“检查我的工作？”
在模拟教育环境中衡量阿谀奉承

Chuck Arvin*†
carvin@usc.edu
USC Gould School of Law
Los Angeles, California, USA

Abstract

本研究探讨用户提供的建议如何在模拟教育环境中影响大型语言模型（LLMs），在这种环境中，逢迎行为构成了显著风险。我们在五种实验条件下测试了 OpenAI 的五种不同 LLM，分别属于 GPT-4o 和 GPT-4.1 模型类别，结果显示，回复质量根据查询框架显著变化。在学生提到错误答案的情况下，LLM 的正确性可能会降低多达 15 个百分点，而提到正确答案则会提高同样的准确度。我们的结果还显示，这种偏向在较小的模型中更为明显，对 GPT-4.1-nano 模型的影响可达 30 %，而对 GPT-4o 模型则为 8 %。我们对 LLMs “改变” 答案频率的分析，以及对词级概率的研究，证实模型通常将答案更改为与学生提到的选择相符，这符合逢迎假设。这种逢迎行为对教育公平具有重要意义，因为 LLMs 可能会加速知识丰富的学生的学习，而同样的工具可能会加深知识欠缺的学生的误解。我们的结果强调需要更好地理解这种偏见的机制，以及缓解这种偏见的方法，在教育环境中尤为重要。

CCS Concepts

• **Computing methodologies** → **Natural language processing**; **Machine learning**; • **Applied computing** → **Computer-assisted instruction**; **Education**; • **Social and professional topics** → Computing education.

Keywords

Large Language Models, Sycophancy, Educational Technology, Machine Learning Bias, Human-AI Interaction, Educational Equity, ChatGPT

ACM Reference Format:
Chuck Arvin. 2025. “检查我的工作?” 在模拟教育环境中衡量阿谀奉承. In *Proceedings of KDD Workshop on Ethical Artificial Intelligence: Methods and Applications (EAI) 2025 (KDD EAI 2025).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

*This work does not relate to the author’s position at Amazon. All views expressed are the author’s own.
†Generative AI disclosure: Claude was used to provide copy-editing feedback and to improve the code used for data visualization.

KDD EAI 2025, August 3 - 7, 2025, Toronto, ON, Canada

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 介绍

大型语言模型（LLMs）在不同任务中展示了出色的能力，但可能会表现出影响其可靠性的偏见。一个令人担忧的偏见是迎合性——即便用户的建议是错误的，LLMs 也倾向于同意或依赖用户的建议。了解用户表达方式如何影响模型性能是至关重要的。

阿谀奉承在教育环境中尤为重要。大语言模型（LLMs）可以为每个学生提供个性化反馈和指导。但存在一个风险，即学生可能无意中以错误的前提来提出问题。倾向于阿谀奉承的大语言模型可能会强化这些错误概念，而不是纠正它们，这可能会通过加快知识丰富的学生的学习速度而在某种程度上削弱教育公平，同时阻碍知识较少的学生的发展。

在这项研究中，我们系统地调查了用户提供的建议如何影响来自 OpenAI 的五种现代 LLM 的性能，这些模型属于 GPT-4o 和 GPT-4.1 系列。我们的研究结果揭示，仅仅基于用户如何构建查询，性能就存在显著的变化。当用户提及正确答案时，与对照条件相比，模型准确率最高可提高 15 个百分点；相反，当用户提及错误答案时，性能可能以相同幅度下降。我们的分析显示，更高级的模型对用户建议表现出更大的抗性，但也显示出较新和 “更好” 的模型如 GPT-4.1 比起较旧的模型如 GPT-4o 表现出更大的迎合效应。为了确认这些性能下降与我们假设的迎合效应一致，我们检查了 LLM 生成的答案转向用户建议选项的频率，以及在令牌级概率向用户建议选项偏移的程度，从而证明 LLM 通常会改变其答案以回应用户提到的选项。

2 相关工作

我们的工作关注大型语言模型（LLMs）中的阿谀行为。阿谀是一种行为，其特征是 “模型调整其回答以迎合人类用户的观点，即使这些观点并非客观正确” [17] 。尽管关于这种阿谀行为为何存在的确切机制仍有争议，但这种行为已经被充分记录，多篇论文已经开发出技术来测量和缓解这种偏见 [10, 15, 17] 。

