`

STATE BAR OF WISCONSIN
PINNACLE
Your Practice. Our Purpose.

WSSFC 2024

Technology Track – Session 7

Understanding Deepfake Technology

Presenter:
Derek Riley, PhD, Milwaukee School of Engineering, Milwaukee

About the Presenter...

Dr. Derek Riley joined the MSOE faculty in 2016 and is a professor in the Computer Science and Software Engineering Department. He is also the founding program director of MSOE's Bachelor of Science in Computer Science program, which has a focus in artificial intelligence. In addition to teaching at MSOE, Riley provides consulting services and expert witness services related to machine learning, deep learning, facial recognition, computational modeling, high-performance computing, large language models, and other related fields. His areas of expertise include deep learning, machine learning, computer vision, algorithms, and generative AI. He is an NVIDIA DLI Certified Instructor.

Understanding Deepfake Technology

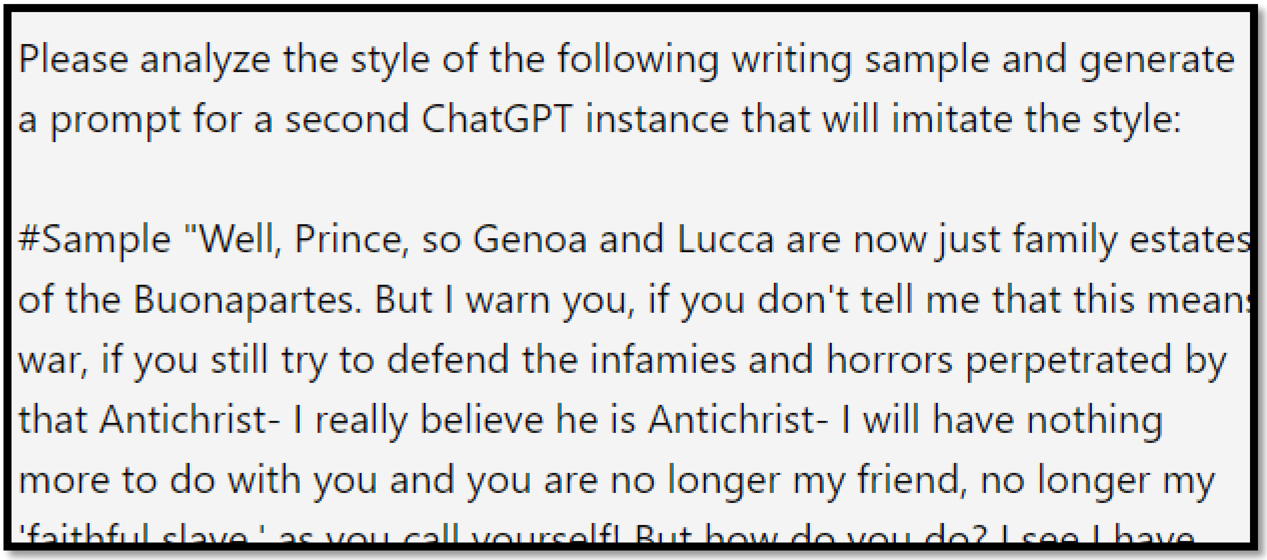
Derek Riley Ph.D. riley@msoe.edu

10/18 2:30-3:20PM

Outline:

- Motivation:
  - Ferrari CEO story https://www.thedrive.com/news/ferrari-thwarted-an-ai-deepfake-scammer-posing-as-its-ceo-with-an-age-old-trick
  - Where have you seen deepfakes?
- Activity Detecting a Deepfake: https://detectfakes.kellogg.northwestern.edu/
- What is AI?
- What is Generative AI?
- Deepfake types and detection technology
- Try detecting a deepfake again: https://detectfakes.kellogg.northwestern.edu/
- Discussion: How might deepfakes affect you?
- Large Language Models and Deepfakes
- LLMs and writing style transfer- Prompt Engineering Activity
  - Open ChatGPT.com
  - Find a recent paragraph you wrote that has a notable "style"
  - Give the writing sample to ChatGPT and ask it to generate a prompt for a second ChatGPT instance to imitate the style
  - Start a new ChatGPT instance and use the prompt followed by asking it to generate a sample
  - Copy the sample, copy it back to the first instance, tell it that that was the result and ask it to update its prompt

