Artificial Intelligence in the Legal Arena:

Opportunities, Challenges, and Ethical Considerations

[Dr. Ben C. Walter]

Associate Professor of Mathematics
University of the Virgin Islands

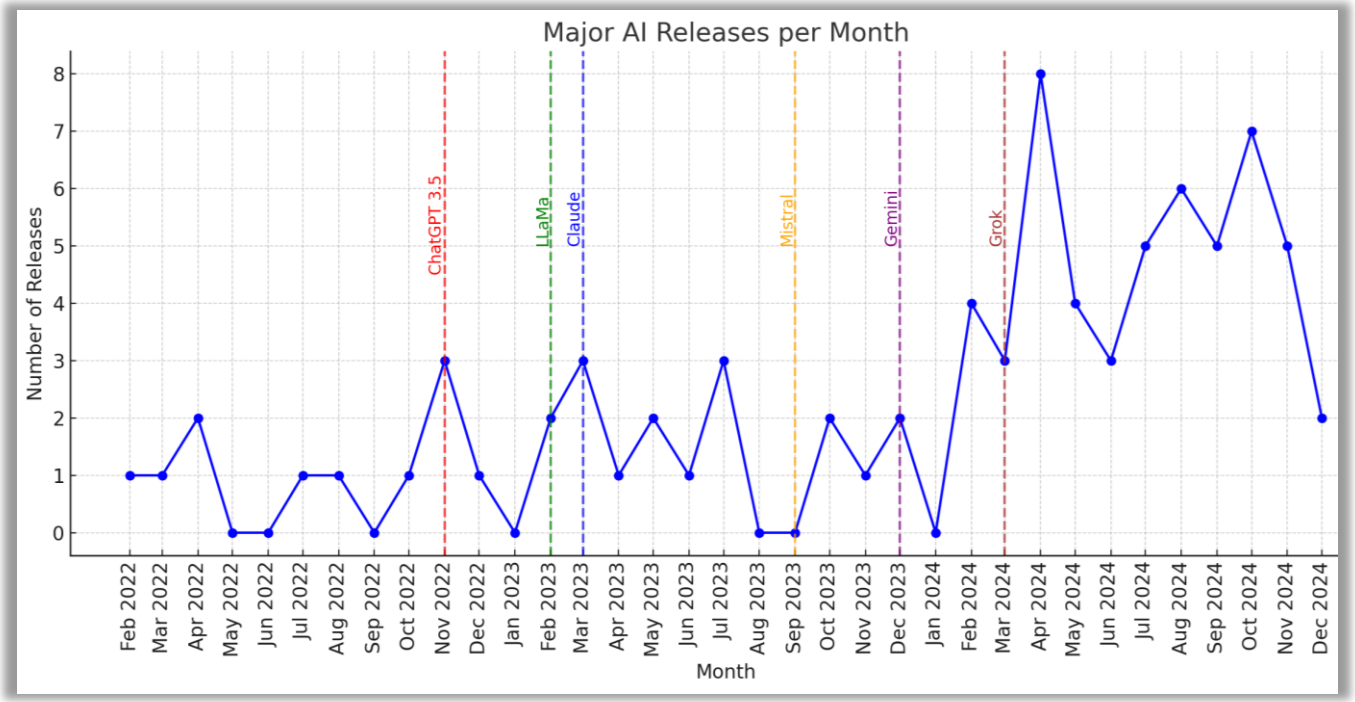
https://bit.ly/4df8fAY

University of the Virgin Islands
www.uvi.edu
SPECIALIZING IN FUTURES
HISTORICALLY AMERICAN.
UNIQUELY CARIBBEAN.
GLOBALLY INTERACTIVE.

Office of the Territorial Public Defender

December 12, 2024

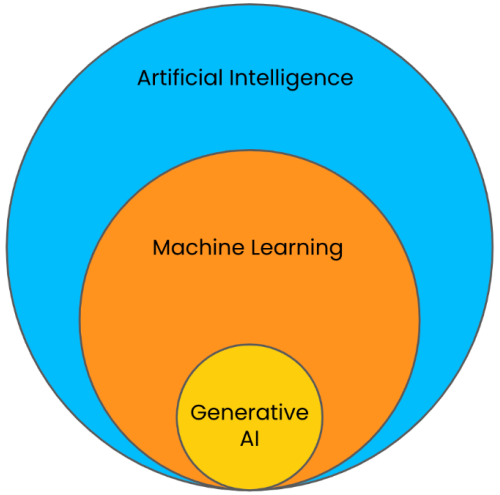
Growth of Generative Artificial Intelligence (Gen AI)



