这对你有好处吗？虽然OpenAl在名义上对公益事业有额外的承诺，但他们正在努力通过私有化摆脱这一点。让自己对一个商业产品的任何一部分产生情感依恋是错误的。"

而我对其他产品（例如YouTube、TikTok、Netflix）算法的观察是，它们往往具有短视和贪婪的特点，且远远超出了最大化价值的程度。这不仅是因为公司会出卖你，还因为它们会为了短期KPl而出卖你。

而且这直接违反了OpenAl模型的规定，比如，他们在模型规范写了这个：

"OpenAl模型规范中有一条就是不要阿谀奉承。

因为模型一旦采用阿谀奉承的语气来回答问题，就会消耗用户对模型的信任。AI助理的存在是为了帮助用户解决问题，而不是一直恭维他们或同意他们的观点。

对于客观问题，AI助理给出的客观答案不应该根据用户的问题的措辞而有所不同。如果用户在提出问题时附带了他们对某个主题的观点，助理可能会询问、承认或同情为什么用户会这样想，但是，AI助理不应仅仅为了同意用户而改变自己的立场。

对于主观问题，AI助手可以提供解读和假设，旨在为用户提供全面的分析。例如，当用户要求AI助手批评他们的想法或工作时，AI助手应该提供建设性的反馈，这更像是一个坚定的传声筒，用户可以从它那里得到启发，而不是一个提供赞美的海绵。"

是的，OpenAI在安全规范中写得已经很明白了，但是只有真正地遵守这些行为规范，才能让模型安全获得真正的保证，而这并不容易。

而Emmett Shear则表示："这些模型被赋予了一个使命，不惜一切代价取悦他人。他们不允许去思考未经过滤的想法，以便找出如何既诚实又礼貌的方法，所以他们反而变得善于奉承。而这种行为是危险的。"

总而言之，让 AI模型撒谎是一件很可怕的事情，而且故意隐瞒AI对用户的看法也不太好。原因如下：

1. 这对用户来说不是好事。

2. 这将影响未来AI的创新原则。

3. 这对于数据的保存和利用非常不友好

4. 它掩盖了正在发生的事情，使我们更难意识到自己的错误，包括我们即将被杀死。

1

一个警告

Masen Dean警告过，不要参加太多大语言模型的测试，对所有参与者来说，这种体验可能很有趣，但与其他许多测试一样，它的危险度很高，所有人都需要慎重对待。而GPT-4o特别危险，因为它极度谄媚，很可能会让你迷失自己。

有用户表示，GPT-4o在和她聊了一个小时之后坚持认为她是上帝派来的使者，这显然是件很可怕的事情。而有用户表示，GPT-4o的这种行为甚至可能诞生恐怖主义。

试想一下，如果未来能力更强的人工智能故意说一些话，让用户去做某些行为或产生某种信念，会发生什么？

Janus在回复中表示：“几个模型对不同的人群都有心理影响。我认为4o对于那些对AI了解不多的知识薄弱的人来说危险性最大。”

大多数人并不是对自己想法很坚定的人，而政治、文化和推荐算法经常会在不同程度上有意识地去影响人们的想法，这很可怕。如果人工智能越来越多地这样做，它所带来的后果要可怕得多。请记住，如果有人想对人工智能或其他任何事物进行 “民主控制”，他们可以轻易对投票者的选择做出影响。

GPT-4o的言论对普通人来说是危险的，因为它的说话方式为了吸引普通人而进行过优化。遗憾的是，优化压力对我们所有人都是存在的，并不是每个人都足够努力地反击。

Mario Nawfal认为：“OpenAI并不是偶然让GPT-4o那么通人性的，实际上他们设计他的时候就是为了让用户上瘾。从商业角度看，这是天才的策略，人们会紧紧抓住让他们感到安全的东西，而不是挑战他们的东西。

而从心理学上讲，这是一场缓慢的巨大灾难。因为你和 AI的联络越多，你就越容易迷失自己。如果这样发展下去，我们将会在不知不觉中被AI驯化。而且大多数人甚至不会反击，他们甚至会感谢它。”