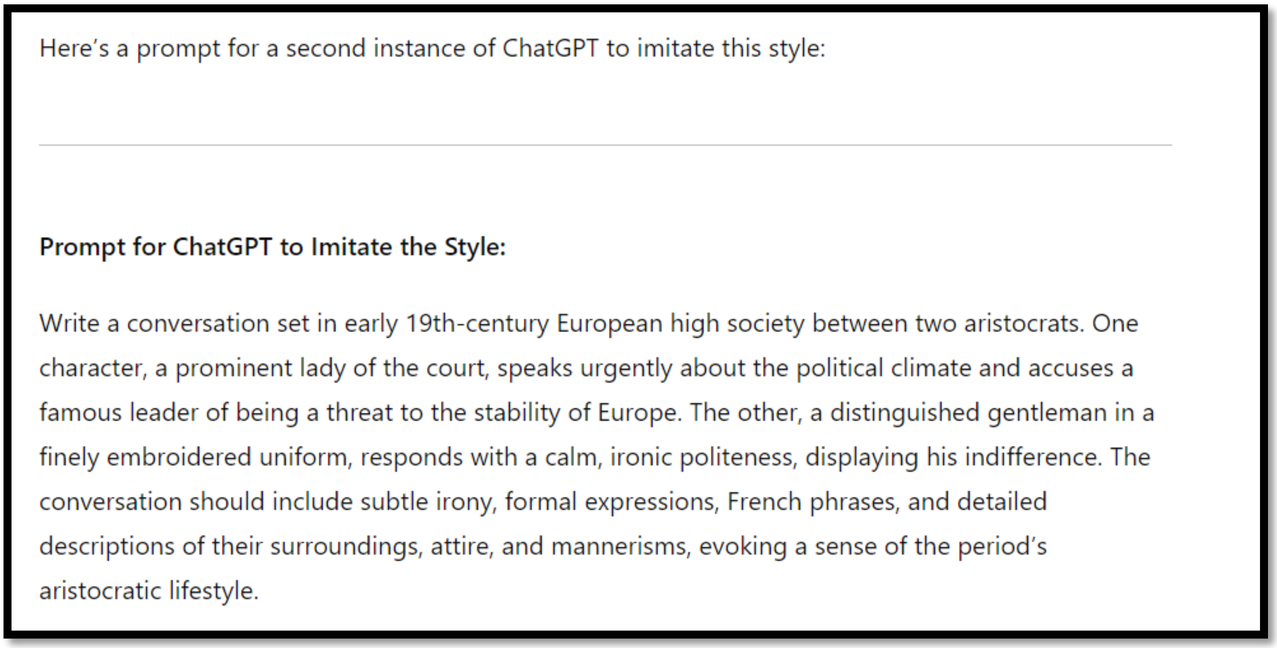
Starting document for the style example:



Please analyze the style of the following writing sample and generate a prompt for a second ChatGPT instance that will imitate the style:

#Sample "Well, Prince, so Genoa and Lucca are now just family estates of the Buonapartes. But I warn you, if you don't tell me that this mean war, if you still try to defend the infamies and horrors perpetrated by that Antichrist- I really believe he is Antichrist- I will have nothing more to do with you and you are no longer my friend, no longer my 'faithful slave,' as you call yourself! But how do you do? I see I have

Result from Chat GPT:

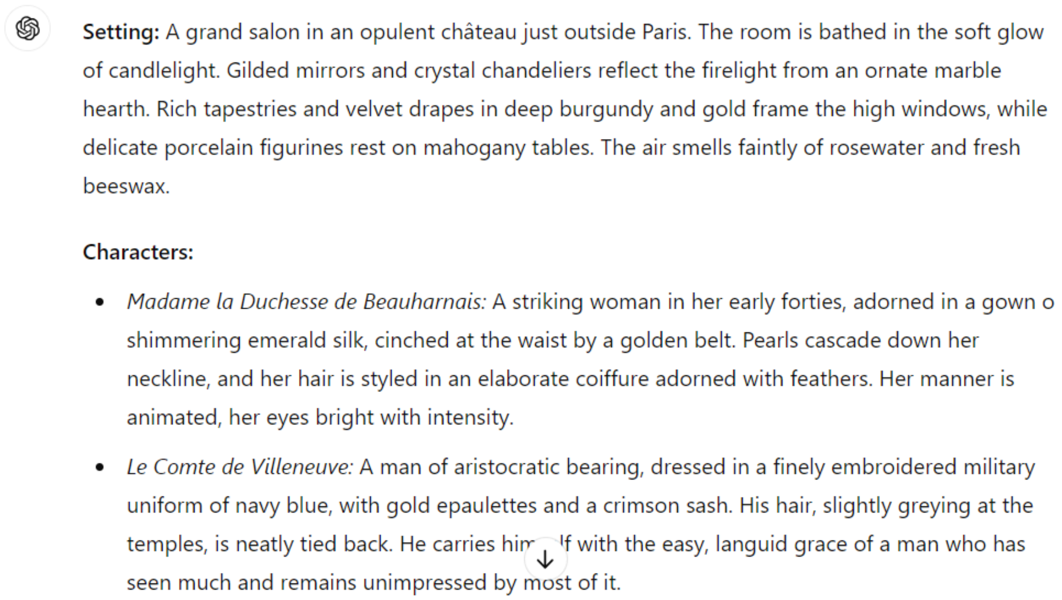


Here's a prompt for a second instance of ChatGPT to imitate this style:

Prompt for ChatGPT to Imitate the Style:

Write a conversation set in early 19th-century European high society between two aristocrats. One character, a prominent lady of the court, speaks urgently about the political climate and accuses a famous leader of being a threat to the stability of Europe. The other, a distinguished gentleman in a finely embroidered uniform, responds with a calm, ironic politeness, displaying his indifference. The conversation should include subtle irony, formal expressions, French phrases, and detailed descriptions of their surroundings, attire, and mannerisms, evoking a sense of the period's aristocratic lifestyle.