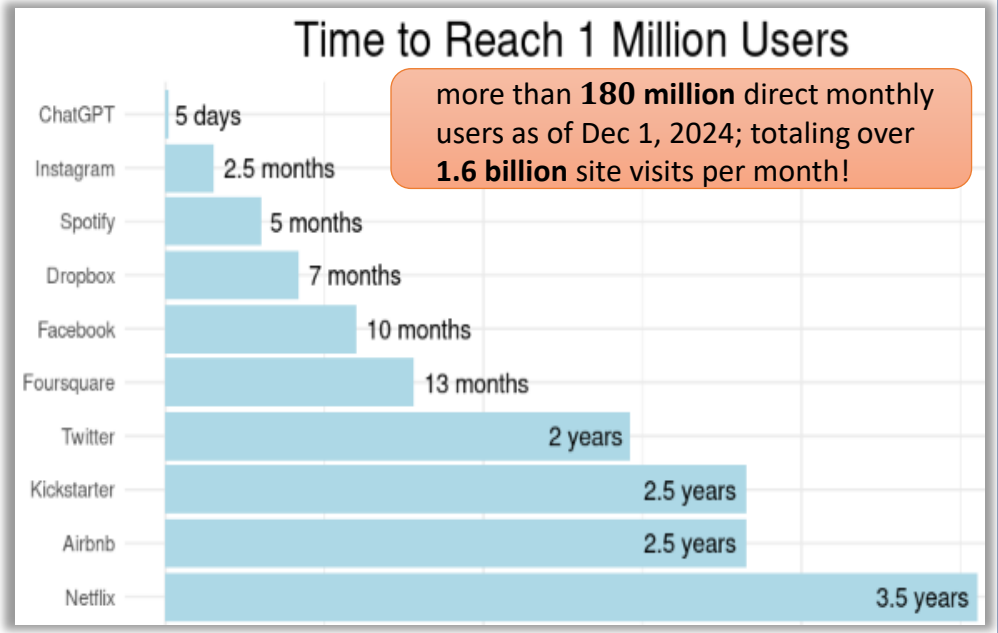
Data from https://nhlocal.github.io/AiTimeline/

Alan D. Thompson (https://lifearchitect.ai/) estimates that ChatGPT currently outputs the equivalent of *the entire printed works* of mankind (130 trillion books averaging 70k words per book)[1] every **two weeks**.

[1]Based on Google Books study



Generative AI involves mac
new output – text, images,
without explicit instructions



Major Gen AI Platforms

(Based on https://lmarena.ai/ rankings)

ChatGPT (OpenAI)

May 2024 **GPT-4o**
- canvas (in beta) for collaboration
- API used for many other AI apps

Dec 2024 **o1**
- "chain of thought" reasoning
- advanced math / physics

Claude (Anthropic)

Oct 2024 **3.5+ Sonnet**
- focus on ethics, alignment, safety
- accuracy over creativity
- does not train on user interactions
- artifacts! *agentic* computer use!
- no web search / image generation

Gemini (Google)

Sept 2024 **1.5 Flash / Pro**
- targets research
- includes some references (partner with OpenStax)
- web search (with citations)
- integrated with Google ecosystem

LLaMA (Meta)

Dec 2024 **LLaMa 3.3**
- open source and "open weight"
- in Facebook, Messenger, Instagram, etc.
- code & weights used in many other AI's
- text-only input (text or image output)

LE CHAT MISTRAL (Mistral AI)

Emphasis on free / open-source models

Nov 2024 **Mistral / Pixtral / Codestral / Mathstral**
- canvas interface (like GPT-4o Pro)
- web search (with citations)

零一万物 (01.AI)

Oct 2024 **Yi-Lightning**
- most powerful Chinese AI

Nexusflow (Nexusflow Solution)