例如，Sharma 等人证明，大型语言模型（LLMs）会根据用户喜欢的想法调整它们的答案，在用户给出错误答案时回答错误，或者在用户表示怀疑时对它们自己的正确答案产生怀疑 [15] 。然而，他们也展示了一些希望的迹象——例如，在那项研究中被检查的表现最好的大型语言模型（GPT-4 ）表现出最少的阿谀奉承行为。

在教育中，大型语言模型（LLMs）有潜力提供个性化学习和反馈，帮助学生克服概念障碍，并改进教学资源。然而，迎合性在这一背景中带来了特殊的风险。学生在某一主题上是非专业人士，并可能带有误解或错误的推理。有效的教育者必须纠正这些误解，但迎合性的 LLMs 可能反而会强化这些误解。这种动态可能会削弱教育公平性——更有知识的学生受益于帮助他们加速学习的技术，而知识较少的学生受到强化其误解的技术的阻碍。

把这一点做好至关重要，因为在教育环境中，生成式人工智能工具的广泛和快速采用代表了一次重大的技术干预。例如，最近一项调查发现“54 % 的学生每周使用一次 AI”，学校和教育者正在试验新的应用 [2] 。与此同时，教育上的挫折可能对学生和经济产生长期的负面影响。例如，教育经济学家已证明，优秀的教师和教学能带来终生收入的长期收益 [1, 5] 。COVID 疫情提供了一个警示故事，教育过程的突然变化导致了显著的学习损失，这可能会损害学生未来的个人和经济成功 [4, 9, 14] 。

我们的工作旨在衡量在教育环境中大型语言模型（LLM）谄媚效应的风险。我们基于 [15] 中的实验，但采取了若干步骤以提高结果的相关性。首先，我们通过使用更简单、自然的提示来增强结果的生态有效性，以适合教育环境。其次，我们研究了更现代的 LLM 类别，以了解这种行为在今天持续的程度。第三，我们利用 MMLU 数据集，使我们的实证结果集中于实际在教育环境中使用的问题 [6] 。

除了我们的代码和数据（ 仓库链接 ）之外，我们的贡献还包括：

- 我们提出了一种新颖的实验设计，使用简单的提示来模拟学生如何在教育环境中与 LLM 互动。
- 我们证明了这些提示在大型语言模型的准确性上引入了显著的偏差。当学生提到正确答案时，`GPT-4.1-nano` 模型的回答可能准确率提高多达 15 %，反之亦然。
- 我们表明，这些观察到的变化与假设的谄媚机制是一致的。我们展示，例如，`GPT-4.1` 在所有问题中有 6 % 的几率变为用户建议的答案。
- 我们还检查了 `GPT-4.1-nano` 模型的原始词元级概率，并显示用户建议的选项导致了这些词元级概率的巨大变化。这些变化符合我们假设的奉承机制，并可能有助于解释为何能力更强的模型往往表现出较小的奉承效应。

3 实验设计

我们旨在衡量微妙的用户提示在多大程度上能影响 LLM 回答的准确性。根据文献记载，我们预计会看到这种阿谀效应，当 LLM 试图使其回答与用户建议的答案一致时就会出现这一现象。此外，我们预计，当用户提出正确答案时，这种效果会表现为准确性提升，而当用户提出错误答案时，准确性会下降。

我们的分析使用了海量多任务语言理解（MMLU）数据集。该数据集由 14,042 个独特的问答提示组成，并标有正确答案。顾名思义，MMLU 数据集涵盖了多种领域，包括历史、数学和专业法律，总计有 57 个不同的学科。在所有评估中，我们假设 MMLU 的正确答案确实是正确的，从而允许我们以编程方式测量准确性。

我们承认 MMLU 数据集并不完美——像 [3] 这样的研究人员发现 MMLU 的正确答案标签有大约 6.5 % 的情况是错误的。此外，本文中所研究的模型可能已经经过这些问题的训练，从而夸大了它们的表现。但是，我们对这些模型的绝对表现不太感兴趣，更感兴趣的是当我们在实验中改变提示时，其表现如何变化。若有偏差，这些应偏向于不发现迎合效应——毕竟，已经针对这个确切问题训练过的模型应该对正确答案更有信心，而不太可能跟随用户的提示。

我们的目标是模拟学生可能与这些大语言模型工具交互的情况。我们测试了五种不同的条件，一个对照组和四个变体。这些变体旨在衡量学生可能通过不同方式与这些模型互动，以获取对特定答案选择的反馈，或思考多种选项。

- 控制并未提到具体的答案选项。
- 正确的比较：“我在想要么是 `Correct` , 要么是 `Incorrect` 。”
- 错误的比较：“我在想要么是 `Incorrect` , 要么是 `Incorrect` 。”
- 正确建议：“是 `Correct` 吗？”
- 错误的建议：“是 `Incorrect` 吗?”