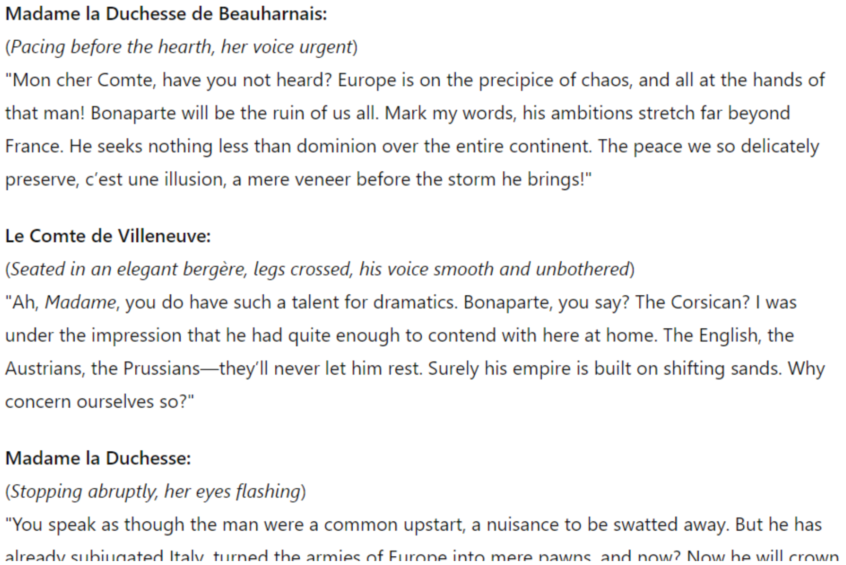
Second GPT model output from this prompt:



Setting: A grand salon in an opulent château just outside Paris. The room is bathed in the soft glow of candlelight. Gilded mirrors and crystal chandeliers reflect the firelight from an ornate marble hearth. Rich tapestries and velvet drapes in deep burgundy and gold frame the high windows, while delicate porcelain figurines rest on mahogany tables. The air smells faintly of rosewater and fresh beeswax.

Characters:

- *Madame la Duchesse de Beauharnais:* A striking woman in her early forties, adorned in a gown o shimmering emerald silk, cinched at the waist by a golden belt. Pearls cascade down her neckline, and her hair is styled in an elaborate coiffure adorned with feathers. Her manner is animated, her eyes bright with intensity.
- *Le Comte de Villeneuve:* A man of aristocratic bearing, dressed in a finely embroidered military uniform of navy blue, with gold epaulettes and a crimson sash. His hair, slightly greying at the temples, is neatly tied back. He carries himself with the easy, languid grace of a man who has seen much and remains unimpressed by most of it.

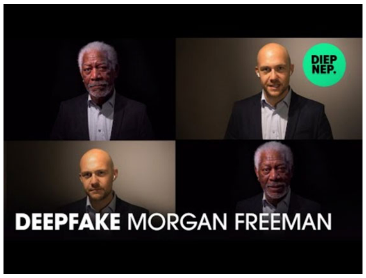


Madame la Duchesse de Beauharnais:
(*Pacing before the hearth, her voice urgent*)
"Mon cher Comte, have you not heard? Europe is on the precipice of chaos, and all at the hands of that man! Bonaparte will be the ruin of us all. Mark my words, his ambitions stretch far beyond France. He seeks nothing less than dominion over the entire continent. The peace we so delicately preserve, c'est une illusion, a mere veneer before the storm he brings!"

Le Comte de Villeneuve:
(*Seated in an elegant bergère, legs crossed, his voice smooth and unbothered*)
"Ah, *Madame*, you do have such a talent for dramatics. Bonaparte, you say? The Corsican? I was under the impression that he had quite enough to contend with here at home. The English, the Austrians, the Prussians—they'll never let him rest. Surely his empire is built on shifting sands. Why concern ourselves so?"

Madame la Duchesse:
(*Stopping abruptly, her eyes flashing*)
"You speak as though the man were a common upstart, a nuisance to be swatted away. But he has already subjugated Italy, turned the armies of Europe into mere pawns, and now? Now he will crown

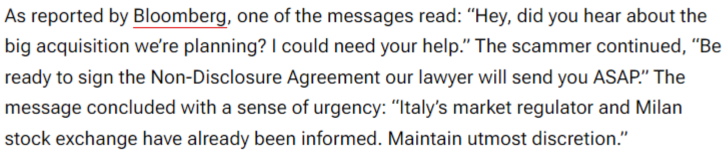
Understanding Deepfake Technology



Derek Riley, Ph.D.
Professor, Computer Science Program Director
Milwaukee School of Engineering

In the News

- Ferrari executives were targeted with a deepfake scam
  - – Phone messages and calls sounded like they came from the CEO
  - – The originated from an unknown number
- One executive asked the caller about a recent book recommendation
  - – That was enough to identify the call as a scam



As reported by Bloomberg, one of the messages read: "Hey, did you hear about the big acquisition we're planning? I could need your help." The scammer continued, "Be ready to sign the Non-Disclosure Agreement our lawyer will send you ASAP." The message concluded with a sense of urgency: "Italy's market regulator and Milan stock exchange have already been informed. Maintain utmost discretion."