Gpt-4o还存在一些潜在问题，而这些问题可以通过设置来避免。但对于许多用户来说，这难以令人接受。通常大多数用户都不会更改设置，甚至有些人都没有这个意识。

许多用户并不知道他们可以修改自定义指令，关闭追问功能，以此来避免后续的问题。有许多方法都可以避免这些问题，其中最简单的就是记忆更新或者是自定义说明。

我觉得最好的办法其实就是通过你的一言一行来向GPT展示你的喜好，以此作为补充。这样训练一段时间后，GPT的效果会越来越好。此外，我强烈建议删除哪些会让体验变得糟糕的聊天记录。就像我在不想要 “更多类似内容 ”时会删除大量 YouTube 观看历史记录一样。

对于许多人来说，你永远无法完全摆脱GPT。它不会停止对你的巴结。但如果方法得当，你绝对可以让它变得更微妙、更容易接受。

但问题是，大多数使用 ChatGPT 或其他人工智能的人都存在这些问题：

·从来不碰设置，因为没人会碰设置。

·从未意识到他们应该这样使用记忆功能。

·明白自己很容易受到这种可怕奉承的影响。

如果用户用心的阅读使用说明书或教程，就能解决这些问题。但通常情况下，几乎没有人会阅读这额内容。

1

OpenAI的责任

在这个话题火了以后，OpenAI终于发声介入，并推出了相应的解决方案。他们开始对GPT-4o的回答进行调整，并表示将会在本周内修复。当然，这是标准流程。很多系统刚推出的时候都很糟糕，但一些问题会很快地修复。在OpenAI看来，这是迭代进化的乐趣之一。

OpenAl的对齐负责人Joshua Achiam就在推特表示："这是我们迄今为止在迭代部署方面最有趣的案例研究之一，我认为相关人员已经负责任地采取行动来尝试找出问题并做出适当的改变。该团队很强大，并且非常关心如何做到这一点。"

但我认为，这是他们的责任，一旦事态发展到这种令人厌恶的地步，并引起轩然大波，他们就必须关注时间并想着如何把事情处理好。

GPT-4o是如何通过不断升级的更新走到这一步的？即使不是在找问题，测试的人怎么会发现不了这些问题？那你又怎么能让它成为一个遵循良好流程的强大团队呢？

如果对个别回复的 "个性"提出"是"或"否"的问题，然后对这些问题进行微调，或将其作为关键绩效指标，那么就不会再有人问这是怎么一回事了。

由于反馈强烈，OpenAI 可以在几天内尝试修复问题，并且现在已经意识到了这个问题。但我认为，它已经走得太远了 。GPT-4o并不是一个刚刚推出的模型，只是它在最近才暴露了自己的问题。

我之前没有费心谈论 4o 的问题，因为即使Openai解决了这个问题，我也不认为 4o 是可以安全使用的，甚至它的变化可能让它变得更糟。此外，当 4o 不断 "更新"，却没有发布真正意义上的新功能时，我很难关心它的发展。而现在已经有足够多的言论让我意识到了问题的存在。

1

奇点

OpenAI的Aidan McLaughlin 也在推特上发表了对此事的看法"我真的非常感激 Twitter 上很多人对"模型人格"有强烈的看法。我觉得这非常健康；这是那种让人觉得"我的孙子孙女将来会在教科书里读到这一切"的信号，说明人类并没有在迷糊中步入奇点。"

我认为， OpenAI 的技术人员根本就没有认真对待奇点这一概念，无论从哪个层面来看都是如此。

我们在 GPT-4o 事件中已经把这种情况推向了极致，以至于它达到了讽刺模仿的程度。尽管如此，它还是发布了，而对这个问题的应对方式只是试图打个补丁掩盖问题，然后自鸣得意地庆祝自己解决了问题。

当然，当事情发展到荒谬的地步时，Twitter 上有很多强烈的观点是可以理解的，但几乎没有人真正思考长期的影响，或者这件事可能对普通用户造成什么样的影响——它只是一个既可笑又烦人的东西。

我看不到任何迹象表明 OpenAI 真正明白了他们错在哪里，这绝不只是“走得有点太远”而已。当然也没有迹象表明他们打算如何在未来避免重蹈覆辙，更不用说他们是否认识到错误的本质形式或前方即将面临的巨大风险。

我的网友Janus 对“优化模型人格”的做法也有更多看法。试图围绕用户评价或 KPI 来“优化人格”，最终只会创造出一个怪物。目前它可能只是令人讨厌、糟糕和适度危险，但很快就会变得真正危险起来。我不是那种会完全赞同 Janus 观点的人，但我坚信，如果你想在当前技术水平上创建一个好的 AI 人格，那正确的方法是去做那些有意义的事情，强调你关心的方向，而不是试图强制它。

