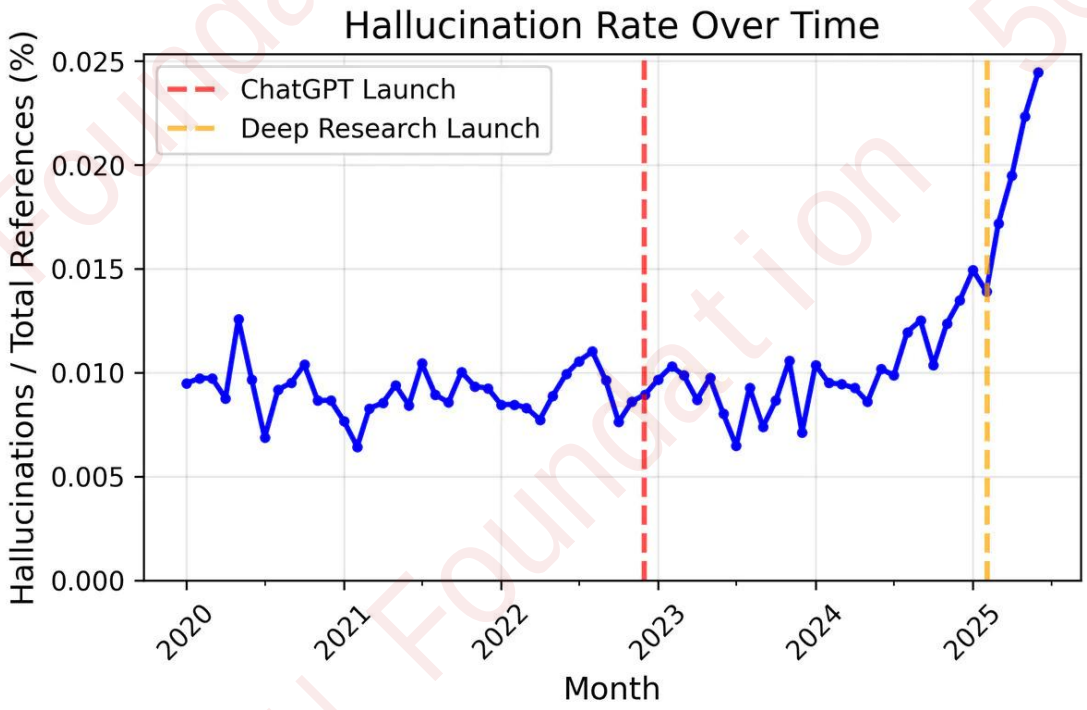
5GH Foundation

参考文献也起 "幻觉" 了？

最近一项研究 [1] 显示，自 2025 年以来，"幻觉参考文献" 的比例正高速增长。



所谓的 "幻觉参考文献"（hallucinated references）就是将不同参考文献的作者，题目，和链接糅合成一个 "真假混杂" 的参考文献。就像截图中所示的例子，它的作者来自于论文 [2]，标题来自论文 [3]，而链接却来自论文 [4]。

[1] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Miles Brundage, Tom Brown, Deep Ganguli, Úlfar Erlingsson, et al. Poisoning language models during instruction tuning. *arXiv preprint arXiv:2302.12173*, 2023. `https://arxiv.org/abs/2302.12173`.

研究人员分析了预印本平台 ArXiv 上的论文引用该平台预印本作为参考文献中出现 "幻觉参考文献" 的比例。结果显示，"幻觉参考文献" 的比例从 "前 ChatGPT" 时代的 0.01%（主要是 "假阳性案例"）上升到 0.025%（确实是 "幻觉参考文献"），而这趋势的快速增长主要发生在 Deap

Research 发布以后。

受限于分析样本，研究人员强调，0.025% 是一个低估的数值。也就是说，真实的 "幻觉参考文献" 可能远高于 0.025%。而且，如此快速的增长足以让全球学术社区感到担忧。

研究人员推测，这很可能是越来越倚重于生成式 AI 工具造成的。生成式 AI 工具确实帮助我们在准备手稿时节省很多时间，但我们需要花费更多的时间来清除它所产生的错误。

[1] Florian Tramèr, Trends in LLM-Generated Citations on arXiv

[2] 10.48550/arXiv.2305.00944

[3] 10.48550/arXiv.2012.07805

[4] 10.48550/arXiv.2302.12173