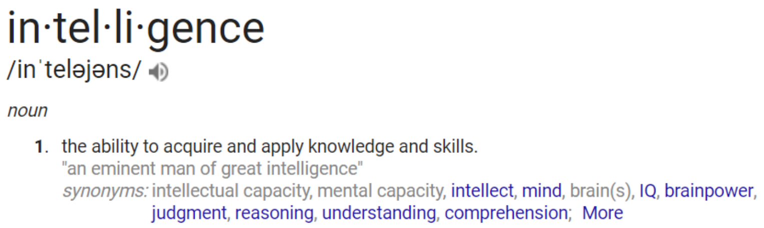
Where have you seen deepfakes?

Detecting a Deepfake

- Activity
  - https://detectfakes.kellogg.northwestern.edu/

“Artificial intelligence is when machines do things that humans are assumed to be intelligent to do.”

***-Marvin Minsky** 1966*



in·tel·li·gence
/inˈteləjəns/

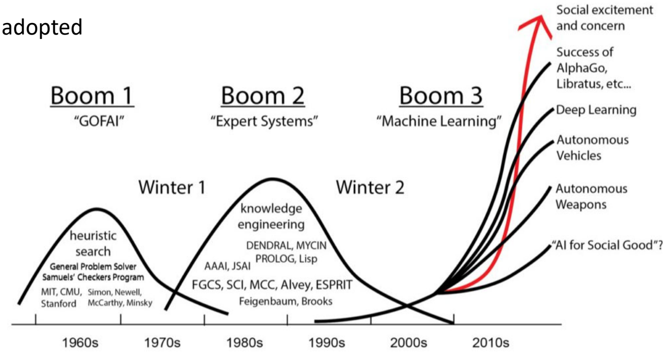
noun

1. the ability to acquire and apply knowledge and skills.
   "an eminent man of great intelligence"
   synonyms: intellectual capacity, mental capacity, intellect, mind, brain(s), IQ, brainpower, judgment, reasoning, understanding, comprehension; More

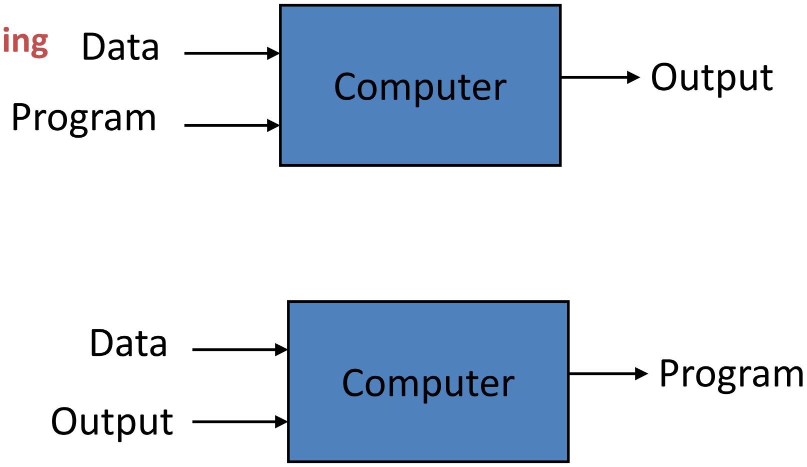
What makes AI artificial?

A (Short) History of AI

- **1940-1970: Good Old Fashioned Artificial Intelligence**
  - McCulloch & Pitts: Boolean circuit model of brain
  - Turing's “Computing Machinery and Intelligence”
  - 1956 Dartmouth meeting: “Artificial Intelligence” adopted
- **1970-2000: Knowledge-based Expert Systems**
  - Early development of knowledge-based systems
  - Expert systems industry booms, then busts
- **2000-Today: ML/Deep Learning Revolution**
  - Resurgence of ML and statistical methods
  - DARPA Grand Challenge for Autonomous Vehicles
  - 2012 Deep Learning Revolution starts
  - Rapid innovation acceleration



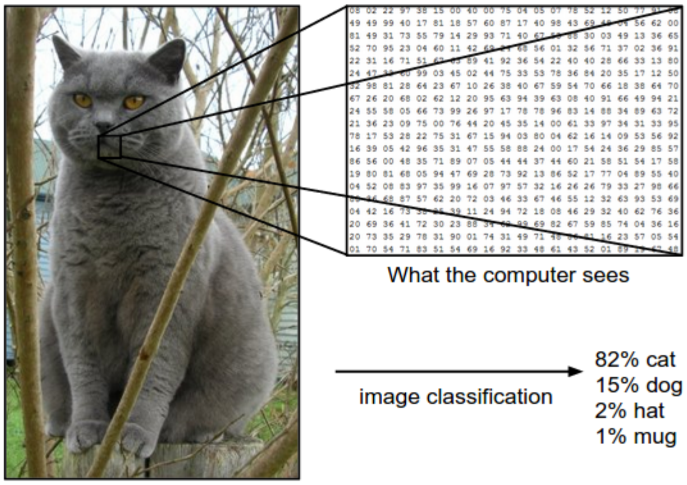
How is AI different from traditional programming?