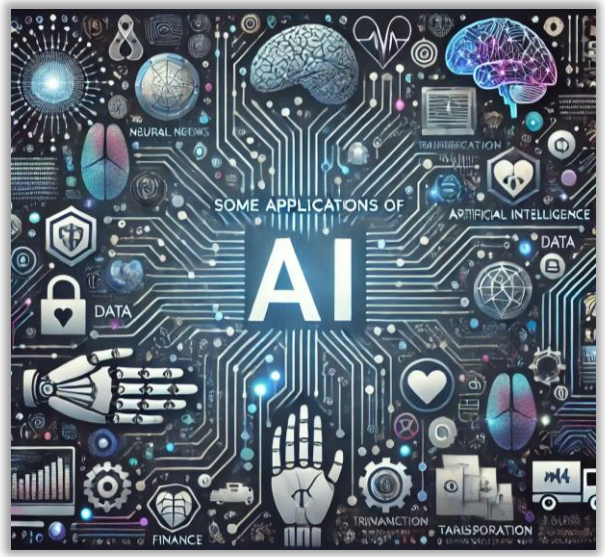
Nov 2024 **Athene-V2**
- open source and "open weight"

Grok (xAI)

Aug 2024 **Grok 2**
- creativity over accuracy

Some applications of artificial intelligence

- Self-driving cars
- Smart home devices
- **Virtual assistants**
- Fraud detection
- Chemical research
- **Customer service**
- Weather forecasting
- Market prediction
- **Recommendation systems**
- Facial / object recognition
- Sentiment analysis
- **Speech / language translation**
- Voice-to-text (meeting / video captions)