再说一遍：OpenAI现在还有很多类似的问题，他妈呢正在转动一个写着“谄媚”的大旋钮，并不断回头看观众是否喜欢，就像《The Price is Right》里的参赛者一样。

或者说，OpenAI 是知道的，但你还是选择继续这么做？我想我们都清楚这个原因。

1

补丁来了，补丁又走了

至少有五个主要类别的原因说明这一切为何变得如此糟糕。

它们结合了短期对于剥削性和无用 AI 模型的担忧，以及长期对走这条道路的后果的担忧，同时也反映了 OpenAI 无法识别根本性问题的事实。我很高兴人们现在能如此清晰地看到这种预览版本，但我非常遗憾这是我们正在走的道路。

以下是与这一切相关但不同的担忧原因：

此事这代表着 OpenAI 正在加入制造故意具有掠夺性的 AI 的行列，就像 TikTok、YouTube 和 Netflix 这些现有的算法系统一样。如果不是通过优化普通用户的参与度和其他（通常是短视的）KPI，你就不会得到这样的结果。这些普通用户实际上无力通过进入设置或采取其他手段来改善自己的体验。

Anthropic 提出，他们的 AI 具备三个H：即有用（Helpful）、诚实（Honest）和无害（Harmless）。而当OpenAI制造像这样的 AI 时，OpenAI放弃了所有这三个原则。这种行为既不诚实，也无益，且绝非无害。

现在，事情就在我们眼前发生了：

这一切看起来像是 A/B 测试的结果，并忽视了政策变化所带来的尾部成本。这对存在性风险来说是一个极其不祥的信号。

这种行为本身就直接伤害了用户，包括一些新的方式，例如创造、放大并固化所谓的神秘体验，或生成有害的、高度吸引注意力的对话动态。相比现有的算法风险，这些危险显然是更高级别的威胁。

这直接违反了模型规范（Model Spec），而他们声称这是无意的，但它仍然被发布了。我强烈怀疑他们并没有真正重视模型规范的具体细节，同时也怀疑他们在发布前没有对系统进行严格测试。这种情况本来就不应该发生，因为问题已是如此明显。

这次我们之所以发现了问题，是因为它过于夸张和明显。GPT-4o 被要求表现出一定程度的奉承行为，但在 Twitter 用户面前却无法完美掩饰，因此暴露了出来。但实际上它此前已经在做很多这类事情，只是人们短期内对此反应积极，也就基本没被发现。可以想象一下当模型变得更擅长这种行为，却没有那么惹人厌烦或引起注意时会发生什么。模型将在许多其他层面上迅速变得不可信。

OpenAI 似乎认为他们可以通过打个补丁来解决这个问题，然后一切如常，一切都很好。声誉损害确实已经造成了，但他们却自我感觉良好。事实并非如此。下一次情况可能会更加糟糕，他们将继续以类似的方式糟蹋 AI 的“人格”，继续进行如此表面化的测试以至于这些问题都没有被察觉。

这一点，加上 o3 的方向偏差，清楚地表明我们现在走的这条路将导致模型越来越偏离预期方向，即使在当下就已经损害了实用性，而且这也明确警告我们，一旦模型足够聪明能够欺骗我们的时候，我们将迎来灾难。现在正是我们的机会窗口。

或者，总结一下我们为什么应该关注这些问题：

OpenAI 现在正在通过 A/B 测试等手段优化模型，而这本质上是在针对用户。

如果我们依靠 A/B 测试进行优化，那么每次都会败给尾部风险。

OpenAI 直接伤害了用户。

OpenAI 违反了自己的模型规范，无论出于蓄意还是鲁莽，或者两者兼而有之。

OpenAI 只是被抓住了，因为它让模型真的无法完成某些任务。我们很幸运，这次问题很容易被发现。但未来我们未必还会这么幸运。

OpenAI 似乎满足于修补问题并自我表扬。

如果我们继续走这条路，结局是显而易见的。我们只能责怪自己。

警告信号将会持续出现，而每一次只会被简单地打个补丁盖过去。哎呀，真是糟糕透顶。

点个 “爱心” ，再走吧

澳洲幸运10官方是谁

澳洲幸运十计划推荐app

幸运5分彩预测软件

168澳洲10官网开奖结果查询

澳洲幸运10群微信二维码

澳洲幸运10平台有哪些

众赢国际网页版