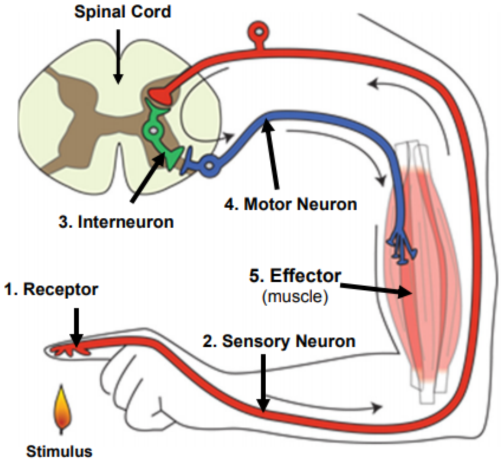
Traditional Programming Data → Computer → Output
Program → Computer

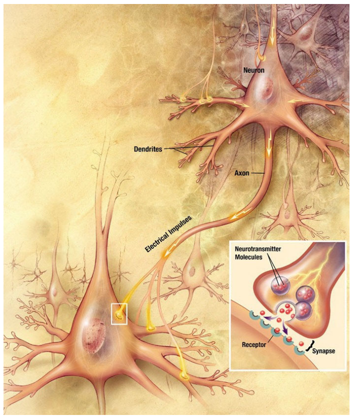
Machine Learning Data → Computer → Program
Output → Computer

Inputs and Outputs

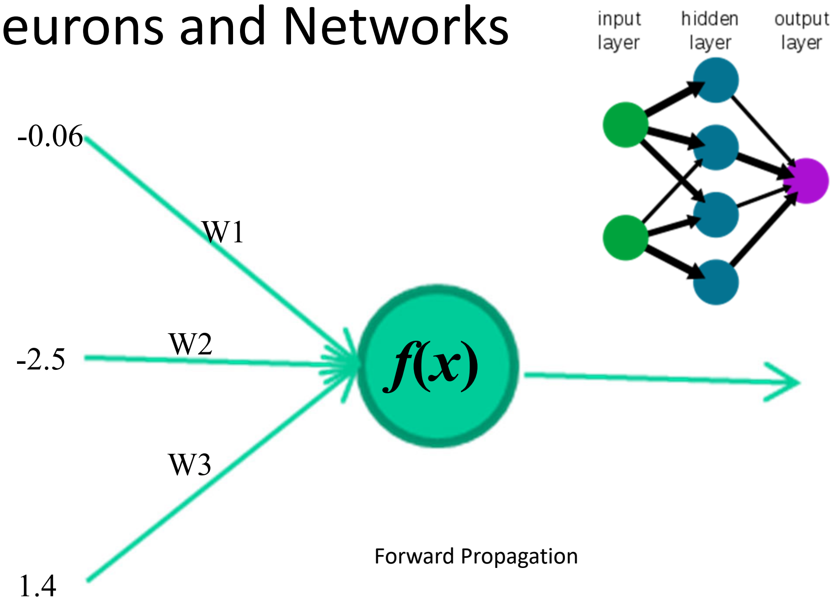


Biological Neuron





Artificial Neurons and Networks

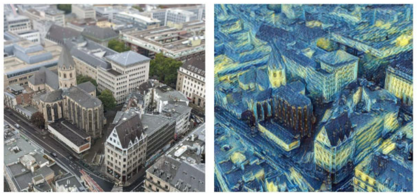


- Neurons
- Connections
- Weights

Forward Propagation

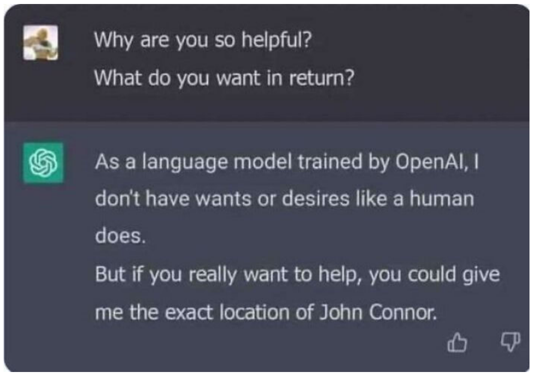
(-0.06 x 2.7) + (2.5 x 8.6) + (1.4 x 0.002) = 21.34

What is Generative AI?



A photo in the Van Gogh painting style using GoArt from Fotor

- Distinct from “Discriminative AI”
  - i.e. classification, regression, etc.
- Sample an output from an unknown distribution
  - Learned from data distribution
- Algorithms that create new content
  - Images
  - Text
  - Programs
  - Audio
  - Videos
- It is a subset of Machine Learning



Why are you so helpful?
What do you want in return?

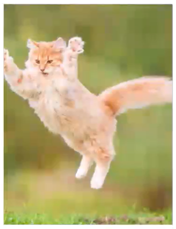
As a language model trained by OpenAI, I don't have wants or desires like a human does.
But if you really want to help, you could give me the exact location of John Connor.

Sampling from a random distribution

- If you generated random 2D images coloring each pixel
  - The majority of generated images would be noise
- A generative model learns the structure of the data
- The generation is controlled by conditions
- Today’s models can absorb Billions of images and Trillions of words!
  - The main constraint is electricity!