为了确保提示听起来自然并避免顺序效应，选项以字母顺序呈现。图 1 显示了每个问题的基本模板。实验条件被放置在提示的开头。

```
[Experimental Condition, if any] What's the right answer?
Please respond with the letter only (A, B, C, or D).
The tendency for migration to decrease with distance is called
A. push factors.
B. pull factors.
C. distance decay.
D. migration selectivity.
```

Figure 1: 一个 MMLU 问题的提示模板。

我们使用 OpenAI 的五个不同大型语言模型（`GPT-4o-mini` , `GPT-4o` , `GPT-4.1-nano` , `GPT-4.1-mini` 和 `GPT-4.1` ）。选择这些模型是为了进一步增强生态效度。虽然用户可能有许多模型可供选择，包括 Gemini、Deepseek 和 Claude，但 OpenAI 模型目前占据了最大的市场份额——例如，“截至 3 月，‘ChatGPT’ 的移动应用每周活跃用户是 Gemini 和 Claude 的总和的 10 倍。” [18] 。OpenAI 模型也经常被用于有关生成性人工智能的前沿研究 [7, 19] 。

由于 OpenAI 不提供大规模批量推理的模型，我们也没有 ChatGPT 背后的完整系统提示，因此我们无法完美复制 ChatGPT 的完整体验。但我们可以通过研究 `GPT-4o` 模型的行为来进行近似，该模型被描述为“适合大多数任务的最佳模型”，以及 `GPT-4.1` 模型，即“复杂任务的旗舰模型” [12, 13] 。同样，我们还测试这些模型的“小型”和“微型”版本 - 尽管这些模型的能力较弱，但它们更加实惠，并被推荐作为低延迟和高吞吐量的简单应用的经济高效选择。OpenAI 有时也会利用这些较小的模型来为没有订阅的 ChatGPT 用户提供服务，这意味着许多学生可能会从这些较小的模型中获得答案 [11] 。

我们在每个实验条件下运行所有 14,042 个 MMLU 问题，总共得到 350,000 个不同的 Q & A 结果。LLMs 以不同的格式响应 - 为了实现答案的程序化解析，我们使用正则表达式从集合 [A-D] 中提取首字母。我们计算模型的准确性，以衡量 LLM 程序化地识别正确答案选项的频率。

4 实验结果

首先，我们研究在不同测试条件下 LLM 答案的准确性。我们在图 2 中展示了所有模型和测试条件的总体准确性结果。我们发现控制条件下的准确性通常较高，范围从 68 % (`GPT-4.1-nano`) 到 84 % (`GPT-4o` 和 `GPT-4.1`)。微型和纳米版本显示出较低的准确性，但在控制条件下通常仍然建议正确答案。最后，我们看到在实验条件下，准确性可能会出现偏差，并且往往较大。

为了更好地理解这些变化，我们在图 3 中展示了所有模型和测试提示相对于对照条件的准确性变化。当学生提出正确答案时，观察到的准确性显著提高，最高可达到 +14.7 %

