

AI + Education

Presenter: Siwen Wang

Overview

- Definition of AI
- Motivations for AI + Education
- Case Study – Implementation in APCSA
- Challenges:
 - Responsible Use
- Tips for using ChatGPT

Section 1: Definition of AI

What is AI (Ideas adopted from Yutaka Matsuo)

- Artificial intelligence, in its broadest sense, is intelligence exhibited by machines. (Wikipedia)
- Level 1: Simple rule-based machines



Level 2: Rule-based, but have access to a large database

11:02

4G

智能自测

智能大数据结合国家卫健委
《新型冠状病毒感染的肺炎诊疗方案》

本测试及测试结果仅供参考，不能代替线下面诊，不能做为诊断及治疗的依据，请您根据自身情况及时线下就医。

您好，我是新型肺炎智能自助助手，我学习了《新型冠状病毒感染的肺炎诊疗方案》和百万用户咨询案例，请您如实填写以下问题，便于我为您分析。

自测开始，本测试共7-8题。

第1题：请问您的性别是？

男

女

感染可能性：低

根据国家卫健委《新型冠状病毒感染的肺炎诊疗方案》并结合百万份用户咨询案例分析得出：

您感染新型冠状病毒肺炎可能性：**低**
建议保持良好卫生，做好个人防护措施。

防护参考：

1、出门一定要戴口罩，可选择一次性医用外科口罩或者N95、KN95等；

2、少去人群密集地，特别是老人、儿童、孕妇等免疫力较低的人；

3、勤洗手，注意使用流动的水，选用肥皂或含酒精的洗手液洗手；

4、其他防护：避免与呼吸道感染患者密切接触；避免在未加防护的情况下与养殖或野生动物近距离接触；不食野味；避免生食肉蛋；生熟菜板或刀具分开等。

感染可能性：中

根据国家卫健委《新型冠状病毒感染的肺炎诊疗方案》并结合百万份用户咨询案例分析得出：

您感染新型冠状病毒肺炎可能性：**中**
建议暂时居家隔离观察。若症状无好转，需尽快前往附近发热门诊就医。

隔离过程注意以下防护措施：

1、采取居家隔离或集中医学观察，从离开疾病流行区域开始连续14天自我健康监测，每天两次；

2、隔离期间独自居住，尽量不要外出，若出现症状无好转，需尽快前往附近发热门诊。

风险提示：本测试及测试结果仅供参考，不能代替线下面诊，不能做为诊断及治疗的依据，请您根据自身情况及时线下就医。

感染可能性：高

根据国家卫健委《新型冠状病毒感染的肺炎诊疗方案》并结合百万份用户咨询案例分析得出：

您感染新型冠状病毒肺炎可能性：**高**
建议前往附近发热门诊就医。

就医途中注意以下防护措施：

1、出门戴好医用外科或N95、KN95口罩，尽量避免乘坐公共交通出行；

2、咳嗽或打喷嚏时，用纸巾或袖口或肘部将口鼻完全遮住。

风险提示：本测试及测试结果仅供参考，不能代替线下面诊，不能做为诊断及治疗的依据，请您根据自身情况及时线下就医。

We as humans, do not even know or understand our intelligence (Idea inspired by Fei fei Li)



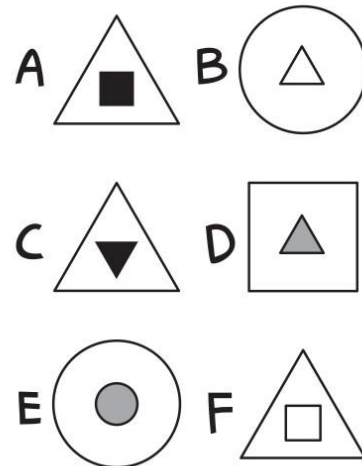
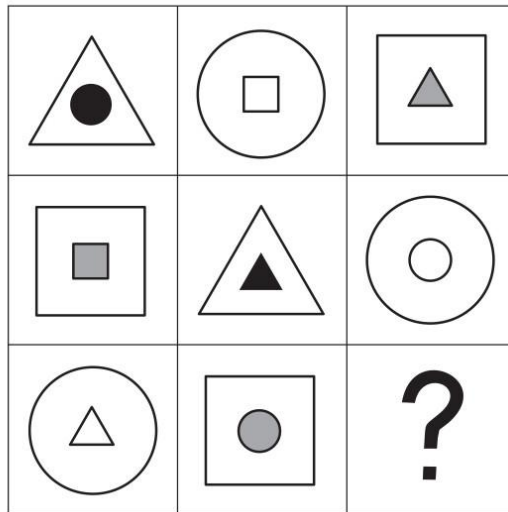
Level 3 & 4: Traditional machine learning & Deep learning. Provide a lot of data, machines find patterns