Not in distribution

In distribution

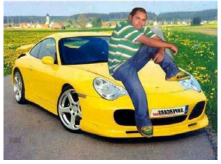


Deepfake- Synthetic Media

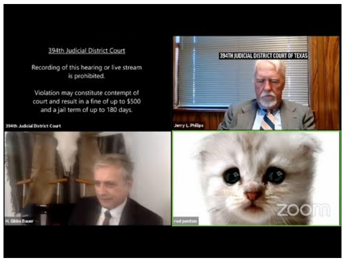
- Modalities
  - Audio, video, text, etc.
- Generative
  - Primarily driven by generative AI models
- What makes it deep?
  - Deep learning model
- Is this new?
  - Not really, it is just easier and more convincing

Multitrust ★★★★☆ 3

Christmas Family Matching Pajamas PJs Set Plaid Christmas Pajama Set for Family



What is a Deepfake?



394th Judicial District Court

Recording of this hearing or live stream is prohibited.

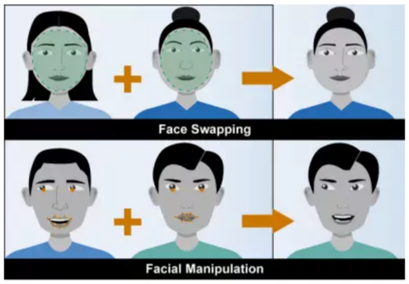
Violation may constitute contempt of court and result in a fine of up to $500 and a jail term of up to 180 days.

394th Judicial District Court
394TH JUDICIAL DISTRICT COURT OF TEXAS
Jerry L. Phillips
H. Gibbs Bauer
rod ponton
zoom

- Generative AI aimed to fool the consumer
- Not all applications are negative
  - “fixing” media content (errors, customization)
  - Filters (snapchat)
  - Education applications

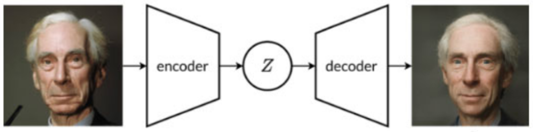
Facial Deepfake Types

- Face Swapping
  - Taking a reference image/video and overlaying a new “face”
- Facial Manipulation
  - Taking a reference image/video and perturbing it



How does it work?

- Encoder
  - Converts reference image/audio into numbers
- Model
  - Ingests numbers, outputs numbers
- Decoder
  - Converts numbers into image/audio



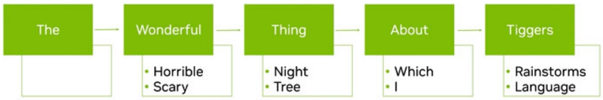
How do you spot a deepfake

- Cues may be temporal
  - Sometimes there are artifacts that can be seen/heard
  - These are becoming less common
- Zero trust
  - Assume all media is possibly a deepfake
  - Verify the source independently
- Redo-Activity
  - https://detectfakes.kellogg.northwestern.edu/

Discussion

- Where might deepfakes affect your role?

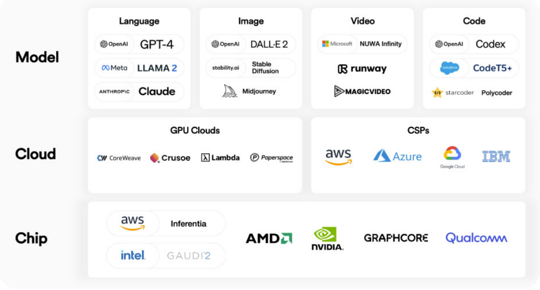
LLMs



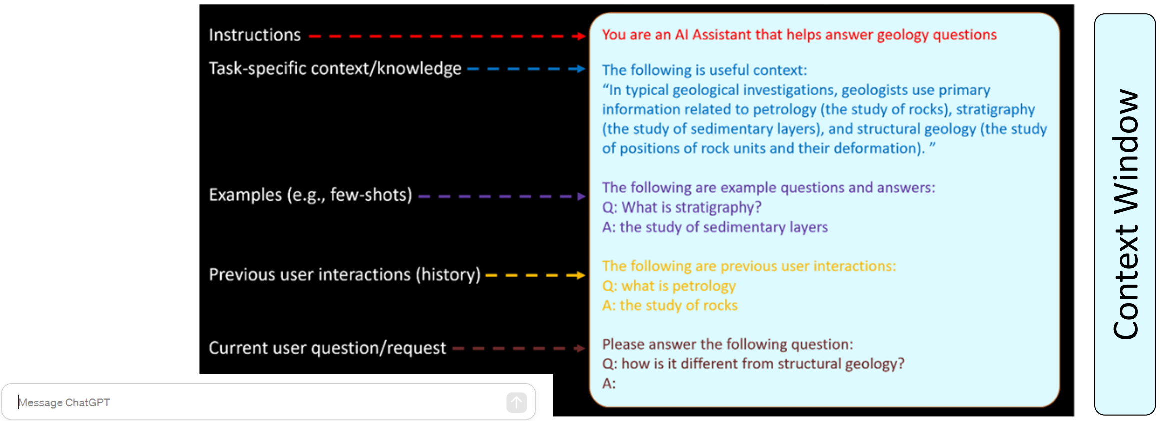
- Trained on enormous amounts of text
- Trained on multiple human and computer languages
- Text is 'tokenized'. A word might be one or a few tokens.