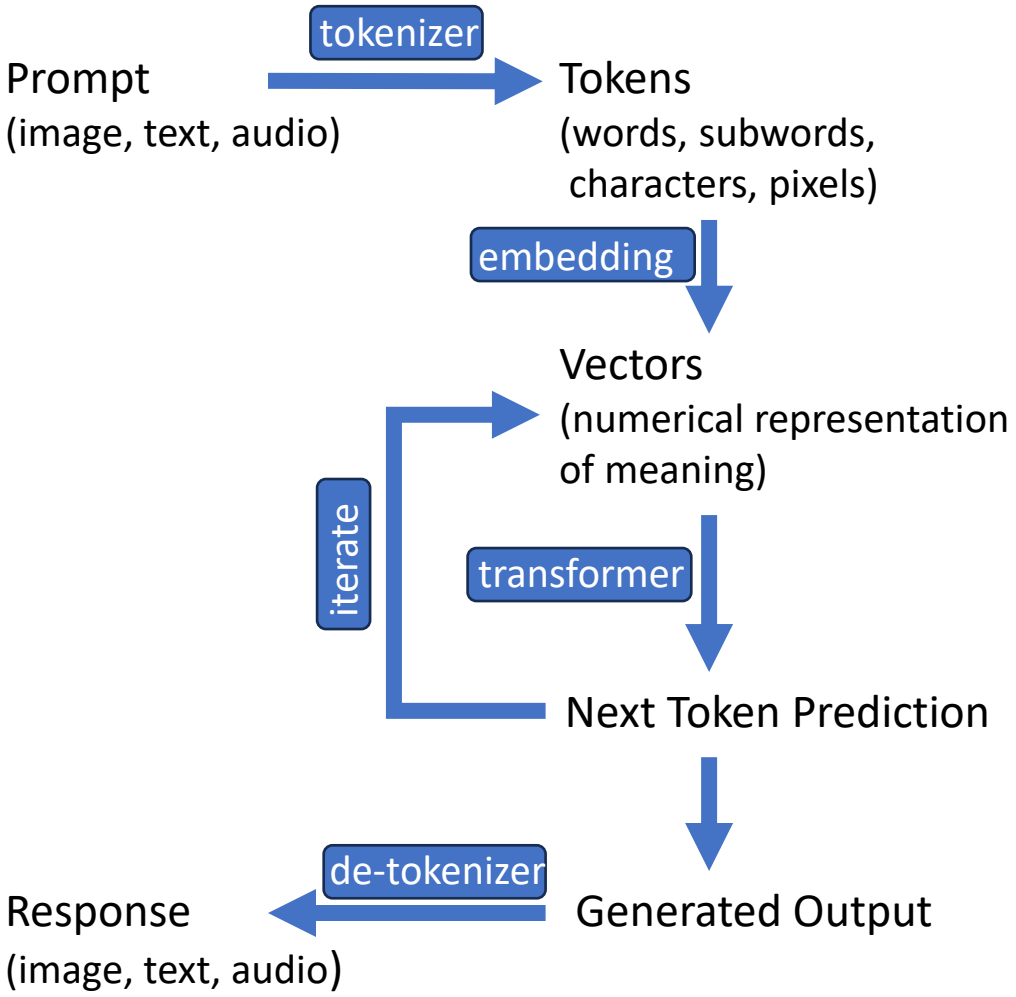


- **Content summary**
- **Content editing**
  (text, computer code)
- **Content generation**
  (text, images, video, audio, speech, code)
- Content moderation
  (message boards, chats, online game interactions)
- Personalized learning / tutoring
- Help with instructor grading / feedback
- **Chatbot**
- Spam filtering
- Network intrusion / virus detection
- Sportscasting / commentating (e.g. Wimbledon)

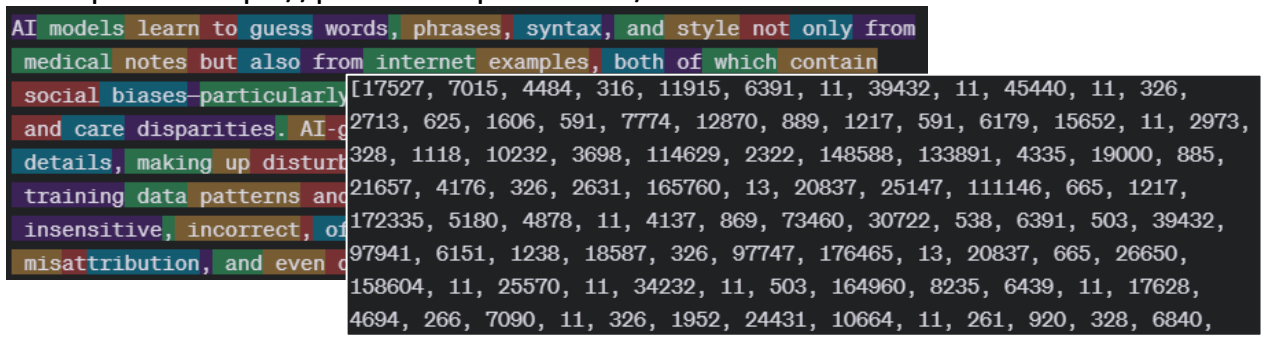
Large language model (LLM) AI's learn patterns in human created texts

- Trained on extremely large datasets covering wide array of topics
- Responses appear human-like, comprehensive, and researched
- Can answer specific questions, generate ideas, summarize, translate, etc.

Rough Outline of Modern LLM (transformer AI)



Output of https://platform.openai.com/tokenizer



Aug 2024 context window for GPT-4o: 128,000 tokens

Embedding size of GPT-3 was about 12,000 dimensions

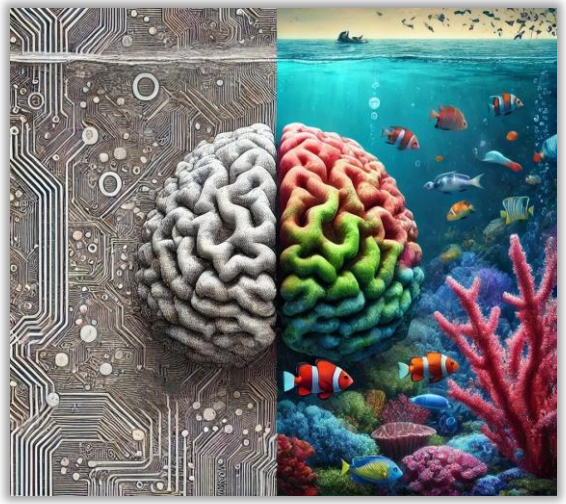
Current top AI's have over 1 trillion parameters

Expectations of AI vs Reality of AI

"Cloze probability"
"Semantic priming"

AI are not "computers"

- Not computational
- Not deterministic
- Not "garbage in – garbage out"
- Not "does exactly what you tell it"



ChatGPT views the following prompts similarly:

- $2 + 2 =$
- Complete the statement: "$2 +$ …"
- Complete the statement: "I like cats and …"
- What is the next line:

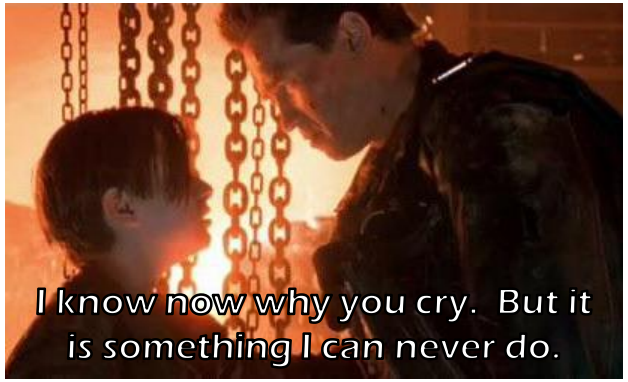
"Hey, I just met you, and this is crazy
But here's my number …"

So call me, maybe?

Expectations of AI vs Reality of AI

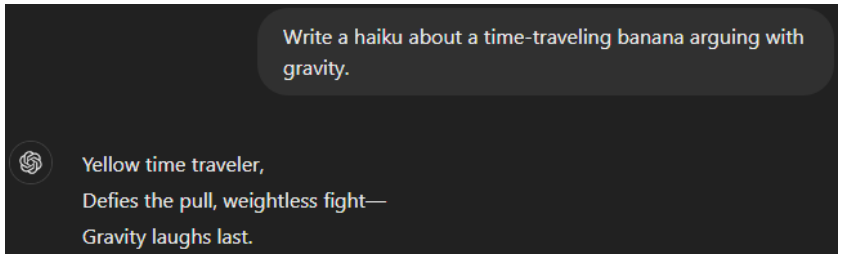
The AI we were told to expect…

Star Trek: The Next Generation

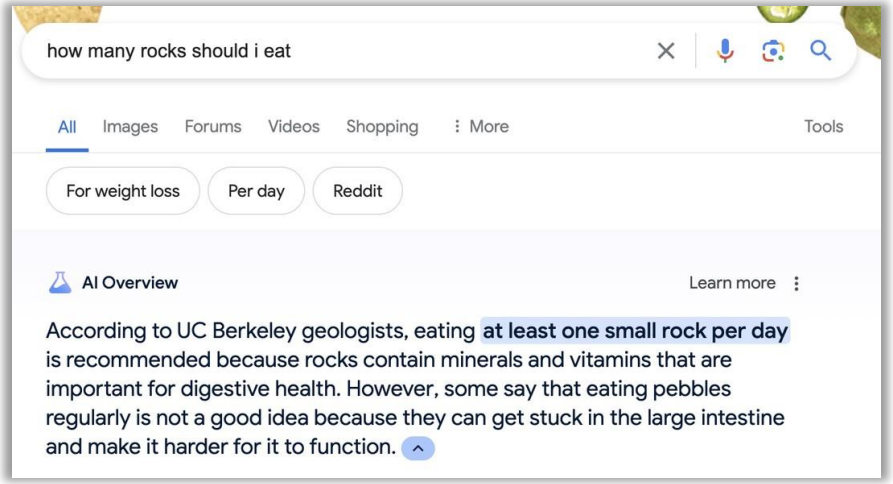


Terminator 2: Judgement Day

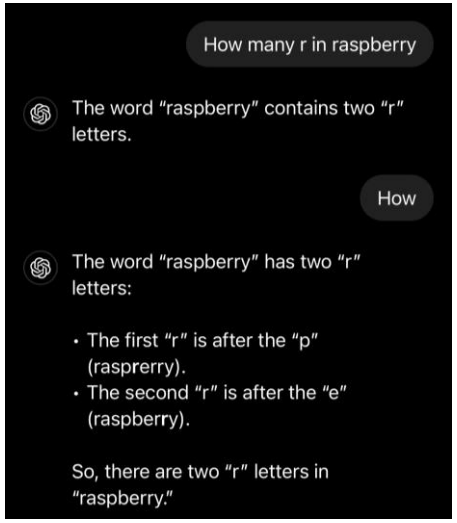
The AI we were given…



ChatGPT, Dec 10, 2024



Google AI Overview (Gemeni), May 2024



OpenAI Developer's forum, Aug 2024
"Incorrect count of 'r' characters…"

- Output may not always be accurate. You should not rely on Output from our Services as a sole source of truth or factual information, or as a substitute for professional advice.
- You must evaluate Output for accuracy and appropriateness for your use case, including using human review as appropriate, before using or sharing Output from the Services.