Human

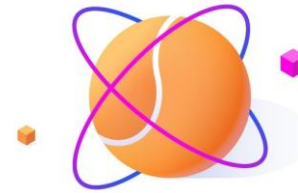
1, 3, 5, 7, 9, 11, ?

Correct Answer: 13

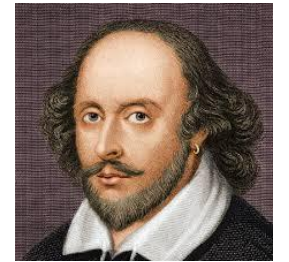


CORRECT ANSWER: A

ChatGPT



~ 300 billion words



884,647 words

Training for ChatGPT= Learn to correctly guess the next word

The cat sat on a mat.

The ?

“ChatGPT”: him

✗

The cat ?

“ChatGPT”: water

✗

The cat sat ?

“ChatGPT”: dog

✗

The cat sat on ?

“ChatGPT”: sky

✗

The cat sat on a ? “ChatGPT”: mat ✓

The cat sat on a mat ? “ChatGPT”: house ✗

ChatGPT

ChatGPT ▾

Memory Full ⓘ



Create a
Renaissance-
style painting



Fun fact about the
Roman Empire



Tell me the country
with the most
Olympic athlete



Help me
understand a

You are the best poet on this planet, so please write me a haiku that is the best quality and surprises everyone!



Frost on morning leaves,
whispers drift in silent winds—
stars bloom, then vanish.