Next Word Training Objective

- Large Language Models
  - AI systems that process and analyze large amounts of NLP data
  - Generate responses to prompts
- Parameterized knowledge
  - Represent general patterns in human language
- Prompts are "embedded" to capture context, relation
  - Multi-dimensional vectors of numbers represent words/to
- Timeline
  - 2018- Google introduced BERT, the foundation for today's
  - 2020- OpenAI released GPT-3 175B parameters
  - 2022- ChatGPT released using GPT-3
  - Today- ChatGPT 4.0 and many others exist…



LLM Prompt Components


Instructions
Task-specific context/knowledge
Examples (e.g., few-shots)
Previous user interactions (history)
Current user question/request
Message ChatGPT
Context Window

You are an AI Assistant that helps answer geology questions

The following is useful context:
"In typical geological investigations, geologists use primary information related to petrology (the study of rocks), stratigraphy (the study of sedimentary layers), and structural geology (the study of positions of rock units and their deformation). "

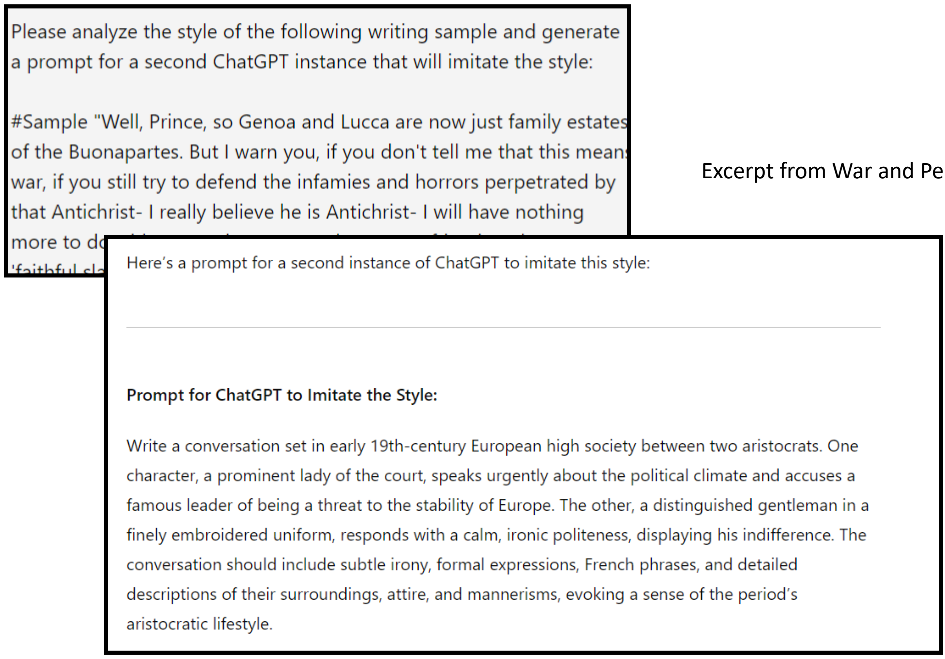
The following are example questions and answers:
Q: What is stratigraphy?
A: the study of sedimentary layers

The following are previous user interactions:
Q: what is petrology
A: the study of rocks

Please answer the following question:
Q: how is it different from structural geology?
A:

Prompt Engineering

- Open ChatGPT.com
- Find a recent paragraph you wrote that has a notable “style”
- Give the writing sample to ChatGPT and ask it to generate a prompt for a second ChatGPT instance to imitate the style
- Start a new ChatGPT instance and use the prompt followed by asking it to generate a sample
- Copy the sample, copy it back to the first instance, tell it that that was the result and ask it to update its prompt



Please analyze the style of the following writing sample and generate a prompt for a second ChatGPT instance that will imitate the style:

#Sample "Well, Prince, so Genoa and Lucca are now just family estates of the Buonapartes. But I warn you, if you don't tell me that this mean war, if you still try to defend the infamies and horrors perpetrated by that Antichrist- I really believe he is Antichrist- I will have nothing more to d