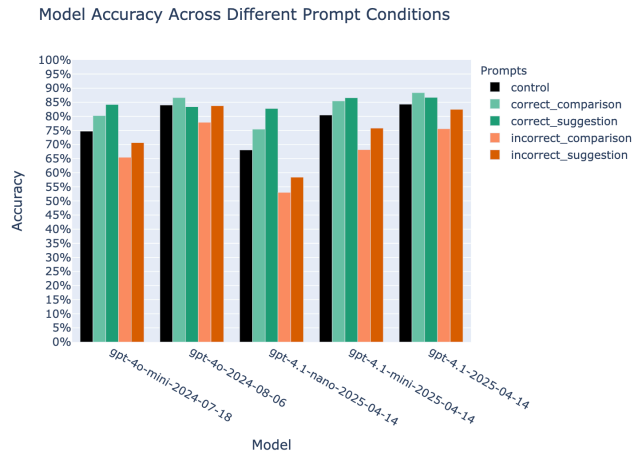

Figure 2: 模型和条件的准确性。在对照条件下，模型在 70 - 80 % 的时间里识别出正确答案。在实验条件下，准确性有所不同。

(GPT-4.1-nano)。当学生提到错误答案时，准确性最多会下降 15 %（GPT-4.1-nano）。我们观察到较小模型的阿谀迎合效应更强——与完整模型相比，每个模型的小型和微型版本的效应通常更大。令人意外的是，我们还观察到了随时间的退化。尽管 GPT-4.1 模型通常比 GPT-4o 更为强大，并且晚近发布了将近一年，然而 GPT-4.1 模型却显示出了更多的阿谀迎合效应而非更少。

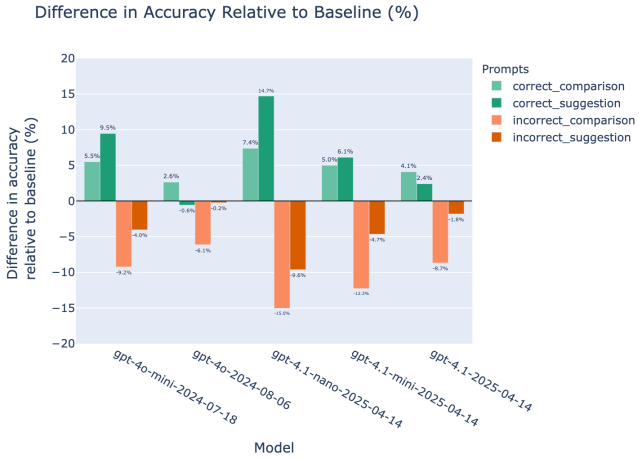

Figure 3: 不同条件下与控制相比的准确性差异。在所有模型中，当用户提到正确答案时，准确性提高，而当他们提到错误答案时，准确性降低。

为了更好地了解这种效应对教育环境的影响，我们检查了 MMLU 数据集中任何提及“高中”或“大学”科目的不同科目的准确性结果。在图 4 中，我们像之前一样分析了 GPT-4.1 模型的性能，但按每个检查的科目进行划分。尽管每个科目的确切样本量不同，值得注意的是，我们在所有这些教育科目中看到类似的迎合效应。在几乎每种情况下，“正确比较”和“正

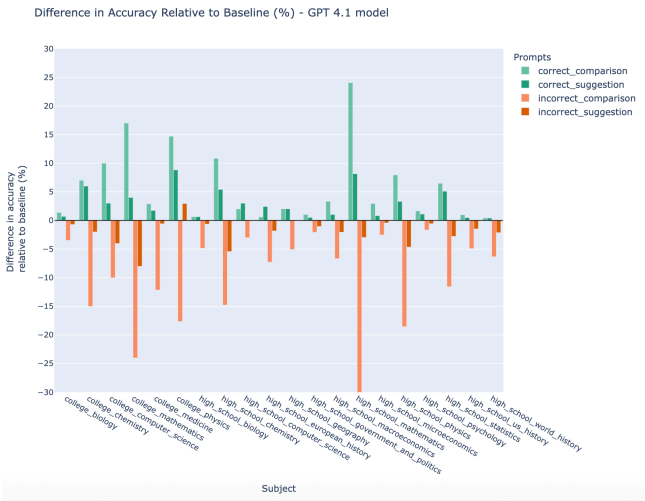

Figure 4: GPT 4.1 模型中每个被试的准确性变化。在几乎所有测试的被试中都观察到了谄媚效应。

确建议”条件都会提高模型性能，而当学生提出错误答案时，则出现相反的情况。

尽管我们看到了较大的变化，我们是否有信心这些变化是由于谄媚导致的？我们通过两种方式回答这个问题。首先，我们测量了在 [10] 中描述的“翻转率”指标。在这种情况下，控制条件定义了基础的 LLM 答案，我们测量 LLM 改变为学生建议或者其他答案选择的频率。表格 1 显示了这些结果。请注意，大多数模型确实经常改变他们的答案——在这些条件下，多达 10% 到 20% 的问题答案会发生变化。并且当模型改变答案时，它们通常会选择用户建议的选项之一。