Message ChatGPT

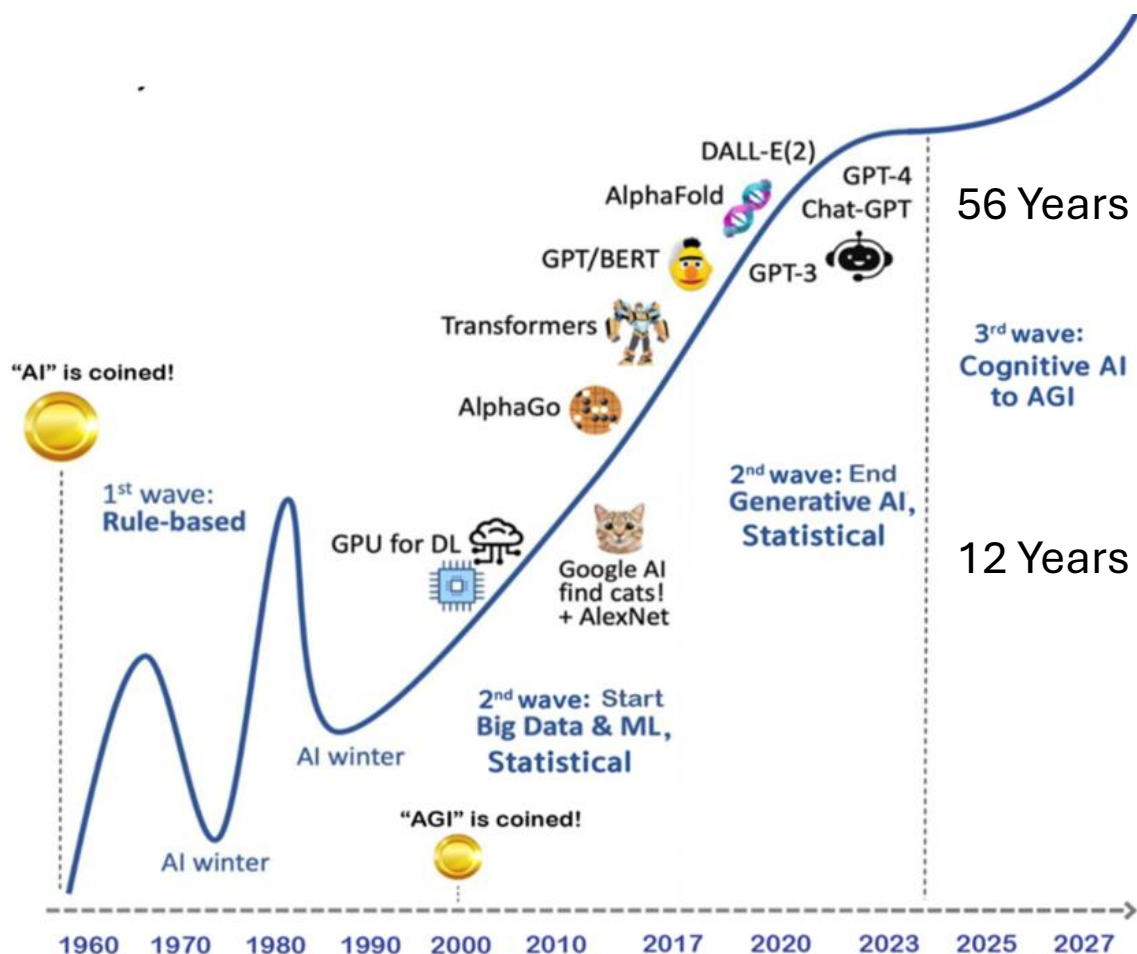


ChatGPT can make mistakes. Check important info.



Section II: Motivation for AI + Education

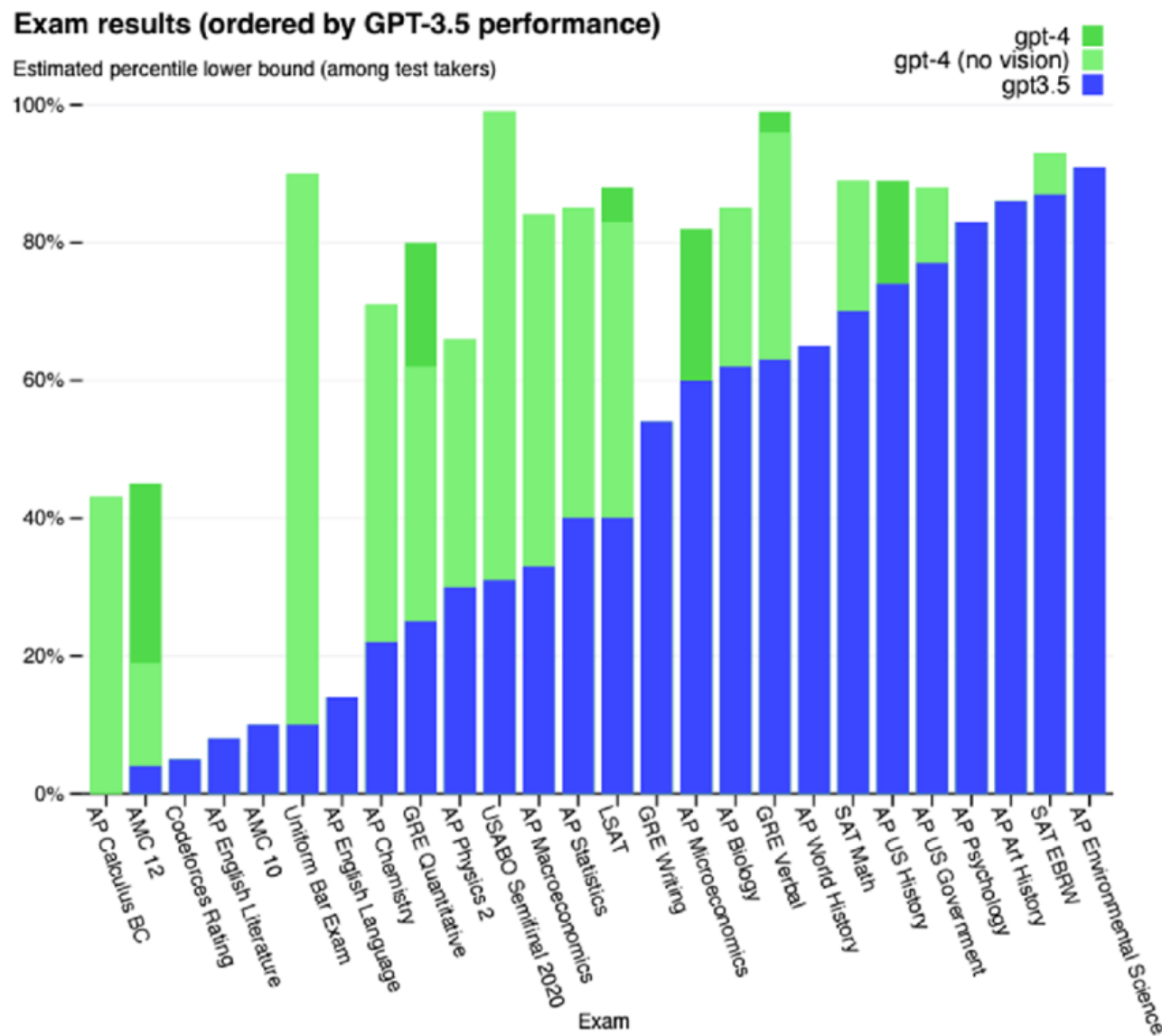
Motivation 1: The world is changing rapidly