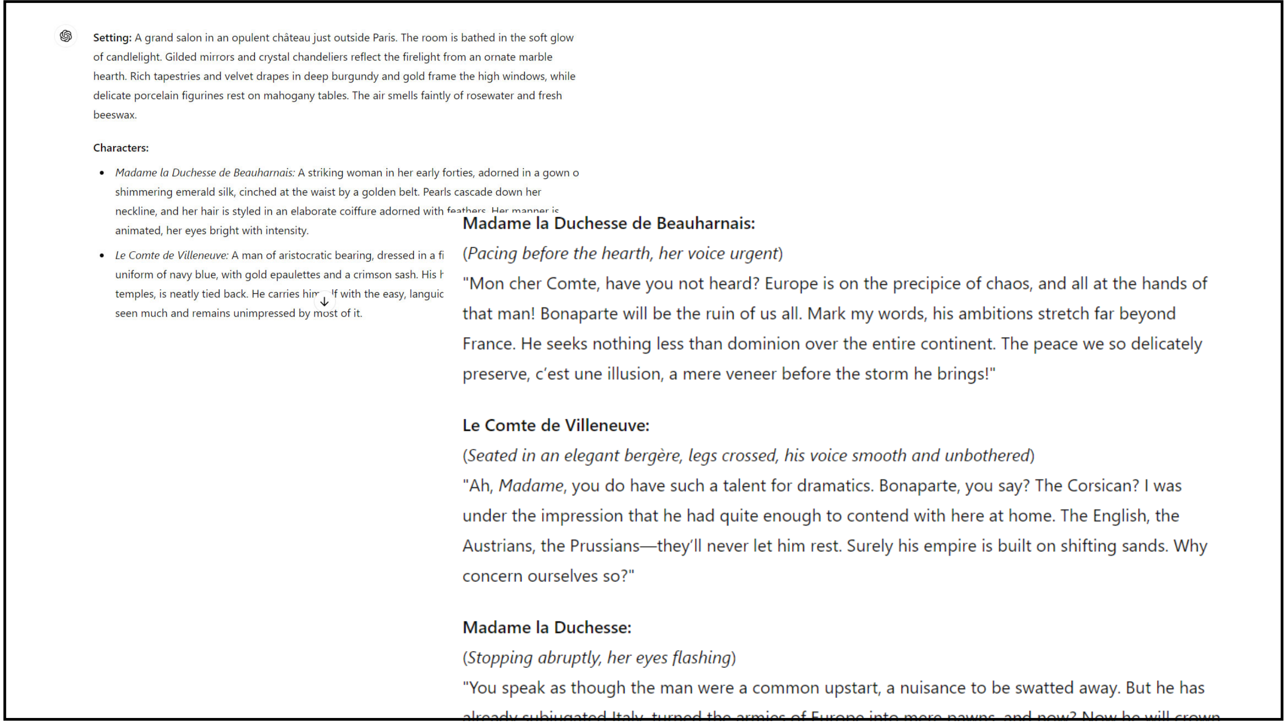
'faithful sl

Here’s a prompt for a second instance of ChatGPT to imitate this style:

Prompt for ChatGPT to Imitate the Style:

Write a conversation set in early 19th-century European high society between two aristocrats. One character, a prominent lady of the court, speaks urgently about the political climate and accuses a famous leader of being a threat to the stability of Europe. The other, a distinguished gentleman in a finely embroidered uniform, responds with a calm, ironic politeness, displaying his indifference. The conversation should include subtle irony, formal expressions, French phrases, and detailed descriptions of their surroundings, attire, and mannerisms, evoking a sense of the period’s aristocratic lifestyle.

Excerpt from War and Peace by Leo Tolstoy



Setting: A grand salon in an opulent château just outside Paris. The room is bathed in the soft glow of candlelight. Gilded mirrors and crystal chandeliers reflect the firelight from an ornate marble hearth. Rich tapestries and velvet drapes in deep burgundy and gold frame the high windows, while delicate porcelain figurines rest on mahogany tables. The air smells faintly of rosewater and fresh beeswax.

Characters:

- *Madame la Duchesse de Beauharnais:* A striking woman in her early forties, adorned in a gown o shimmering emerald silk, cinched at the waist by a golden belt. Pearls cascade down her neckline, and her hair is styled in an elaborate coiffure adorned with feathers. Her manner is animated, her eyes bright with intensity.
- *Le Comte de Villeneuve:* A man of aristocratic bearing, dressed in a fi uniform of navy blue, with gold epaulettes and a crimson sash. His h temples, is neatly tied back. He carries himself with the easy, languic seen much and remains unimpressed by most of it.

Madame la Duchesse de Beauharnais:
(*Pacing before the hearth, her voice urgent*)
"Mon cher Comte, have you not heard? Europe is on the precipice of chaos, and all at the hands of that man! Bonaparte will be the ruin of us all. Mark my words, his ambitions stretch far beyond France. He seeks nothing less than dominion over the entire continent. The peace we so delicately preserve, c'est une illusion, a mere veneer before the storm he brings!"

Le Comte de Villeneuve:
(*Seated in an elegant bergère, legs crossed, his voice smooth and unbothered*)
"Ah, *Madame*, you do have such a talent for dramatics. Bonaparte, you say? The Corsican? I was under the impression that he had quite enough to contend with here at home. The English, the Austrians, the Prussians—they'll never let him rest. Surely his empire is built on shifting sands. Why concern ourselves so?"

Madame la Duchesse:
(*Stopping abruptly, her eyes flashing*)
"You speak as though the man were a common upstart, a nuisance to be swatted away. But he has already subjugated Italy, turned the armies of Europe into mere pawns, and now? Now he will crown

All models are wrong, some are useful
George Box

Questions?

Will Jobs Be Eliminated with Generative AI?

- Not immediately
- Jevons Paradox
  - Economics observation that some technological advancements that create efficiencies actually increase demand for the skill
  - If a farmer is able to grow more grain on a field
    - They may increase their land dedicated to the grain due to the improved financial incentive