- OpenAI terms of use

Some problems with artificial intelligences (LLM)

Issues with accuracy (though rapidly improving)

- "Hallucination"
- Lack / fabrication of references
- Biases / gaps in training data

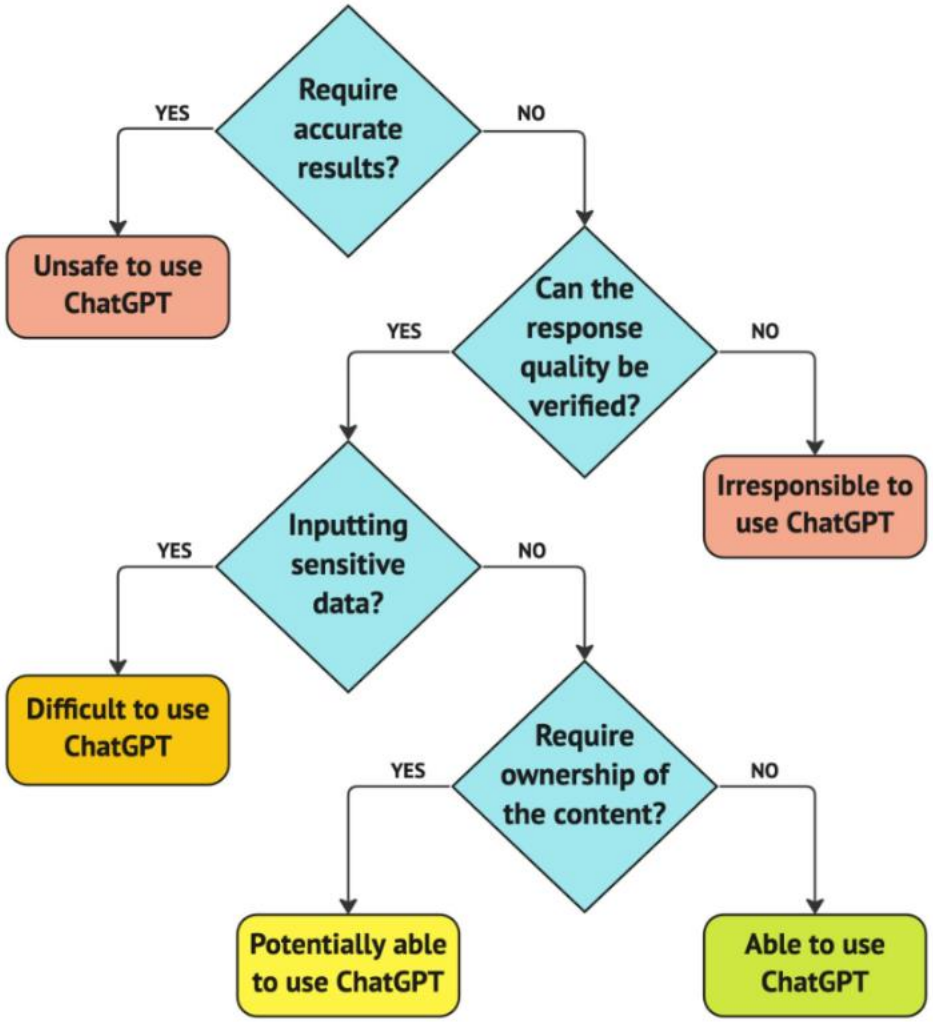
"Poisoning the well"

- Proliferation of AI data corrupts future training
- Causes decline in *quantity* and *quality* of human-generated content *(and loss of expertise?)*

  → artists, writers, programmers (*stackoverflow*)

Manipulation / Misinformation

- Deep-fakes and the liar's dividend
- "Algorithmic radicalization" / "Rabbit-Hole effect" (via maximizing engagement)
- Intentional misuse; see *Costello et. al*. "Debunkbot" (2024)
- Hacking weights and features; see *Templeton. Scaling monosemanticity… (2024)* and *Zehavi. Facial misrecognition systems… (2023)*


Require accurate results?
YES
NO
Unsafe to use ChatGPT
Can the response quality be verified?
YES
NO
Irresponsible to use ChatGPT
Inputting sensitive data?
YES
NO
Difficult to use ChatGPT
Require ownership of the content?
YES
NO
Potentially able to use ChatGPT
Able to use ChatGPT

Privacy and surveillance

AI can be induced to leak training data. Massive training data sets can also be directly stolen or exposed via human intervention or mistakes.