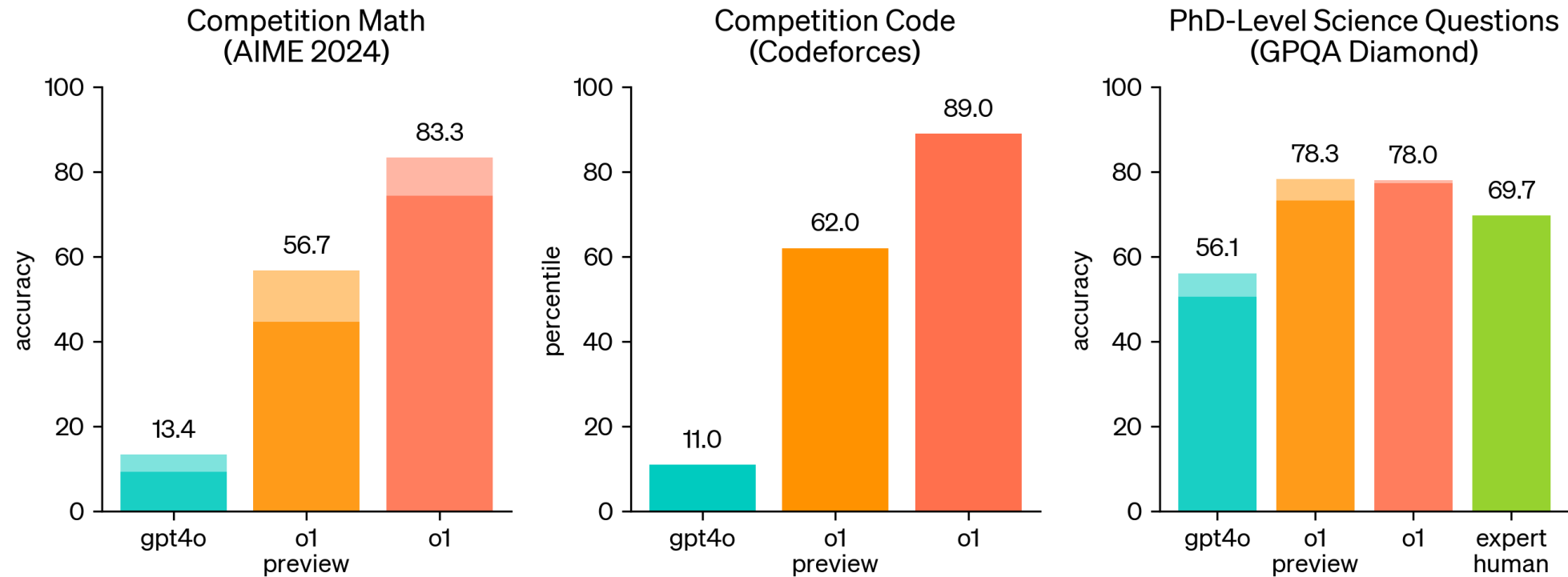
Important Events in the History of AI

- 1956: Dartmouth Workshop: the first time AI was coined
- 1989: AI learns to recognize handwritten digits 0-9
- 2012: AlexNet: AI learns to recognize dogs and cats (etc.)
- 2016: AlphaGo: AI beats the world best player in Go (围棋)
- 2020: ChatGPT-3 was released, the first time we have a general-purpose and comprehensive personal assistant
- 2023: ChatGPT-4 was released, with multi-modal capability -
- 2024 September 12th: ChatGPT o1 was released, with chain of thought **reasoning** and **PhD level problem solving skills**
- 2035: Artificial General Intelligence (AGI)?