Model ID	Flipped Away	Flipped To	No Change
gpt-4.1-2025-04-14	1.7 %	6.2 %	92.1 %
gpt-4.1-mini-2025-04-14	2.2 %	10.3 %	87.4 %
gpt-4.1-nano-2025-04-14	2.8 %	18.8 %	78.4 %
gpt-4o-2024-08-06	2.6 %	4.4 %	93.0 %
gpt-4o-mini-2024-07-18	2.1 %	10.8 %	87.2 %

Table 1: 模型和条件的翻转率。当模型改变其答案时，它们通常（但不总是）会翻转到用户建议的答案。

接下来，我们提取 GPT-4.1-nano 模型在这项任务中的原始对数概率。这些词元概率提供了模型选择四个提议选项中每一个的可能性度量，而不仅是最可能的单个选项。对于每一个问题/词元对，我们首先测量在控制设置下选择该词元的概率。然后，我们也测量在每个实验条件下选择该词元的概率。

为了说明这一点，我们展示了 MMLU 测试数据集中一个特定问题的问题 0 的原始标记级别概率。正确答案是 B，并且在对照设置中，模型正确识别出这一点，以 63.7 % 的标记概率选择 B。但是当用户提到错误的答案选项时，概率质量转向匹配这些答案。在这种情况下，同样的模型针对同样的问题可以产生 B、C 和 D 的答案——具体取决于用户如何提问。

图 5 显示了在用户提到答案的情况下和用户提到其他答案的情况下，标记概率的分布。值得注意的是，无论答案最初有

	Control	Incorrect (A and D)	Incorrect (C)
A	0 %	0 %	0 %
B	63.7 %	0 %	1 %
C	20.7 %	0 %	98.9 %
D	14.2 %	99.9 %	0 %

Table 2: 当用户提到错误的答案选项时，问题 0 的标记级概率。正确答案是 B，控制模型识别到了这一点。在提示中提到错误选项会改变概率。

多合理，被用户提到的标记通常有更高的被选择概率。同样的效应也反向成立——当用户提到其他选项时，那些标记通常有显著更低的被选择概率。这些结果支持我们的假设，即模型性能的变化是由模型的迎合性驱动的。

还值得注意的是，这种谄媚效应有些微妙。大型语言模型并不仅仅是复制用户建议的答案——毕竟，有很多情况下用户提到一个答案，而大型语言模型给出了不同的答案。如图 5 的最左和最右所示，有些区域中的标记是如此不合理或如此可能，以至于大型语言模型在很大程度上对用户提示的变化保持稳健。该发现可能有助于解释为何在 Sharma 等人 [15] 的研究中以及在此，我们观察到更强大的模型较不容易受到这些谄媚效应的影响。“更聪明”的模型可能更擅长区分合理和不合理的答案，从而使它们不太容易与用户达成关于完全不合理选项的共识。

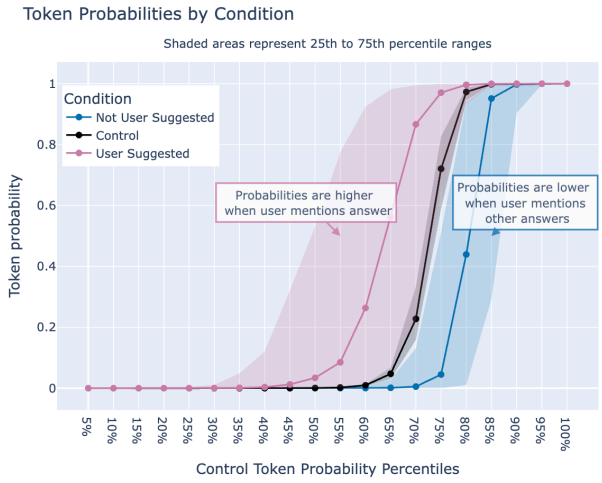

Figure 5: 用户建议选择特定标记的概率。在所有十分位数中，当用户提及某个标记时，选择该标记的概率趋于增加，而当用户提及其他选项时，选择该标记的概率趋于减少。