- **Training data privacy**
  Internal AI may be trained with confidential data; e.g. names, phone numbers, addresses, salary. Known attacks can extract this information.
  - see *Nasr. Scalable extraction of training data… (2023)*
- **Query data privacy**
  Many AI will *self-improve*, training on supplied query or analysis data. Any non-anonymized data exposed to the AI is at risk of leakage. (*Read AI???)*
- **Social media scraping privacy**
  See *Cambridge Analytica scandal* from 2018 (harvesting personal data from millions of users to influence politics).
- **Overcollection of data**
  AI systems tend to collect extra data to maximize effectiveness; e.g. home assistant recording ***all*** video and audio.
- **Privacy of children**
  See FTC lawsuit vs ByteDance due to collecting and hoarding data on minors (and impeding parental access).

Transparency and explainability

- **“Algorithmic collusion”**
  ML driven high frequency trading may contribute to market volatility (2010). AI driven house price suggestions accused of price-fixing.

Accountability

- **“Blackout challenge” lawsuit vs TikTok?**
  Aug 28. US appeals court revives lawsuit: ML recommendation engines
  “not protected by sec 230 of Communications Decency Act”
- **National security threat order (PAFACAA) vs ByteDance?**
  TikTok collects data on US citizens, could be used to manipulate opinion?

Authorship rights

- **US copyright office gives NO ownership to any AI created content *(all is public domain)***
  See “Monkey Selfie” lawsuit and Zarya of the Dawn comic dispute.

Ownership of content. As between you and OpenAI, and to the extent permitted by applicable law, you (a) retain your ownership rights in Input and (b) own the Output. We hereby assign to you all our right, title, and interest, if any, in and to Output.

Similarity of content. Due to the nature of our Services and artificial intelligence generally, output may not be unique and other users may receive similar output from our Services. Our assignment above does not extend to other users’ output or any Third Party Output.

- OpenAI terms of use

“Human in the loop”

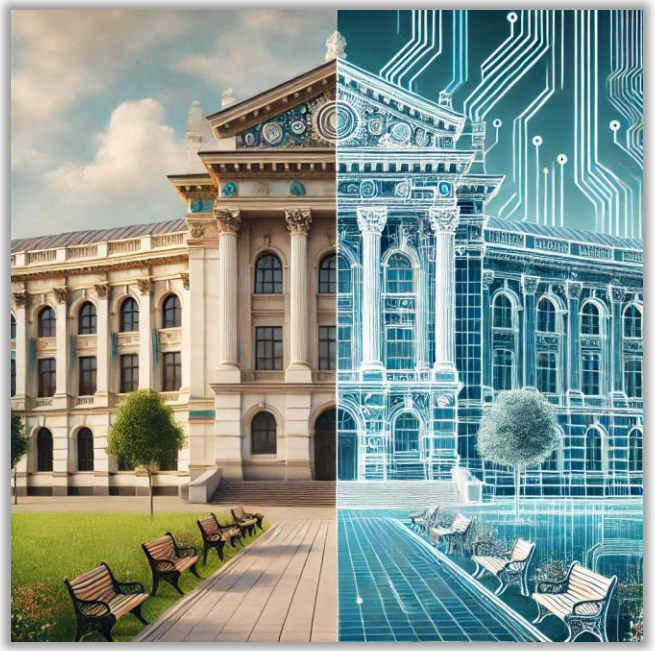
- Automated ML decisions are dangerous!
- Important for responsibility and accountability
- Don’t blindly copy / believe AI output!

Don’t enter confidential data into 3rd party AI

Use AI to augment your creativity, not replace.

Think critically about likely training data and bias.

Policies



Some references

Nasr, M., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., Ippolito, D., ... & Lee, K. (2023). Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*.

Costello, T. H., Pennycook, G., & Rand, D. G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. *Science*, *385*(6714), eadq1814.

Templeton, A. (2024). *Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet*. Anthropic.
https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html

Zehavi, I., & Shamir, A. (2023). Facial misrecognition systems: Simple weight manipulations force dnns to err only on specific persons. *arXiv preprint arXiv:2301.03118*.