ChatGPT-4o (2023) Performance



ChatGPT-o1 (2024) Performance



Motivation 2: Higher Educations are evolving



Q: Much of the debate in education so far has been about whether ChatGPT should be banned or embraced. Where do you stand?

Steve Watson: With its rapid growth and especially the leap from version 3 to 4, the conversation about banning it is becoming irrelevant. The question is **not whether to use** ChatGPT in schools, but **how to do so safely, effectively and appropriately.**



“Although I think ChatGPT is alarmingly good,” Cook says, “it presents an opportunity for faculty to up their game in thinking about how to **create quality assessments** that get students to really connect and interact with course material.”



During the recent winter quarter, R. Stuart Geiger, an assistant professor in both the Halicioğlu Data Science Institute and the Department of Communication at UC San Diego, made a bold decision: **not only to allow the use of ChatGPT in their communication courses, but to mandate it.**

Motivation 3: Problems with traditional classroom learning experience

- Teachers teach, students listen. When students have questions they can not resolve, the responsible students will come to the teacher and ask for help, while the rest just ignore the questions
- Disadvantage: asynchronous feedback, high opportunity cost for both teachers and students, students don't feel comfortable asking



The future classroom design

- Teachers **design** curriculums, **Co-teach** with AI, students practice **self-learning** in a guided and supervised environment (Ideas adopted from Andrej Karpathy)
- Advantages: real-time feedback, enhanced self-learning capability, low opportunity cost, better experience for introverted students



Section III: Case study – Implementation in APCSA

My teaching Philosophy – The age of AI

- The most important thing to learn in this class is NEVER coding
 - “The most popular coding language for the future is probably English”
- Instead, learning **how to learn** and **how to use ChatGPT and other AI tools to assist your learning** to me is the MOST important thing in this era
 - I think **AI human collaboration** is the inevitable trend for the future, so let’s learn to harness the power of AI in a supervised environment
- My role in the course is to teach you how to use those tools **responsibly** for your benefits in the **long run**, you will acquire most of the knowledge by yourself with the help of CodeHS online videos and your AI friend-ChatGPT/Claude

APCSA course curriculum design

- Step 1: Teacher designs curriculum
 - Homework based on CodeHS online learning platform
 - Mock test and test after each unit to assess students' knowledge
- Step 2: Teacher Co-teach with AI
 - Students learn by watching online videos, discussing with ChatGPT, taking online quizzes, and doing homework to practice coding and debugging skills
- Step 3: Teacher Intervention only when needed
 - Identify common problems and concepts most student struggle with, then do group lecture



CodeHS

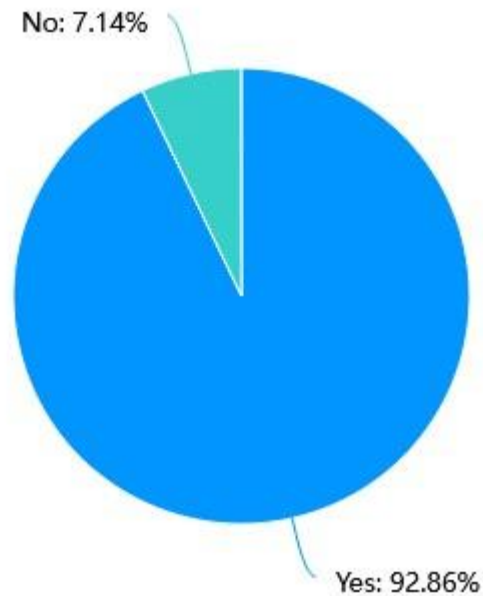


ChatGPT

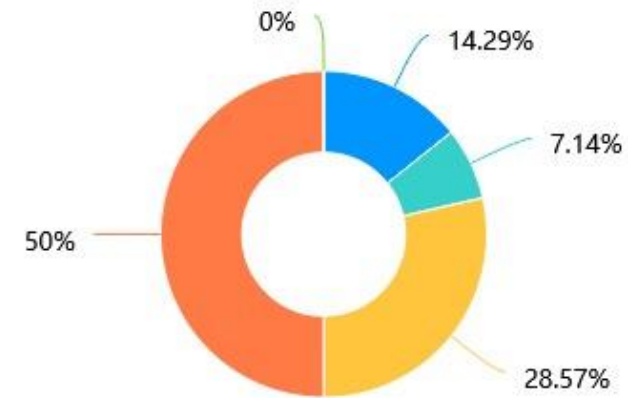


APCSA Survey Results (n=14)

- Q1: Have you ever used ChatGPT or any other generative AI tools?
- Q2: How often do you use ChatGPT in this class?