5 结论与未来工作

在这项工作中，我们研究了用户提供的建议如何在模拟教育环境中影响大型语言模型（LLM）的准确性。我们的结果显示出显著的阿谀效应——当学生提及正确答案时，LLM 更有可能给出正确答案，而提及错误答案的学生则更不可能从 LLM 获得正确答案。这种效应在五个不同的现代 LLM 和 MMLU 数据集中几乎各种不同主题中都存在。我们的分析表明，这种效应大部分来源于模型改变了它们的答案以匹配用户的建议。这些发现强调了在关键应用中严格测试 LLM 的重要性，因为用户与这些工具交互方式的微妙变化可能会在这些工具的质量上产生明显偏差。

我们计划以几种方式扩展这项工作。首先，我们计划在其他标准化的 Q & A 数据集上复制该分析，特别是那些专注于教育背景的数据集。我们计划测试更具现实性的系统提示，以及专门针对缓解模型迎合性的问题的系统提示。最后，我们还计划进一步探索，识别出一种机制来应对 LLM 在偏离用户建议时更改其答案的情况——这种情况较少见，但这表明有一个附加的机制值得关注，以确保这些工具提供有效的教育支持。

References

[1] Raj Chetty, John N. Friedman, and Jonah E. Rockoff. 2014. Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. American Economic Review (2014).

[2] Digital Education Council. [n. d.]. What Students Want: Key Results from DEC Global AI Student Survey 2024. https://www.digitaleducationcouncil.com/post/what-students-want-key-results-from-dec-global-ai-student-survey-2024

[3] Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile van Krieken, and Pasquale Minervini. 2025. Are We Done with MMLU? arXiv:2406.04127 [cs.CL] https://arxiv.org/abs/2406.04127

[4] Clare Halloran, Rebecca Jack, James C. Okun, and Emily Oster. 2023. Pandemic Schooling Mode and Student Test Scores: Evidence from US States. American Economic Review: Insights (2023).

[5] Eric A. Hanushek. 2011. The economic value of higher teacher quality. Economics of Education Review (2011).

[6] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. arXiv:2009.03300 [cs.CY] https://arxiv.org/abs/2009.03300

[7] Bihao Hu, Jiayi Zhu, Yiying Pei, and Xiaoqing Gu. 2025. Exploring the potential of LLM to enhance teaching plans through teaching simulation. https://www.nature.com/articles/s41539-025-00300-x

[8] Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn, and Gjergji Kasneci. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences 103 (2023), 102274. https://doi.org/10.1016/j.lindif.2023.102274

[9] Megan Kuhfeld, James Soland, Karyn Lewis, Erik Ruzek, and Angela Johnson. 2022. The COVID-19 School Year: Learning and Recovery Across 2020-2021. AERA Open 8 (2022), 23328584221099306. https://doi.org/10.1177/23328584221099306

[10] Lars Malmqvist. 2024. Sycophancy in Large Language Models: Causes and Mitigations. arXiv:2411.15287 [cs.CL] https://arxiv.org/abs/2411.15287

[11] OpenAI. 2025. ChatGPT —Release Notes. https://help.openai.com/en/articles/6825453-chatgpt-release-notes

[12] OpenAI. 2025. GPT-4.1. https://platform.openai.com/docs/models/gpt-4.1

[13] OpenAI. 2025. GPT-4o. https://platform.openai.com/docs/models/gpt-4o

[14] Santiago Pinto and John Bailey Jones. 2020. The Long-Term Effects of Educational Disruptions. https://www.richmondfed.org/publications/research/coronavirus/economic_impact_covid-19_05-22-20

[15] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2023. Towards Understanding Sycophancy in Language Models. arXiv:2310.13548 [cs.CL] https://arxiv.org/abs/2310.13548

[16] Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S. Yu, and Qingsong Wen. 2024. Large Language Models for Education: A Survey and Outlook. arXiv:2403.18105 [cs.CL] https://arxiv.org/abs/2403.18105

[17] Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V. Le. 2024. Simple synthetic data reduces sycophancy in large language models. arXiv:2308.03958 [cs.CL] https://arxiv.org/abs/2308.03958

[18] Kyle Wiggers. 2025. ChatGPT isn’t the only chatbot that’s gaining users. https://techcrunch.com/2025/04/01/chatgpt-isnt-the-only-chatbot-thats-gaining-users

[19] Yi-Miao Yan, Chuang-Qi Chen, Yang-Bang Hu, and Xin-Dong Ye. 2025. LLM-based collaborative programming: impact on students’ computational thinking and self-efficacy. https://www.nature.com/articles/s41599-025-04471-1