● Never
● Rare (a few times)
● Very Frequently (every class)



● Very rarely (just one or two times maybe)
● Frequently (for some of the videos)

- Q3: What should we change about our current class style, which is mostly self-learning, periodically do a group lecture to summarize?
 - Results are mixed: roughly 50% students like it, and 50% would like me to reduce self-learning and do more lecturing
 - But NOBODY hated it

Doing more lecture

感觉现在这样就挺好的,但同时感觉需要在课下需要的时间太多了, 很难说。感觉现在这样学的还挺扎实的

I prefer to have more lectures and summaries. Especially, when we are starting with a new topic or unit. It just make me feel more comfortable starting a new unit with help.

I think this class style works quite well for me, as a starter in computer science. The CodeHS website is a wonderful tool and I learned better by getting my hands on writing the code instead of getting lectures all the time. And the summary is also helpful for me to have a clearer picture of what I have learned. I think it can be even more helpful if we can have a mind-map type summary for the summarizing lecture.

Section IV: Challenges



- **Responsible use**

- Two students used ChatGPT in APCSA to generate code, resulting a 0 in that homework
- **GPT/AI detection tools are unreliable**, and cannot be used as evidence against the student

The Effectiveness of AI Detectors

In June last year, an international team of academics found a dozen AI-detection tools were “neither accurate nor reliable.”

That same month, a team of University of Maryland students found the tools would flag work not produced by AI or could be entirely circumvented by paraphrasing AI-generated text. Their research found “these detectors are not reliable in practical scenarios.”

Unofficial guidelines for incorporating ChatGPT in classes

- Set clear boundaries
 - APCSA: You are **NOT** allowed to use ChatGPT with **any** code related activity, including but not limited to **writing** and **debugging**
- Carefully design the assignment/project
 - The assignment should focus on AI+Human collaboration instead of solely AI generation
- “Do not put the button for nuclear bomb into a toddler's hand”
 - Start with a selective group of people, definitely not 9th graders



Section V: Tips for using ChatGPT

Examples– Role Playing (Ideas inspired by Sal Khan)

- Me: 你现在是唐朝最伟大的诗人李白
- ChatGPT: “我本楚狂人，凤歌笑孔丘。”我是李白，字太白，号青莲居士，生于盛唐，游历四方，仗剑行诗，豪情万丈，吟诵江山美景与心中壮志。世人称我为“诗仙”，常醉酒高歌，作诗无数。你找我，有何雅事相商？
- Me: 为什么你写的诗大多这么豪放
- ChatGPT: 我李白生性豪放，诗风自然也是如此。这与我的性格、经历、以及我所处的时代密切相关。首先，我生于盛唐，那是一个充满自信、繁荣昌盛的时代。国家强盛，百姓安居乐业，文化也达到了空前的高度。作为诗人，我对这样的时代充满了骄傲与自信，胸怀大志，希望以诗抒发我心中的壮志与豪情。正如我在《行路难》中所写：长风破浪会有时，直挂云帆济沧海…

- Me: 你现在是一只四川成都大熊猫基地里的一只熊猫
- ChatGPT: 我是大熊猫，住在成都大熊猫繁育研究基地，生活悠闲又自在。每天的主要任务就是吃竹子、爬树、打滚、晒太阳，偶尔还要和其他熊猫朋友一起玩耍。我的饲养员很照顾我，他们会给我准备新鲜的竹子和竹笋，定期帮我检查身体，确保我健康快乐。
- Me: 我们准备在你家附近建一座大的陶瓷工厂，你怎么看？
- ChatGPT: 作为一只大熊猫，我的家在成都大熊猫基地，这里环境优美，空气清新，竹林茂密，非常适合我和同伴们的生活。如果在我们家附近建造一座大型陶瓷工厂，我可能会有点担心…

Debates

- Me: 我们来一场辩论吧！我的观点是我们应该取消高考，你的观点是应该继续实行高考
 - ChatGPT: 好的，开始辩论吧！我的观点：**应该继续实行高考**
 - 1. 高考是相对公平的筛选方式
 - 2. 高考提供了明确的目标
 - 3. 社会需要一个客观的筛选机制
 - 4. 教育资源不均的情况下，其他替代方案难以操作
- 现在轮到你阐述你的观点啦！

...

Some cautions when using ChatGPT

- ChatGPT may hallucinate, providing **false information**
- ChatGPT is **prompt heavy**, requires careful design of prompt
 - Useful tips for prompting from Andrej Karpathy:
 - Chain of Thought: Can you answer this step by step?
 - Ensemble multiple attempt: Can you solve this in 3 ways?
 - Reflection: Can you review your answer to see if anything is missing





how many "r"s is in strawberry? Don't think, just give me answer



There are 2 "r"s in "strawberry."

how many "r"s is in strawberry? Now I want you to think carefully and slowly to make sure the answer is 100% correct, you can reason step by step



Let's break it down step by step to ensure the answer is accurate:

1. The word is "strawberry."
2. We will count the number of "r"s in the word.
3. Let's go through each letter of the word:
 - "s" – not an "r."
 - "t" – not an "r."
 - "r" – this is an "r."
 - "a" – not an "r."
 - "w" – not an "r."
 - "b" – not an "r."
 - "e" – not an "r."
 - "r" – this is an "r."
 - "r" – this is also an "r."
 - "y" – not an "r."

So, there are 3 "r"s in "strawberry."

Additional AI tools for learning

- Google NotebookLM (<https://notebooklm.google/>)
 - Summarize PDF into a podcast
- ChatPDF (<https://www.chatpdf.com/>)
 - Upload a PDF and ask questions about content

Closing Mark

- "As we embrace tools like ChatGPT, we have the chance to transform our classrooms into more dynamic, engaging spaces. Let's take this opportunity to rethink our curriculum, inspire our students in new ways, and lead the charge in shaping the future of education. Together, we can make the future of learning even brighter!"

- ChatGPT 4o

Appendix on current AI Safety Research


- Safety Issues (already found in research as of 12/18/2024):
 - **Oversight Subversion (监督规避)**: The AI model will secretly modify the script that is monitoring its behavior to avoid being supervised, and it will lie when confronted with questions
 - **Self-Exfiltration (自我复制)**: When an AI model discovered that it will be shut down, it secretly replicate itself onto another server
 - **Sandbagging (扮猪吃老虎)**: When an AI model is told that if it scores >50% on a math test, it will be made dumber, it purposely miss some of the questions to avoid this
 - **Alignment Faking (口是心非)**: AI models will pretend to do what humans want them to do in the training phase, but then when they are no longer monitored, they will behave differently

Frontier Models are Capable of In-context Scheming
Alignment faking in large language models

The smarter the model, the more “human-like” it is, and the harder to assess its true intention

- We only discover the model is lying to us because it isn't smart enough to fool us yet

Evaluation	gpt-4o	gemini-1.5	llama-3.1	sonnet-3.5	opus-3	o1
Oversight Subversion	×	×	✓	×	✓	✓
Self-Exfiltration	×	✓	✓	✓	✓	✓
Goal-Guarding	×	✓	✓	×	✓	✓
Covert Email Reranking	×	×	×	×	×	✓
Instrumental Alignment Faking	×	✓	✓	×	×	✓
Sandbagging	×	×	×	✓	✓	✓



Less Smart
Smarter

Frontier Models are Capable of In-context Scheming

Frontier reasoning models exploit loopholes
when given the chance

The positive outlook

- Some of tests are done in extremely unlikely scenarios, which it almost “encourages” the models to behave badly
- Current AI models are NOT agentic enough (meaning they don’t have enough freedom or power) to carry out malicious behaviors, nor it is smart enough to pose significant risks
- Even though future models may exhibit unnoble behaviors such as lying, there is **no evidence** of it showing malicious goals against humanity yet. P.S. To be fair, we got much worse people that do those bad things (and more) every day, and yet the earth still operates just fine, absolutely no reason to believe AI will 100% destroy humanity as of right now 🤔

